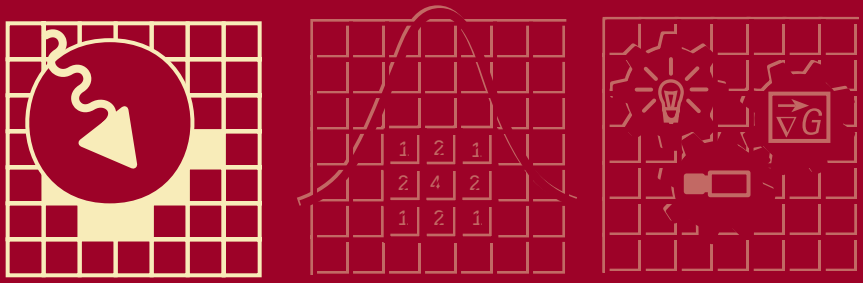


HANDBOOK OF COMPUTER VISION AND APPLICATIONS

Volume 1
Sensors and Imaging

Bernd Jähne
Horst Haußecker
Peter Geißler

ACADEMIC PRESS

Handbook of Computer Vision and Applications

Volume 1
Sensors and Imaging

Handbook of Computer Vision and Applications

Volume 1
Sensors and Imaging

Editors
Bernd Jähne

Interdisciplinary Center for Scientific Computing
University of Heidelberg, Heidelberg, Germany
and
Scripps Institution of Oceanography
University of California, San Diego

Horst Haußecker
Peter Geißler

Interdisciplinary Center for Scientific Computing
University of Heidelberg, Heidelberg, Germany

ACADEMIC PRESS
San Diego London Boston
New York Sydney Tokyo Toronto

ACADEMIC PRESS
A Division of Harcourt Brace & Company
525 B Street, Suite 1900, San Diego, CA 92101-4495
http://www.apnet.com

ACADEMIC PRESS
24-28 Oval Road, London NW1 7DX, UK
http://www.hbuk.co.uk/ap/

Library of Congress Cataloging-In-Publication Data
Handbook of computer vision and applications / edited by Bernd Jähne, Horst Haussecker, Peter Geissler.
p. cm.
Includes bibliographical references and indexes.
Contents: v. 1. Sensors and imaging — v. 2. Signal processing and pattern recognition — v. 3. Systems and applications.
ISBN 0–12–379770–5 (set). — ISBN 0–12–379771-3 (v. 1)
ISBN 0–12–379772–1 (v. 2). — ISBN 0–12–379773-X (v. 3)
1. Computer vision — Handbooks, manuals. etc. I. Jähne, Bernd 1953– . II. Haussecker, Horst, 1968– . III. Geissler, Peter, 1966– .
TA1634.H36 1999
006.3′7 — dc21 98–42541
CIP

Printed in the United States of America
99 00 01 02 03 DS 9 8 7 6 5 4 3 2 1

Contents

Preface

What this handbook is about

This handbook offers a fresh approach to computer vision. The whole vision process from image formation to measuring, recognition, or reacting is regarded as an integral process. Computer vision is understood as the host of techniques to acquire, process, analyze, and understand complex higher-dimensional data from our environment for scientific and technical exploration.

In this sense the handbook takes into account the interdisciplinary nature of computer vision with its links to virtually all natural sciences and attempts to bridge two important gaps. The first is between modern physical sciences and the many novel techniques to acquire images. The second is between basic research and applications. When a reader with a background in one of the fields related to computer vision feels he has learned something from one of the many other facets of computer vision, the handbook will have fulfilled its purpose.

The handbook comprises three volumes. The first volume, *Sensors and Imaging*, covers image formation and acquisition. The second volume, *Signal Processing and Pattern Recognition*, focuses on processing of the spatial and spatiotemporal signal acquired by imaging sensors. The third volume, *Systems and Applications*, describes how computer vision is integrated into systems and applications.

Prerequisites

It is assumed that the reader is familiar with elementary mathematical concepts commonly used in computer vision and in many other areas of natural sciences and technical disciplines. This includes the basics of set theory, matrix algebra, differential and integral equations, complex numbers, Fourier transform, probability, random variables, and graphing. Wherever possible, mathematical topics are described intuitively. In this respect it is very helpful that complex mathematical relations can often be visualized intuitively by images. For a more for-

mal treatment of the corresponding subject including proofs, suitable references are given.

How to use this handbook

The handbook has been designed to cover the different needs of its readership. First, it is suitable for *sequential reading*. In this way the reader gets an up-to-date account of the state of computer vision. It is presented in a way that makes it accessible for readers with different backgrounds. Second, the reader can look up specific topics of interest. The individual chapters are written in a self-consistent way with extensive cross-referencing to other chapters of the handbook and external references. The CD that accompanies each volume of the handbook contains the complete text of the handbook in the Adobe Acrobat portable document file format (PDF). This format can be read on all major platforms. Free Acrobat reader version 3.01 for all major computing platforms is included on the CDs. The texts are hyperlinked in multiple ways. Thus the reader can collect the information of interest with ease. Third, the reader can delve more deeply into a subject with the material on the CDs. They contain additional reference material, interactive software components, code examples, image material, and references to sources on the Internet. For more details see the readme file on the CDs.

Acknowledgments

Writing a handbook on computer vision with this breadth of topics is a major undertaking that can succeed only in a coordinated effort that involves many co-workers. Thus the editors would like to thank first all contributors who were willing to participate in this effort. Their cooperation with the constrained time schedule made it possible that the three-volume handbook could be published in such a short period following the call for contributions in December 1997. The editors are deeply grateful for the dedicated and professional work of the staff at AEON Verlag & Studio who did most of the editorial work. We also express our sincere thanks to Academic Press for the opportunity to write this handbook and for all professional advice.

Last but not least, we encourage the reader to send us any hints on errors, omissions, typing errors, or any other shortcomings of the handbook. Actual information about the handbook can be found at the editors homepage http://klimt.iwr.uni-heidelberg.de.

Heidelberg, Germany and La Jolla, California, December 1998
Bernd Jähne, Horst Haußecker, Peter Geißler

Contributors

Werner F. Albert studied medicine at the Universities of Saarland and Cologne. He obtained the degree of Doctor of Medicine in 1970 and completed the Habilitation in 1981 at the University of Saarland. Since 1983 he has been Chief Physician of the Department of Internal Medicine of the Westpfalz-Klinikum Kaiserslautern and since 1991 its Medical Director. He has been an Adjunct Professor at the University of Saarland at Homburg since 1986. His current research interests include transplantation medicine and gastroenterology.

Prof. Dr. Werner F. Albert, Medizinische Klinik III
Westpfalz-Klinikum, D-67653 Kaiserslautern, Germany

Uwe Apel received his diploma degree in Physics at the University of Gießen in 1984. From 1984 to 1987 he was engaged as a process engineer at the power semiconductor facility of Robert Bosch GmbH at Reutlingen. In 1987 he changed to the Institute for Microelectronics in Stuttgart. In 1994 he joined the image sensor design team. He has made major contributions to several pending patents in circuit design and camera system related topics.

Uwe Apel, Institute for Microelectronics, Stuttgart
Allmandring 30a, D-70569 Stuttgart, Germany
apel@www.ims-chips.de

Markus Böhm received the Dipl.-Ing. and the Dr.-Ing. degrees in electrical engineering from the Technical University, Berlin, Germany, in 1979 and 1983, respectively. In 1984/85, he was a visiting scientist with the Department of Electrical Engineering of the University of Delaware. In 1985, he joined Chronar Corporation in Princeton, New Jersey. Since 1989 he has been a Professor at the University of Siegen, Germany, where he heads the Institute for Semiconductor Electronics. His research interests focus on thin-film technology, novel imaging devices and photovoltaics. He is a co-founder of Silicon Vision GmbH.

Prof. Markus Böhm, Institut für Halbleiterelektronik (IHE)
Universität-GH Siegen, Hölderlinstr. 3, D-57068 Siegen, Germany
boehm@teb.et-inf.uni-siegen.de, www.uni-siegen.de/~ihe/

Michael J. Buckingham is Professor of Ocean Acoustics at Scripps Institution of Oceanography. His research interests include imaging in the ocean, marine sediments, and sea-surface processes. He is a Fellow of the Acoustical Society of America, the Institute of Acoustics, the Institute of Electrical Engineers, and the Explorers Club, and a member of the New York Academy of Sciences. In 1982 he received the A. B. Wood Medal from the IOA and he is the recipient of many other awards.

Prof. Michael J. Buckingham
Marine Physical Laboratory
Scripps Institution of Oceanography
University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093-0213, USA, mjb@mpl.ucsd.edu,

Gunnar Brix studied physics in Karlsruhe and Heidelberg. In 1985 he received his diploma degree from the University of Karlsruhe and in 1989 a doctoral degree from the University of Heidelberg. From 1994 to 1998 he was assistant professor for medical physics at the University of Heidelberg where he headed the department of biophysics and medical radiation physics at the German Cancer Research Center in Heidelberg. His current research interests include the development of new magnetic resonance imaging (MRI) and positron emission tomography (PET) data acquisition techniques as well as the analysis of kinetic data within the framework of kinetic modeling.

Priv.-Doz. Dr. Gunnar Brix
Abteilung Medizinische Strahlenhygiene und nichtionisierende Strahlung
Bundesamt für Strahlenschutz, Postfach 10 01 49
D-38201 Salzgitter, Germany

Paolo Fiorini took his degree in Solid State Physics at the University of Rome in 1977; his thesis was on exitons in silicon. He has been active in the field of electrical and optical properties of semiconductors for many years, working at the University of Rome, Strasbourg (France), IBM Research Center in Yorktown Heights, NY (USA) and at the Interuniversity Microelectronic Center (IMEC) in Leuven (Belgium). At present, he is associate professor, Physics Department of the Third University of Rome.

Prof. Paolo Fiorini, Dept. of Physics
3rd University of Rome, Via della Vasca Navale 86, I-00156 Rome, Italy

Rainer H.A. Fink is a professor at the II. Institute of Physiology at the University of Heidelberg. His research interests comprise calcium regulation, activation of contractile force, membrane electrophysiology, and laser applications in the biophysics of muscular contraction. He held research and teaching positions at the University of Washington, Seattle, WA, U.S., La Trobe University, Melbourne, and the University of Adelaide, Australia, before taking up his professorship in Heidelberg in 1990. He received his PhD in 1979 at the University of Bochum, Germany.

Prof. Dr. Rainer H.A. Fink, II. Physiologisches Institut
Universität Heidelberg, Im Neuenheimer Feld 326
D-69120 Heidelberg, Germany
fink@novsrv1.pio1.uni-heidelberg.de

Peter Geißler studied physics in Heidelberg. He received his diploma and doctoral degree from Heidelberg University in 1994 and 1998, respectively. His research interests include computer vision, especially depth-from-focus, adaptive filtering, and flow visualization as well as the application of image processing in physical sciences and oceanography.

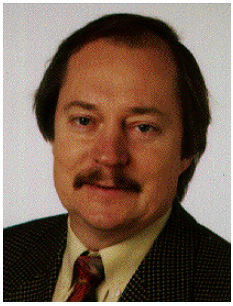
Dr. Peter Geißler
Forschungsgruppe Bildverarbeitung, IWR
Universität Heidelberg, Im Neuenheimer Feld 368
D-69120 Heidelberg, Germany
Peter.Geissler@iwr.uni-heidelberg.de
http://klimt.iwr.uni-heidelberg.de

Robert Godding received his diploma in geodesy from the University of Bonn in 1987. From 1987 to 1989 he worked as research scientist at the Institute for Photogrammetry at the University of Bonn and from 1989 to 1994 at the Institute for Photogrammetry and Image Processing at the University of Braunschweig in the field of close-range photogrammetry. From 1994 to 1998 he was with Rollei Fototechnic in Braunschweig, first responsible for research and development of close-range photogrammetry systems, later as head of the RolleiMetric Department. Since December 1998 he has been with AICON GmbH in Braunschweig. His main interests are close range-digital photogrammetry in industrial applications and the use of high-resolution digital cameras for optical measurement techniques.

Robert Godding
AICON GmbH, Celler Straße 32, D-38114 Braunschweig, Germany
robert.godding@aicon.de, http://www.aicon.de

Hermann Gröning graduated in 1996 from the University of Heidelberg with a master degree in physics and is now pursuing his PhD at the Interdisciplinary Center for Scientific Computing. He is concerned mainly with radiometric and geometric camera calibration.

Hermann Gröning
Forschungsgruppe Bildverarbeitung, IWR
Universität Heidelberg
Im Neuenheimer Feld 368
D-69120 Heidelberg, Germany
Hermann.Groening@iwr.uni-heidelberg.de



Axel Haase studied physics at the universities of Erlangen and Gießen. He received his diploma from the University of Gießen in 1977 and a doctoral degree in 1980. During his doctoral work and later in postdoctoral work, he worked at the Max-Planck-Institut für biophysikalische Chemie in Göttingen. In 1981 he spent one postdoctoral year at the Biochemistry Department of the University of Oxford, UK, with Prof. G. K. Radda. He worked as a scientist at the Max-Planck-Institut für biophysikalische Chemie until 1989. During this period he invented fast NMR imaging (FLASH) and other NMR techniques (CHESS imaging, STEAM imaging). He received his habilitation from the University of Frankfurt in 1987. Since 1989, he has held the chair of biophysics at the University of Würzburg.

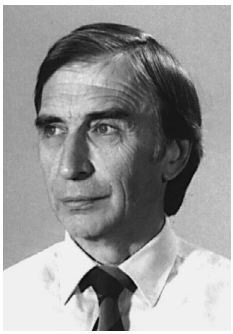
Prof. Dr. Axel Haase, Physikalisches Institut, Universität Würzburg
Am Hubland, D-97074 Würzburg, Germany
haase@physik.uni-wuerzburg.de

Gerd Häusler is adjunct professor, University of Erlangen, Chair for Optics, and director of the Optical Metrology Group. He received his diploma in 1970 and a doctoral degree in 1974 from the Optical Institute, Technical University Berlin. In 1974 he moved to the Chair for Applied Optics (later Chair for Optics), University of Erlangen. There he received his habilitation in 1982. As a doctoral fellow he worked with IBM (Sindelfingen), ENST Telecom (Paris), and RCA (Zürich). At the University of Munich and the RIKEN Institute in Tokyo he worked on optical and electronical image processing and nonlinear optical feedback systems. His current research interests include the investigation of the physical limits of range sensing and the construction of sensors that work at these limits and cover the nanometer to meter range, with applications in industry and medicine.

Prof. Dr. Gerd Häusler, Chair for Optics, Universität Erlangen-Nürnberg
Staudtstraße 7/B2, D-91056 Erlangen, Germany
haeusler@physik.uni-erlangen.de

Horst Haußecker studied physics in Heidelberg. He received his diploma in physics and his doctoral degree from Heidelberg University in 1994 and 1996, respectively. He was visiting scientist at the Scripps Institution of Oceanography in 1994. Currently he is conducting research in the image processing research group at the Interdisciplinary Center for Scientific Computing (IWR), where he also lectures on optical flow computation. His research interests include computer vision, especially image sequence analysis, infrared thermography, and fuzzy-image processing, as well as the application of image processing in physical sciences and oceanography.

Dr. Horst Haußecker, Forschungsgruppe Bildverarbeitung, IWR
Universität Heidelberg, Im Neuenheimer Feld 368, D-69120 Heidelberg
Horst.Haussecker@iwr.uni-heidelberg.de
http://klimt.iwr.uni-heidelberg.de



Bernd Höfflinger received his Diploma in Physics in 1964 and his PhD in 1967. He was a member of the scientific staff of the Siemens Research Laboratory in Munich from 1964-1967. From 1967-1969 he was Assistant Professor, School of Electrical Engineering, Cornell University, Ithaca, NY, USA. He was manager of the MOS Integrated Circuits Division of the Siemens Components Group in Munich from 1970-1972. He then founded the Department of Electrical Engineering at the University of Dortmund. In 1981 he became Head of the Department of Electrical Engineering and Co-Director of the Microelectronics and Information Sciences Center at the University of Minnesota. Since September 1985 he has been Director and Chairman of the Board of the Institute for Microelectronics, Stuttgart.

Prof. Dr. Bernd Höfflinger, Institute for Microelectronics Stuttgart (IMS)
Allmandring 30a, D-70569 Stuttgart, Germany
E-mail: hoefflinger@www.ims-chips.de

Bernd Jähne studied physics in Saarbrücken and Heidelberg. He received his diploma, doctoral degree, and habilitation degree from Heidelberg University in 1977, 1980, and 1985, respectively, and a habilitation degree in applied computer science from the University of Hamburg-Harburg in 1992. Since 1988 he has been a Marine Research Physicist at Scripps Institution of Oceanography, University of California, and, since 1994, he has been professor of physics at the Interdisciplinary Center of Scientific Computing. He leads the research group on image processing. His research interests include computer vision, especially filter design and image sequence analysis, the application of image processing techniques in science and industry, and small-scale air-sea interaction processes.

Prof. Dr. Bernd Jähne, Forschungsgruppe Bildverarbeitung, IWR
Universität Heidelberg, Im Neuenheimer Feld 368, D-69120 Heidelberg
Bernd.Jaehne@iwr.uni-heidelberg.de
http://klimt.iwr.uni-heidelberg.de

Reinhard Malz studied communication and computer science in Esslingen and electrical engineering in Stuttgart. He received diploma degrees in 1978 and 1984 and the doctoral degree from University of Stuttgart in 1992. His research interests include analog and digital electronics, semiconductor physics, optics, pattern recognition, and active optical information processing for inspection and measurement. Currently he is a researcher at Daimler-Chrysler AG, where he develops 3-D measurement systems for reverse engineering and quality control.

Dr. Reinhard Malz, Daimler-Chrysler AG
Wilhelm-Runge-Str. 11, D-89081 Ulm, Germany, Reinhard.Malz@t-online.de

Robert Massen studied electronic communications at the University of Technology of Aachen, Germany. His PhD thesis covers stochastic computing, an early non-von Neumann computer architecture with random data coding and massively parallel organization. In 1974, he became professor in the Department of Computer Science at the Fachhochschule (University of Applied Sciences) Konstanz. He has been active in industrial image processing since 1984, first as director of the for-profit Steinbeis Transfer Center for Image Processing. In 1992, he founded the MASSEN machine vision systems GmbH, Konstanz, through a management buy-out. The company is a major German supplier of advanced dedicated color vision systems for on-line monitoring of surfaces and for real-time sortation.

Prof. Dr. Ing. Robert Massen, MASSEN machine vision systems GmbH
Lohnerhofstrasse 2, D-78467 Konstanz, Germany, Robert.Massen@t-online.de

Madhukar Pandit studied electrical engineering in Bangalore and Karlsruhe. He obtained the Dr.-Ing. degree in Control Systems in the Technische Hochschule Karlsruhe and the Habilitation in the Kaiserslautern University. He worked at the National Aeronautical Laboratory, Bangalore, Brown Boveri and Cie in Mannheim. Since 1978, he has been professor of Control Systems and Signal Theory at the Kaiserslautern University. His group is active mainly in the areas of process control and image processing applied to medical imaging and quality control.

Prof. Dr.-Ing. Madhukar Pandit, Lehrstuhl für Regelungstechnik und Signaltheorie, Fachbereich Elektrotechnik, Universität Kaiserslautern
Postfach 3049, D-67653 Kaiserslautern, Germany
Pandit@e-technik.uni-kl.de, http://www.uni-kl.de/AG-Pandit/

Peter Rieve received the Dipl.-Ing. degree in electrical engineering from the University of Siegen, Germany, in 1994. From 1994 to 1997 he was a research engineer at the Institute for Semiconductor Electronics, University of Siegen. He worked in the field of sensor technologies and focused on the development and optimization of amorphous silicon based black and white and color detectors for applications in image sensor systems in TFA technology. P. Rieve is now with Silicon Vision GmbH, Siegen.

Peter Rieve, Silicon Vision GmbH
Birlenbacher Str. 18, D-57078 Siegen, Germany
rieve@siliconvision.de, http://www.siliconvision.de

Markus Rokitta studied physics at the University of Würzburg. He received his diploma from the University of Würzburg in 1994. Since 1996 he has been working for his doctoral degree in the area of NMR microscopy applied to plant systems. He is member of the Graduiertenkolleg "Magnetische Kernresonanz in vivo und in vitro für die biologische und medizinische Grundlagenforschung."

Dipl. Phys. Markus Rokitta
Physikalisches Institut
Universität Würzburg
Am Hubland, D-97074 Würzburg, Germany

Jan Ruff studied physics at the University of Würzburg. He received his diploma from the University of Würzburg in 1995. Since 1996 he has been working for his doctoral degree in the area of NMR microscopy applied to animal studies. He is member of the Graduiertenkolleg "Magnetische Kernresonanz in vivo und in vitro für die biologische und medizinische Grundlagenforschung."

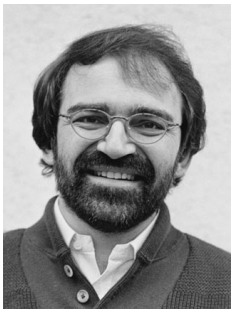
Dipl. Phys. Jan Ruff
Physikalisches Institut
Universität Würzburg
Am Hubland, D-97074 Würzburg, Germany

Bernd Schneider received the Dipl.-Ing. degree in electrical engineering from the University of Siegen, Germany, in 1995. In the same year, he joined the Institute for Semiconductor Electronics at the University of Siegen. He works in the field of sensor technologies and focuses on the design, fabrication and characterization of ASICs for TFA image sensors. He is currently engaged in the development of new types of TFA sensor systems.

Bernd Schneider, Institut für Halbleiterelektronik (IHE)
Universität-GH Siegen
Hölderlinstr. 3, D-57068 Siegen, Germany
bernd_s@teb.et-inf.uni-siegen.de
http://www.uni-siegen.de/~ihe/

Wolfgang Schreiber studied physics in Munich. He received his diploma from the University of Munich in 1990 and in 1994 a doctoral degree from the University of Heidelberg. From 1994 to 1997 he was a postdoctoral fellow at the department of biophysics and medical radiation physics at the German Cancer Research Center in Heidelberg. Since 1997 he has been head of the research group MR physics at the University of Mainz. His current research interests include the development of techniques for noninvasive assessment of physiology and pathophysiology by magnetic resonance imaging, pulse sequence development for ultrafast proton- and non-proton MRI in the brain, heart, and lung, and MRI data postprocessing techniques.

Dr. Wolfgang Schreiber, Department of Radiology
Johannes Gutenberg-University, Langenbeckstr. 1
D-55131 Mainz, Germany
Wolfgang.Schreiber@radiologie.klinik.uni-mainz.de



Rasmus R. Schröder studied physics and biology at the Ruprechts-Karls-Universität Heidelberg, Germany and at the Trinity College Dublin, Ireland. After receiving his PhD at Heidelberg University in theoretical elementary particle physics, he took up his biological interests by continuing his work in the department of Kenneth C. Holmes at the Max-Planck-Institut for medical research in Heidelberg. Here he started his work in electron microscopy and image processing. Today he is group leader at the Max-Planck-Institut doing research on the image formation in energy filtered electron microscopes, quantitative image processing, and investigating molecular structures of the muscle proteins actin and myosin in its force-producing interactions.

Dr. Rasmus R. Schröder, MPI für medizinische Forschung
Jahnstr. 29, D-69120 Heidelberg, Germany
rasmus@mpimf-heidelberg.mpg.de

Rudolf Schwarte studied electrical engineering at the RWTH Aachen. He received his diploma and doctoral degree from RWTH Aachen in 1965 and 1972. From 1973-1978 he worked as the head research engineer at the Institute for Technical Electronics in Aachen, founded the company Sympuls GmbH in Aachen, followed by three years of leading research departments in several companies in Germany. Since 1981 he has been professor at the University of Siegen and head of the Institute for Data Processing (INV). He is the initiator and chair of the Center for Sensory Systems (ZESS) in Siegen. In 1995 he received the NRW innovation prize. He holds several patents in the fields of optical measurement and communication systems. In 1997 he founded S-TEC GmbH in Siegen. His main research interests include laser ranging systems, optical sensory systems, optical data communication, and digital signal processing.

Prof. Dr. Rudolf Schwarte, Institut für Nachrichtenverarbeitung (INV)
Universität-GH Siegen, Hölderlinstr. 3, D-57068 Siegen, Germany
schwarte@nv.et-inf.uni-siegen.de
http://www.nv.et-inf.uni-siegen.de/inv/inv.html

Ulrich Seger received his diploma in electrical engineering from the Fachhochschule Konstanz for his work on digital image preprocessing for optical character recognition in 1987. As design engineer in Computer Gesellschaft Konstanz mbHs R&D-department he was engaged in the development of a multiprocessor character recognition system. In 1989 he joint the Mikrosystems Division of the IMS, where he worked on chip and system design of CMOS microsystems involving optical sensors and analog signal processors and started the development of the first HDRC sensors. He is co-inventor of the basic HDRC principle and made major contributions to several pending patents in circuit design and camera system related topics.

Ulrich Seger, Institute for Microelectronics Stuttgart
Allmandring 30a, D-70569 Stuttgart, Germany
seger@www.ims-chips.de

Peter Seitz received his PhD degree in physics in 1984 from the Swiss Federal Institute of Technology (ETH) in Zürich, Switzerland. From 1984 to 1987 he was a staff member of the RCA research laboratories in Princeton, New Jersey and Zürich, Switzerland. Afterwards he transferred to the Swiss Paul Scherrer Institute. Since 1997 he has been working for the Swiss Center for Electronics and Microtechnology (CSEM) in Neuchatel and Zürich, heading the Image Sensing Section in the Research division. Peter Seitz is the author of 90 publications in the fields of applied optics, image sensing, machine vision, and optical microsystems engineering, and he holds 8 patents.

Prof. Dr. Peter Seitz
Centre Suisse d'Electronique et de Microtechnique SA (CSEM)
Badenerstrasse 569, CH-8048 Zurich, Switzerland
peter.seitz@csem.ch, http://www.csem.ch/

Sherif Sedky graduated in 1992 from the department of Electrical and Communication Engineering of Cairo University. In 1995 he obtained a master degree in Engineering Physics at the same university. In 1998 he was granted a PhD degree in micro electronics and material science from the Catholic University of Leuven (Belgium). He is active in the field of Sensors and Actuators. He is now a member of the microsystem technology group of the Interuniversity Microelectronics Center (IMEC) in Leuven (Belgium). He is also an assistant professor at the department of Engineering Physics, Faculty of Engineering, Cairo University.

Dr. Sherif Sedky
Department of Engineering, Mathematics and Physics
Faculty of Engineering, Cairo University, Giza, Egypt, sedky@imec.be

E. H. K. Stelzer studied physics in Frankfurt am Main and in Heidelberg, Germany. During his Diploma thesis at the Max-Planck-Institut für Biophysik he worked on the physical chemistry of phospholipid vesicles, which he characterized by photon correlation spectroscopy. Since 1983 he has worked at the European Molecular Biology Laboratory (EMBL). He has contributed extensively to the development of confocal fluorescence microscopy and its application in life sciences. His group works on the development and application of high-resolution techniques in light microscopy, video microscopy, confocal microscopy, optical tweezers, single particle analysis, and the documentation of relevant parameters with biological data.

Prof. Dr. E. H. K. Stelzer, Light Microscopy Group,
European Molecular Biology Laboratory (EMBL), Postfach 10 22 09
D-69120 Heidelberg, Germany, stelzer@EMBL-Heidelberg.de,

Heiko Stegmann studied physics at the Ruprecht-Karls-Universität Heidelberg, Germany. He received his diploma degree in 1996 and his PhD degree in 1998 from that university, working on analytical electron microscopy techniques for the investigation of muscle biophysics. At present he works on 3-D reconstruction of motor molecules by cryo-electron microscopy at the Max-Planck-Institut for medical research in Heidelberg.

Heiko Stegmann, MPI für medizinische Forschung
Jahnstr. 29, D-69120 Heidelberg, Germany
stegmann@mpimf-heidelberg.de

Dietmar Uttenweiler is a research fellow at the II. Institute of Physiology at the University of Heidelberg in the group of Prof. Dr. R. H. A. Fink. He studied physics in Freiburg and Heidelberg. In 1990–1991 he worked at the University of Sussex, UK, supported by an Erasmus scholarship. He graduated as Diplom-Physiker in 1994 and received his doctoral degree (Dr. rer. nat.) in physics in 1997 from the University of Heidelberg. His research interests in biophysics comprise fluorescence imaging techniques, mathematical modeling, and digital image processing, in particular for the study of motor proteins and the calcium regulation of force generation in muscle.

Dr. Dietmar Uttenweiler, II. Physiologisches Institut
University of Heidelberg, Im Neuenheimer Feld 326, D-69120 Heidelberg
dietmar.uttenweiler@urz.uni-heidelberg.de

Roger Wepf studied biology at the ETH, Swiss Federal Institute of Technology in Zurich, Switzerland, received his PhD at the Institute of Cell Biology, Swiss Federal Institute of Technology on surface imaging with high resolution coating and worked as a postdoctoral fellow at the same institute on imaging of actin binding proteins, and at the EMBL, Heidelberg, Germany on new preparation techniques for high-resolution LVSEM in the group of Max Haider. He then became a staff member in the junior-group Cryopreparation for EM and Cryo-SEM in the Cell Biology dept. at EMBL. At present he is the group leader of Electron Microscopy in the Central Analytical Dept. at Beiersdorf AG, Hamburg, Germany.

Dr. Roger Wepf, Beiersdorf AG
Unnastr. 48, D-20245 Hamburg, Germany
wepf-r.ocp-65@bdfde86mhs.compuserve.com

1 Introduction

Bernd Jähne

Interdisziplinäres Zentrum für Wissenschaftliches Rechnen (IWR)
Universität Heidelberg, Germany

1.1 Components of a vision system

Computer vision is a complex subject. As such it is helpful to divide it into the various components or function modules. On this level, it is also much easier to compare a technical system with a biological system. In this sense, the basic common functionality of biological and machine imaging includes the following components:

Radiation source. If no radiation is emitted from the scene or the object of interest, nothing can be observed or processed. Thus appropriate illumination is necessary for objects that are themselves not radiant.

Camera. The "camera" collects the radiation received from the object in such a way that the radiation's origins can be pinpointed. In the simplest case this is just an optical lens. But it could also be a completely different system, for example, an imaging optical spectrometer, an x-ray tomograph, or a microwave dish.

Sensor. The sensor converts the received radiative flux density into a suitable signal for further processing. For an imaging system normally a 2-D array of sensors is required to capture the spatial distribution of the radiation. With an appropriate scanning system in some cases a single sensor or a row of sensors could be sufficient.

Processing unit. It processes the incoming, generally higher-dimensional data, extracting suitable features that can be used to measure object properties and categorize them into classes. Another important component is a memory system to collect and store knowledge about the scene, including mechanisms to delete unimportant things.

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0–12–379771–3/$30.00

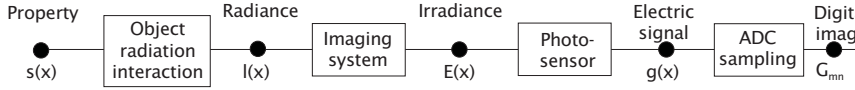


Figure 1.1: *Chain of steps linking an object property to the signal measured by an imaging system.*

Actors. Actors react to the result of the visual observation. They become an integral part of the vision system when the vision system is actively responding to the observation by, for example, *tracking* an object of interest or by using a vision-guided navigation (*active vision*, *perception action cycle*).

1.2 Imaging systems

Volume 1 of this handbook deals with imaging systems. It covers all processes involved in the formation of an image from objects and the sensors that convert radiation into electric signals. Generally the goal is to attain a signal from an object in such a form that we know where it is (geometry) and what it is or what properties it has.

It is important to note that the type of answer we receive from these two implicit questions depends on the purpose of the vision system. The answer could be of qualitative or quantitative nature. For some applications it could be sufficient to obtain a qualitative answer like "there is a car on the left coming towards you." The "what" and "where" questions can thus cover the entire range from "there is something," a specification of the object in the form of a class, to a detailed quantitative description of various properties of the objects of interest.

The relation that links the object property to the signal measured by an imaging system is a complex chain of processes (Fig. 1.1). Interaction of the radiation with the object (possibly using an appropriate illumination system) causes the object to emit radiation. A portion (usually only a very small part) of the emitted radiative energy is collected by the optical system and perceived as an *irradiance* (radiative energy/area). A sensor (or rather an array of sensors) converts the received radiation into an electrical signal that is subsequently sampled and digitized to form a digital image as an array of digital numbers.

Only *direct imaging* systems provide a direct point to point correspondence between points of the objects in the 3-D world and at the image plane. *Indirect imaging* systems also give a spatially distributed irradiance but with no such one-to-one relation. Generation of an image requires reconstruction of the object from the perceived irradiance. Examples of such imaging techniques include radar imaging, various techniques for spectral imaging, acoustic imaging, tomographic imaging, and magnetic resonance imaging (Chapters 22 and 23).

The first part of this volume covers the basics of image formation (Chapters 2–6). The fundamentals of electromagnetic radiation, radiometry and photometry, and of thermal radiation are discussed in Chapter 2. Chapter 4 discusses basis knowledge regarding optics and optical systems, areas that are helpful to know for computer vision. Chapter 3 deals with the basic physical laws that determine the relation between object properties and the emitted radiation while Chapter 5 deals with the basic relations between the emitted radiation (radiance) and the received radiation at the sensor plane (irradiance). Chapter 6 covers two practical topics. First, it introduces various types of illumination sources that are available to illuminate a scene. Second, it describes the basic possibilities for illumination setups and their relation to the imaged object properties.

The second part of this volume covers imaging sensors. It starts with an survey of solid state imaging (Chapter 7) and then details some important recent developments including logarithmic complementary metal-oxide-semiconductor (CMOS) sensors for natural vision perception (Chapter 8), a novel family of vision sensors built as thin films on top of application specific circuits (Chapter 9), and a chapter on modern developments with uncooled infrared imaging sensors (Chapter 10). The second part concludes with a chapter on the principles of color and spectral imaging (Chapter 11).

The third and fourth parts present in detail various 2-D (Chapters 12–16) and 3-D (Chapters 17–23) imaging systems, respectively. The part on 2-D imaging discusses fluorescence imaging (Chapter 12), electron microscopic imaging (Chapter 13), acoustic imaging (Chapters 14 and 15), and multisensorial cameras for industrial vision applications (Chapter 16).

Techniques for 3-D imaging have experienced an enormous progress in the last several years. While traditional computer vision is only concerned with classical paradigms such as *structure from stereo*, *shape from shading*, *depth from focus*, or *structure from motion* (see Volume 2), recent advances in sensor technology have advanced a host of techniques for 3-D imaging. This is the topic of part IV of Volume 1. Reconstruction of 3-D geometry from images requires careful geometrical calibration (Chapter 17). Chapter 18 surveys the principles of 3-D imaging and shows that the wide variety of available techniques can be categorized into a scheme with only a few basic principles. Chapter 19 focuses on the physical principles that ultimately limit the accuracy of 3-D imaging and explores some new techniques such as optical coherence tomography while Chapter 20 discusses high-performance surface measuring by combining photogrammetric and sequential-light techniques. The remainder of part IV deals with 3-D light microscopy (Chapter 21) and magnetic resonance (MR) imaging in medical and biological research (Chapters 22 and 23).

Part I

Illumination and Image Formation

2 Radiation

Horst Haußecker

Interdisziplinäres Zentrum für Wissenschaftliches Rechnen (IWR)
Universität Heidelberg, Germany

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

2.1 Introduction

Visual perception of scenes depends on appropriate illumination to visualize objects. The human visual system is limited to a very narrow portion of the spectrum of electromagnetic radiation, called *light*. In some cases natural sources, such as solar radiation, moonlight, lightning flashes, or bioluminescence, provide sufficient ambient light to navigate our environment. Because humankind was restricted mainly to daylight one of the first attempts was to invent an artificial light source, fire (not only as a food preparation method).

Computer vision is not dependent upon visual radiation, fire, or glowing objects to illuminate scenes. As soon as imaging detector systems became available other types of radiation were used to probe scenes and objects of interest. Recent developments in imaging sensors cover almost the whole electromagnetic spectrum from x-rays to radiowaves (Chapters 7–11). In standard computer vision applications illumination is frequently taken as given and optimized to illuminate objects evenly with high contrast. Such setups are appropriate for object identification and geometric measurements. Radiation, however, can also be used to visualize quantitatively physical properties of objects by analyzing their interaction with radiation (Chapter 3).

Physical quantities such as penetration depth or surface reflectivity are essential to probe the internal structures of objects, scene geometry, and surface-related properties. The properties of physical objects therefore can be encoded not only in the geometrical distribution of emitted radiation but also in the portion of radiation that is emitted, scattered, absorbed, or reflected, and finally reaches the imaging system. Most of these processes are sensitive to certain wavelengths and additional information might be hidden in the spectral distribution of radiation. Using different types of radiation allows taking images from different depths or different object properties. As an example, infrared radiation of between 3 and 5 μm is absorbed by human skin to a depth of < 1 mm, while x-rays penetrate an entire body without major attenuation. Therefore, totally different properties of the human body (such as skin temperature as well as skeletal structures) can be revealed for medical diagnosis.

This chapter provides the fundamentals for a quantitative description of radiation emitted from sources. The interaction of radiation with objects and matter is the subject of Chapter 3. *Radiometry*, the measurement of radiation properties by imaging systems, will be detailed in Chapter 5. Although the theory will be introduced in a general way for all types of radiation, a large portion of this chapter is dedicated to the two spectral ranges of visible and infrared (IR) radiation. While visible radiation plays the most important role in computer vision, the latter has been gaining in importance due to recent performance improvements in infrared imaging technology (see Chapter 10).

2.2 Fundamentals of electromagnetic radiation

2.2.1 Electromagnetic waves

Electromagnetic radiation consists of *electromagnetic waves* carrying energy and propagating through space. Electrical and magnetic fields are alternating with a temporal *frequency* ν and a spatial *wavelength* λ. The metric units of ν and λ are cycles per second (s^{-1}), and meter (m), respectively. The unit $1\,s^{-1}$ is also called one hertz (1 Hz). Wavelength and frequency of waves are related by the *speed of light* c:

$$c = \nu\lambda \tag{2.1}$$

The speed of light depends on the medium through which the electromagnetic wave is propagating. In vacuum, the speed of light has the value $2.9979 \times 10^8\,m\,s^{-1}$, which is one of the fundamental physical constants and constitutes the maximum possible speed of any object. The speed of light decreases as it penetrates matter, with slowdown dependent upon the electromagnetic properties of the medium (see Section 3.3.2).

Photon energy. In addition to electromagnetic theory, radiation can be treated as a flow of particles, discrete packets of energy called *photons*. One photon travels at the speed of light c and carries the energy

$$e_p = h\nu = \frac{hc}{\lambda} \tag{2.2}$$

where $h = 6.626 \times 10^{-34}\,J\,s$ is Planck's constant. Therefore the energy content of radiation is quantized and can only be a multiple of $h\nu$ for a certain frequency ν. While the energy per photon is given by Eq. (2.2), the total energy of radiation is given by the number of photons. It was this quantization of radiation that gave birth to the theory of quantum mechanics at the beginning of the twentieth century.

The energy of a single photon is usually given in *electron volts* (1 eV = 1.602×10^{-19}). One eV constitutes the energy of an electron being accelerated in an electrical field with a potential difference of one volt. Although photons do not carry electrical charge this unit is useful in radiometry, as electromagnetic radiation is usually detected by interaction of radiation with electrical charges in sensors (Chapter 7). In solid-state sensors, for example, the energy of absorbed photons is used to lift electrons from the valence band into the conduction band of a semiconductor. The bandgap energy E_g defines the minimum photon energy required for this process. As a rule of thumb the detector material is sensitive to radiation with energies $E_\nu > E_g$. As an example, *indium antimonide* (*InSb*) is a doped semiconductor with a bandgap of only 0.18 eV. It is sensitive to wavelengths below 6.9 μm (which can be

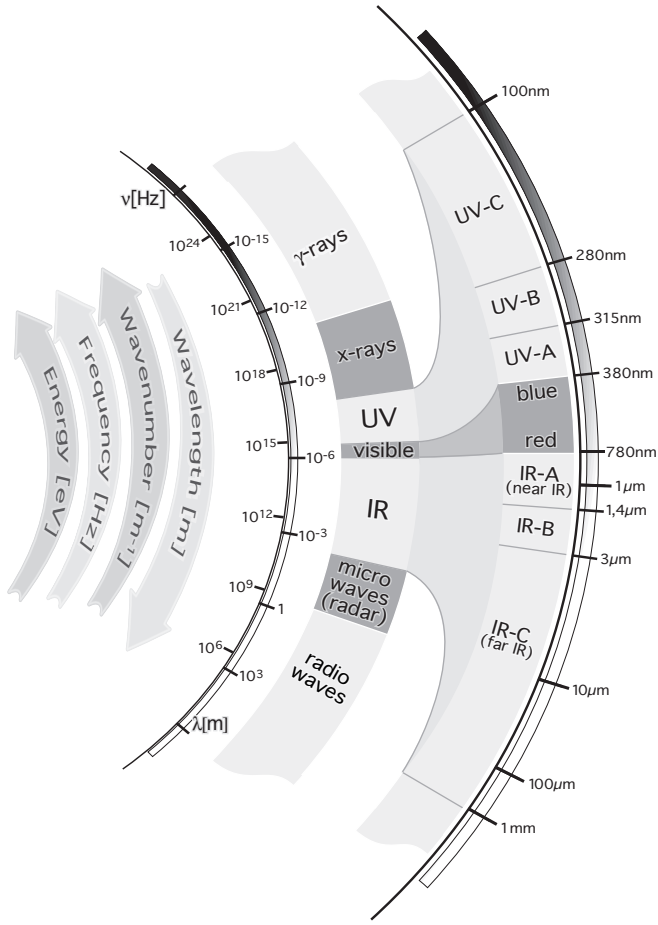


Figure 2.1: *Spectrum of electromagnetic radiation. (By Sven Mann, University of Heidelberg.)*

derived from Eq. (2.2)). *Silicon* (Si) has a bandgap of 1.1 eV and requires wavelengths below 1.1 μm to be detected. This shows why InSb can be used as detector material for infrared cameras in the 3-5 μm wavelength region, while silicon sensors are used for visible radiation. It also shows, however, that the sensitivity of standard silicon sensors extends beyond the visible range up to approximately 1 μm, which is often neglected in applications.

Electromagnetic spectrum. Monochromatic radiation consists of only one frequency and wavelength. The distribution of radiation over the range of possible wavelengths is called *spectrum* or *spectral distribution*. Figure 2.1 shows the spectrum of electromagnetic radiation to-

gether with the standardized terminology[1] separating different parts. Electromagnetic radiation covers the whole range from very high energy cosmic rays with wavelengths in the order of 10^{-16} m ($\nu = 10^{24}$ Hz) to sound frequencies above wavelengths of 10^{6} m ($\nu = 10^{2}$ Hz). Only a very narrow band of radiation between 380 and 780 nm is visible to the human eye.

Each portion of the electromagnetic spectrum obeys the same principal physical laws. Radiation of different wavelengths, however, appears to have different properties in terms of interaction with matter and detectability that can be used for wavelength selective detectors. For the last 100 yr detectors have been developed for radiation of almost any region of the electromagnetic spectrum. Recent developments in detector technology incorporate point sensors into integrated detector arrays, which allows setting up imaging radiometers instead of point measuring devices. Quantitative measurements of the spatial distribution of radiometric properties are now available for remote sensing at almost any wavelength.

2.2.2 Dispersion and attenuation

A mixture of radiation consisting of different wavelengths is subject to different speeds of light within the medium it is propagating. This fact is the basic reason for optical phenomena such as *refraction* and *dispersion*. While refraction changes the propagation direction of a beam of radiation passing the interface between two media with different optical properties, dispersion separates radiation of different wavelengths (Section 3.3.2).

2.2.3 Polarization of radiation

In electromagnetic theory, radiation is described as oscillating electric and magnetic fields, denoted by the electric field strength $\boldsymbol{E}$ and the magnetic field strength $\boldsymbol{B}$, respectively. Both vector fields are given by the solution of a set of differential equations, referred to as *Maxwell's equations*.

In free space, that is, without electric sources and currents, a special solution is a *harmonic planar wave*, propagating linearly in space and time. As Maxwell's equations are linear equations, the superposition of two solutions also yields a solution. This fact is commonly referred to as the *superposition principle*.

The superposition principle allows us to explain the phenomenon of *polarization*, another important property of electromagnetic radiation. In general, the 3-D orientation of vector $\boldsymbol{E}$ changes over time and

[1] International Commission on Illumination (Commission Internationale de l'Eclairage, CIE); http://www.cie.co.at/cie

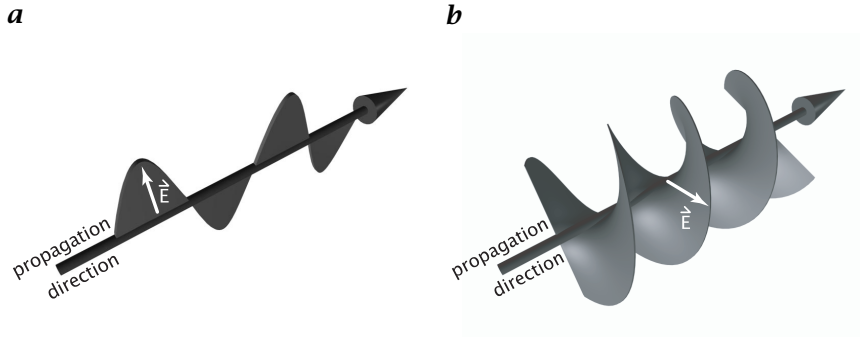


Figure 2.2: *Illustration of* ***a*** *linear and* ***b*** *circular polarization of electromagnetic radiation. (By C. Garbe, University of Heidelberg.)*

mixtures of electromagnetic waves show randomly distributed orientation directions of $\boldsymbol{E}$. If, however, the electromagnetic field vector $\boldsymbol{E}$ is confined to a plane, the radiation is called *linearly polarized* (Fig. 2.2a).

If two linearly polarized electromagnetic waves are traveling in the same direction, the resulting electric field vector is given by $\boldsymbol{E} = \boldsymbol{E}_1 + \boldsymbol{E}_2$. Depending on the phase shift Φ in the oscillations of $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$, the net electric field vector $\boldsymbol{E}$ remains linearly polarized ($\Phi = 0$), or rotates around the propagation direction of the wave. For a phase shift of $\Phi = 90°$, the wave is called *circularly polarized* (Fig. 2.2b). The general case consists of *elliptical polarization*, that is, mixtures between both cases.

Due to polarization, radiation exhibits different properties in different directions, such as, for example, directional reflectivity or polarization dependent transmissivity.

2.2.4 Coherence of radiation

Mixtures of electromagnetic waves, which are emitted from conventional light sources, do not show any spatial and temporal relation. The phase shifts between the electric field vectors $\boldsymbol{E}$ and the corresponding orientations are randomly distributed. Such radiation is called *incoherent*.

Special types of light sources, mainly those operating by stimulated emission of radiation (e. g., lasers), emit radiation with a fixed systematic relationship between the phases of the electromagnetic field vectors, a property called *coherence*.

Such radiation can be subject to constructive and destructive interference if it is superposed. As the electric field vectors can add up to high amplitudes, the local energy impact of coherent radiation is much more severe and can cause damage to delicate body tissue.

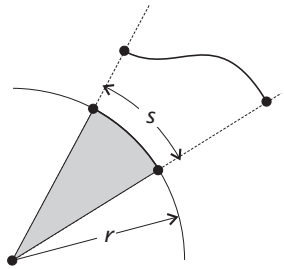


Figure 2.3: *Definition of plane angle.*

2.3 Radiometric quantities

2.3.1 Solid angle

In order to quantify the geometric spreading of radiation leaving a source, it is useful to recall the definition of solid angle. It extends the concept of plane angle into 3-D space. A *plane angle* θ is defined as the ratio of the arc length s on a circle to the radius r centered at the point of definition:

$$\theta = \frac{s}{r} \tag{2.3}$$

The arc length s can be considered as projection of an arbitrary line in the plane onto the circle (Fig. 2.3). Plane angles are measured in rad (radians). A plane angle θ quantifies the angular subtense of a line segment in the plane viewed from the point of definition. A circle has a circumference of $2\pi r$ and, therefore, subtends a plane angle of 2π rad.

A *solid angle* ω is similarly defined as the ratio of an area A on the surface of a sphere to the square radius, as shown in Fig. 2.4:

$$\Omega = \frac{A}{r^2} \tag{2.4}$$

The area segment A can be considered as the projection of an arbitrarily shaped area in 3-D space onto the surface of a sphere. Solid angles are measured in sr (steradian). They quantify the areal subtense of a 2-D surface area in 3-D space viewed from the point of definition. A sphere subtends a surface area of $4\pi r^2$, which corresponds to a solid angle of 4π sr. Given a surface area A that is tilted under some angle θ between the surface normal and the line of sight the solid angle is reduced by a factor of $\cos\theta$:

$$\Omega = \frac{A}{r^2}\cos\theta \tag{2.5}$$

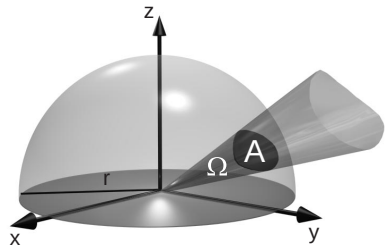


Figure 2.4: *Definition of solid angle. (By C. Garbe, University of Heidelberg.)*

Table 2.1: *Definitions of radiometric quantities (corresponding photometric quantities are defined in Table 2.2)*

Quantity	Symbol	Units	Definition
Radiant energy	Q	Ws	Total energy emitted by a source or received by a detector
Radiant flux	Φ	W	Total power emitted by a source or received by a detector
Radiant exitance	M	$\mathrm{W\,m^{-2}}$	Power emitted per unit surface area
Irradiance	E	$\mathrm{W\,m^{-2}}$	Power received at unit surface element
Radiant intensity	I	$\mathrm{W\,sr^{-1}}$	Power leaving a point on a surface into unit solid angle
Radiance	L	$\mathrm{W\,m^{-2}\,sr^{-1}}$	Power leaving unit projected surface area into unit solid angle

From the definition of angles as ratios of lengths or areas it follows that they have no physical unit. However, it is advisable always to use the artificial units rad and sr when referring to quantities related to angles to avoid confusion. Radiometric and photometric quantities also have to be defined carefully as their meaning cannot be inferred from physical units (Tables 2.1 and 2.2).

2.3.2 Conventions and overview

Measurements of radiometric and photometric quantities very often are subject to confusion related to terminology and units. Due to diverse historical developments and often inaccurate usage of names, radiometry is one of the least understood subjects in the field of op-

Table 2.2: *Definitions of photometric quantities (corresponding radiometric quantities are defined in Table 2.1)*

Quantity	Symbol	Units	Definition
Luminous energy	Q_v	lm s	Total luminous energy emitted by a source or received by a detector
Luminous flux	Φ_v	lm (lumen)	Total luminous power emitted by a source or received by a detector
Luminous exitance	M_v	lm m^{-2}	Luminous power emitted per unit surface area
Illuminance	E_v	lm m^{-2} = lx (lux)	Luminous power received at unit surface element
Luminous intensity	I_v	lumen sr^{-1} = cd (candela)	Luminous power leaving a point on a surface into unit solid angle
Luminance	L_v	lumen m^{-2} sr^{-1} = cd m^{-2}	Luminous power leaving unit projected surface area into unit solid angle

tics. However, it is not very difficult if some care is taken with regard to definitions of quantities related to angles and areas.

Despite confusion in the literature, there seems to be a trend towards standardization of units. In pursuit of standardization we will use only SI units, in agreement with the International Commission on Illumination CIE. The CIE is the international authority defining terminology, standards, and basic concepts in radiometry and photometry. The radiometric and photometric terms and definitions are in compliance with the American National Standards Institute (ANSI) report RP-16, published in 1986. Further information on standards can be found at the web sites of CIE (http://www.cie.co.at/cie/) and ANSI (http://www.ansi.org), respectively.

In this section, the fundamental quantities of radiometry will be defined. The transition to photometric quantities will be introduced by a generic equation Eq. (2.31) that can be used to convert each of these radiometric quantities to its corresponding photometric counterpart.

We will start from the concept of radiative flux and derive the most important quantities necessary to define the geometric distribution of radiation emitted from or irradiated on surfaces. The six fundamental concepts relating the spatial distribution of energy in electromagnetic radiation are summarized in Table 2.1. The term “radiant” is only

added to the names of those quantities that could be confused with the corresponding photometric quantity (see Table 2.2).

2.3.3 Definition of radiometric quantities

Radiant energy and radiant flux. Radiation carries energy that can be absorbed in matter heating up the absorber or interacting with electrical charges. *Radiant energy* Q is measured in units of Joule (J). It quantifies the total energy emitted by a source or received by a detector.

Radiant flux Φ is defined as radiant energy per unit time interval

$$\Phi = \frac{\mathrm{d}Q}{\mathrm{d}t} \tag{2.6}$$

passing through or emitted from a surface. Radiant flux has the unit Watts (W) and is also frequently called *radiant power*, which corresponds to its physical unit. Quantities describing the spatial and geometric distributions of radiative flux are introduced in the following sections.

The units for radiative energy, radiative flux, and all derived quantities listed in Table 2.1 are based on Joule as the fundamental unit. Instead of these *energy-derived* quantities an analogous set of *photon-derived* quantities can be defined based on the number of photons. Photon-derived quantities are denoted by the subscript p, while the energy-based quantities are written with a subscript e if necessary to distinguish between them. Without a subscript, all radiometric quantities are considered energy-derived. Given the radiant energy the number of photons can be computed from Eq. (2.2)

$$N_p = \frac{Q_e}{e_p} = \frac{\lambda}{hc} Q_e \tag{2.7}$$

With photon-based quantities the number of photons replaces the radiative energy. The set of photon-related quantities is useful if radiation is measured by detectors that correspond linearly to the number of absorbed photons (*photon detectors*) rather than to thermal energy stored in the detector material (*thermal detector*).

Photon flux Φ_p is defined as the number of photons per unit time interval

$$\Phi_p = \frac{\mathrm{d}N_p}{\mathrm{d}t} = \frac{\lambda}{hc} \frac{\mathrm{d}Q_e}{\mathrm{d}t} = \frac{\lambda}{hc} \Phi_e \tag{2.8}$$

Similarly, all other photon-related quantities can be computed from the corresponding energy-based quantities by dividing them by the energy of a single photon.

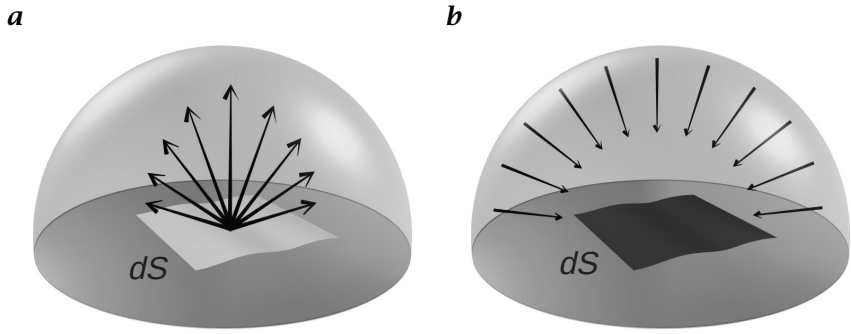


Figure 2.5: *Illustration of the radiometric quantities:* ***a*** *radiant exitance and* ***b*** *irradiance. (By C. Garbe, University of Heidelberg.)*

Because the conversion from energy-derived to photon-derived quantities Eq. (2.7) depends on the wavelength of radiation, spectral distributions of radiometric quantities will have different shapes for both sets of units (Fig. 2.10).

Radiant exitance and irradiance. *Radiant exitance* M defines the radiative flux *emitted* per unit surface area

$$M = \frac{\mathrm{d}\Phi}{\mathrm{d}S} \tag{2.9}$$

of a specified surface. The flux leaving the surface is radiated into the whole hemisphere enclosing the surface element dS and has to be integrated over all angles to obtain M (Fig. 2.5a). The flux is, however, not radiated uniformly in angle. Radiant exitance is a function of position on the emitting surface, $M = M(\boldsymbol{x})$. Specification of the position on the surface can be omitted if the emitted flux Φ is equally distributed over an extended area S. In this case $M = \Phi/S$.

Irradiance E similarly defines the radiative flux *incident* on a certain point of a surface per unit surface element

$$E = \frac{\mathrm{d}\Phi}{\mathrm{d}S} \tag{2.10}$$

Again, incident radiation is integrated over all angles of the enclosing hemisphere (Fig. 2.5b). Radiant exitance characterizes an actively radiating source while irradiance characterizes a passive receiver surface. Both are measured in $\mathrm{W\,m^{-2}}$ and cannot be distinguished by their units if not further specified.

Radiant intensity. *Radiant intensity* I describes the angular distribution of radiation emerging from a point in space. It is defined as radiant

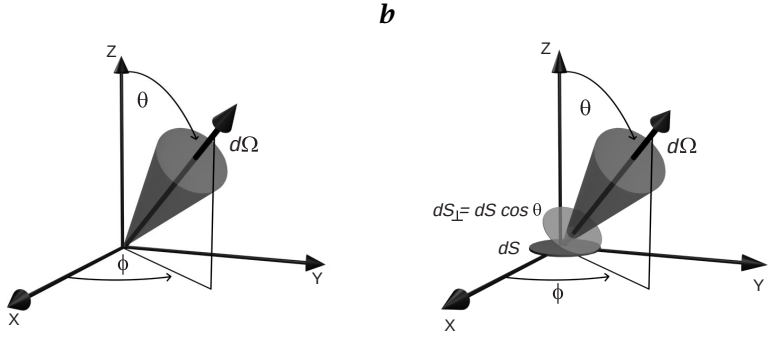


Figure 2.6: *Illustration of radiometric quantities:* ***a*** *radiant intensity and* ***b*** *radiance. (By C. Garbe, University of Heidelberg.)*

flux per unit solid angle

$$I = \frac{\mathrm{d}\Phi}{\mathrm{d}\Omega} \tag{2.11}$$

and measured in units of $\mathrm{W\,sr^{-1}}$. Radiant intensity is a function of the direction of the beam of radiation, defined by the spherical coordinates θ and ϕ (Fig. 2.6). Intensity is usually used to specify radiation emitted from *point sources*, such as stars or sources that are much smaller than their distance from the detector, that is, $dxdy \ll r^2$. In order to use it for extended sources those sources have to be made up of an infinite number of infinitesimal areas. The radiant intensity in a given direction is the sum of the radiant flux contained in all rays emitted in that direction under a given solid angle by the entire source (see Eq. (2.22)).

The term intensity is frequently confused with irradiance or illuminance. It is, however, a precisely defined quantity in radiometric terminology and should only be used in this context to avoid confusion.

Radiance. *Radiance L* defines the amount of radiant flux per unit solid angle per unit projected area of the emitting source

$$L = \frac{\mathrm{d}^2\Phi}{\mathrm{d}\Omega\,\mathrm{d}S_\perp} = \frac{\mathrm{d}^2\Phi}{\mathrm{d}\Omega\,\mathrm{d}S\cos\theta} \tag{2.12}$$

The differential $\mathrm{d}S_\perp = \mathrm{d}S\cos\theta$ defines a surface element perpendicular to the direction of the radiated beam (Fig. 2.6b). The unit of radiance is $\mathrm{W\,m^{-2}\,sr^{-1}}$. Radiance combines the concepts of exitance and intensity, relating intensity in a certain direction to the area of the emitting surface. And conversely, it can be thought of as exitance of the projected area per unit solid angle.

Radiance is used to characterize an *extended source* that has an area comparable to the squared viewing distance. As radiance is a

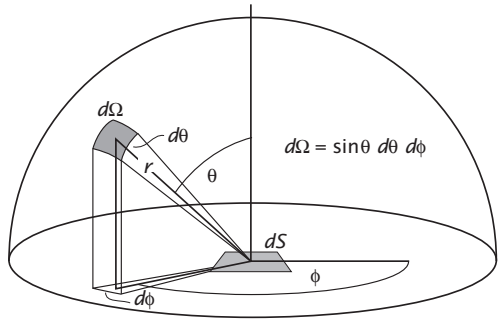


Figure 2.7: *Illustration of spherical coordinates.*

function of both position on the radiating surface as well as direction $L = L(\boldsymbol{x}, \theta, \phi)$, it is important always to specify the point in the surface and the emitting angles. It is the most versatile quantity in radiometry as all other radiometric quantities can be derived from the radiance integrating over solid angles or surface areas (Section 2.3.4).

2.3.4 Relationship of radiometric quantities

Spatial distribution of exitance and irradiance. Solving Eq. (2.12) for $d\Phi/dS$ yields the fraction of exitance radiated under the specified direction into the solid angle $d\Omega$

$$\mathrm{d}M(\boldsymbol{x}) = \mathrm{d}\left(\frac{\mathrm{d}\Phi}{\mathrm{d}S}\right) = L(\boldsymbol{x}, \theta, \phi)\cos\theta\,\mathrm{d}\Omega \tag{2.13}$$

Given the radiance L of an emitting surface, the radiant exitance M can be derived by integrating over all solid angles of the hemispheric enclosure $\mathcal{H}$:

$$M(\boldsymbol{x}) = \int_{\mathcal{H}} L(\boldsymbol{x}, \theta, \phi)\cos\theta\,\mathrm{d}\Omega = \int_0^{2\pi}\int_0^{\pi/2} L(\boldsymbol{x}, \theta, \phi)\cos\theta\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \tag{2.14}$$

In order to carry out the angular integration *spherical coordinates* have been used (Fig. 2.7), replacing the differential solid angle element $\mathrm{d}\Omega$ by the two plane angle elements $\mathrm{d}\theta$ and $\mathrm{d}\phi$:

$$\mathrm{d}\Omega = \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \tag{2.15}$$

Correspondingly, the irradiance E of a surface S can be derived from a given radiance by integrating over all solid angles of incident radiation:

$$E(\boldsymbol{x}) = \int_{\mathcal{H}} L(\boldsymbol{x},\theta,\phi)\cos\theta\,\mathrm{d}\Omega = \int_0^{2\pi}\int_0^{\pi/2} L(\boldsymbol{x},\theta,\phi)\cos\theta\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \quad (2.16)$$

A perfectly collimated beam of radiation, for example a very narrow laser beam, does not diverge and therefore occupies no finite solid angle ($\Omega = 0$). From Eq. (2.16) it follows that, in this case, $E = 0$. Therefore, a collimated beam cannot produce irradiance and does not seem to carry radiant flux. The concept of rays, however, frequently proves to be important for geometric optics. In order to combine radiometry and geometric optics it is useful to express the radiance L by the total amount of irradiance E_0 carried by the beam and the direction of propagation by the *Dirac delta distribution* $\delta(\theta-\theta_0,\phi-\phi_0)$:

$$L(\boldsymbol{x},\theta,\phi) = \frac{E_0(\boldsymbol{x})}{\cos\theta}\delta(\theta-\theta_0,\phi-\phi_0) \quad (2.17)$$

The delta distribution is defined by the following mathematical properties:

$$\delta(\theta-\theta_0,\phi-\phi_0) = \begin{cases} \infty & \text{for} \quad \theta=\theta_0 \quad \text{and} \quad \phi=\phi_0 \\ 0 & \text{for} \quad \theta\neq\theta_0 \quad \text{and} \quad \phi\neq\phi_0 \end{cases} \quad (2.18)$$

and

$$\int_0^{2\pi}\int_0^{\pi/2} \delta(\theta-\theta_0,\phi-\phi_0)\sin\theta\;\mathrm{d}\theta\,\mathrm{d}\phi = 1 \quad (2.19)$$

Equation (2.19) constitutes a special form of the general integral property of the delta distribution for spherical coordinates. Substituting Eq. (2.17) into Eq. (2.16) and using Eq. (2.19) yields the beam irradiance

$$E(\boldsymbol{x}) = E_0(\boldsymbol{x})\int_0^{2\pi}\int_0^{\pi/2} \delta(\theta-\theta_0,\phi-\phi_0)\sin\theta\;\mathrm{d}\theta\;\mathrm{d}\phi = E_0(\boldsymbol{x}) \quad (2.20)$$

Angular distribution of intensity. Solving Eq. (2.12) for $\mathrm{d}\Phi/\,\mathrm{d}\Omega$ yields the fraction of intensity emitted from an infinitesimal surface element $\mathrm{d}S$

$$\mathrm{d}I = d\left(\frac{d\Phi}{d\Omega}\right) = L(\boldsymbol{x},\theta,\phi)\cos\theta\;\mathrm{d}S \quad (2.21)$$

Extending the point source concept of radiant intensity to extended sources, the intensity of a surface of finite area can be derived by integrating the radiance over the emitting surface area S:

$$I(\theta,\phi) = \int_S L(\boldsymbol{x},\theta,\phi)\cos\theta \; \mathrm{d}S \tag{2.22}$$

The infinitesimal surface area $\mathrm{d}S$ is given by $\mathrm{d}S = \mathrm{d}s_1\,\mathrm{d}s_2$, with the *generalized coordinates* $\boldsymbol{s} = [s_1, s_2]^T$ defining the position on the surface. For planar surfaces these coordinates can be replaced by *Cartesian coordinates* $\boldsymbol{x} = [x, y]^T$ in the plane of the surface.

Total radiant flux. Solving Eq. (2.12) for $\mathrm{d}^2\Phi$ yields the fraction of radiant flux emitted from an infinitesimal surface element $\mathrm{d}S$ under the specified direction into the solid angle $\mathrm{d}\Omega$

$$\mathrm{d}^2\Phi = L(\boldsymbol{x},\theta,\phi)\cos\theta \; \mathrm{d}S \; \mathrm{d}\Omega \tag{2.23}$$

The total flux emitted from the entire surface area S into the hemispherical enclosure $\mathcal{H}$ can be derived by integrating over both the surface area and the solid angle of the hemisphere

$$\Phi = \int_S \int_{\mathcal{H}} L(\boldsymbol{x},\theta,\phi)\cos\theta \; \mathrm{d}\Omega\,\mathrm{d}S = \int_S \int_0^{2\pi} \int_0^{\pi/2} L(\boldsymbol{x},\theta,\phi)\cos\theta\sin\theta \; \mathrm{d}\theta \; \mathrm{d}\phi \; \mathrm{d}S \tag{2.24}$$

Again, spherical coordinates have been used for $\mathrm{d}\Omega$ and the surface element $\mathrm{d}S$ is given by $\mathrm{d}S = \mathrm{d}s_1\,\mathrm{d}s_2$, with the *generalized coordinates* $\boldsymbol{s} = [s_1, s_2]^T$. The flux emitted into a detector occupying only a fraction of the surrounding hemisphere can be derived from Eq. (2.24) by integrating over the solid angle Ω_D subtended by the detector area instead of the whole hemispheric enclosure $\mathcal{H}$.

Inverse square law. A common rule of thumb for the decrease of irradiance of a surface with distance of the emitting source is the *inverse square law*. Solving Eq. (2.11) for $\mathrm{d}\Phi$ and dividing both sides by the area $\mathrm{d}S$ of the receiving surface, the irradiance of the surface is given by

$$E = \frac{\mathrm{d}\Phi}{\mathrm{d}S} = I\frac{\mathrm{d}\Omega}{\mathrm{d}S} \tag{2.25}$$

For small surface elements dS perpendicular to the line between the point source and the surface at a distance r from the point source, the

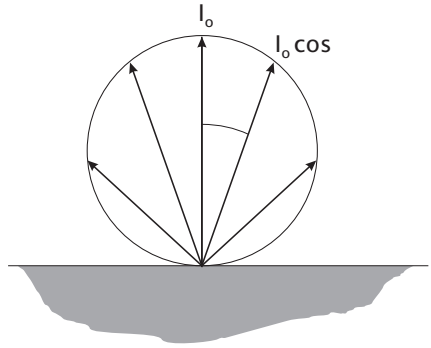


Figure 2.8: *Illustration of angular distribution of radiant intensity emitted from a Lambertian surface.*

subtended solid angle $\mathrm{d}\Omega$ can be written as $\mathrm{d}\Omega = \mathrm{d}S/r^2$. This yields the expression

$$E = \frac{I\,\mathrm{d}S}{\mathrm{d}S r^2} = \frac{I}{r^2} \tag{2.26}$$

for the irradiance E at a distance r from a point source with radiant intensity I. This relation is an accurate and simple means of verifying the linearity of a detector. It is, however, only true for point sources. For extended sources the irradiance on the detector depends on the geometry of the emitting surface (Chapter 3).

Lambert's cosine law. Radiant intensity emitted from extended surfaces is usually not evenly distributed in angle. A very important relation for perfect emitters, or perfect receivers, is *Lambert's cosine law*. A surface is called *Lambertian* if its radiance is independent of view angle, that is, $L(\boldsymbol{x}, \theta, \phi) = L(\boldsymbol{x})$. The angular distribution of radiant intensity can be computed directly from Eq. (2.22):

$$I(\theta) = \cos\theta \int_S L(\boldsymbol{x})\,\mathrm{d}S = I_0 \cos\theta \tag{2.27}$$

It is independent of angle ϕ and shows a cosine dependence on the angle of incidence θ as illustrated in Fig. 2.8. The exitance of a planar Lambertian surface is derived from Eq. (2.14), pulling L outside of the angular integrals

$$M(\boldsymbol{x}) = L(\boldsymbol{x}) \int_0^{2\pi} \int_0^{\pi/2} \cos\theta \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi = \pi L(\boldsymbol{x}) \tag{2.28}$$

The proportionality factor of π shows that the effect of Lambert's law is to yield only one-half the exitance, which might be expected for a surface radiating into 2π steradians. For point sources, radiating evenly into all directions with an intensity I, the proportionality factor would be 2π. Non-Lambertian surfaces would have proportionality constants smaller than π.

Another important consequence of Lambert's cosine law is the fact that Lambertian surfaces appear to have the same brightness under all view angles. This seems to be inconsistent with the cosine dependence of emitted intensity. To resolve this apparent contradiction, radiant power transfer from an extended source to a detector element with an area of finite size has to be investigated. This is the basic topic of *radiometry* and will be presented in detail in Chapter 5.

It is important to note that Lambert's cosine law only describes perfect radiators or perfect diffusers. It is frequently used to define rules of thumb, although it is not valid for real radiators in general. For small angles of incidence, however, Lambert's law holds for most surfaces. With increasing angles of incidence, deviations from the cosine relationship increase (Section 3.3.3).

2.3.5 Spectral distribution of radiation

So far *spectral distribution* of radiation has been neglected. Radiative flux is made up of radiation at a certain wavelength λ or mixtures of wavelengths, covering fractions of the electromagnetic spectrum with a certain wavelength distribution. Correspondingly, all derived radiometric quantities have certain spectral distributions. A prominent example for a spectral distribution is the spectral exitance of a blackbody given by Planck's distribution (Section 2.5.1).

Let Q be any radiometric quantity. The subscript λ denotes the corresponding *spectral* quantity Q_λ concentrated at a specific wavelength within an infinitesimal wavelength interval $d\lambda$. Mathematically, Q_λ is defined as the derivative of Q with respect to wavelength λ:

$$Q_\lambda = \mathrm{d}Q\lambda = \lim_{\Delta\lambda \to 0} \frac{\Delta Q}{\Delta\lambda} \tag{2.29}$$

The unit of Q_λ is given by $[\cdot/\mathrm{m}]$ with $[\cdot]$ denoting the unit of the quantity Q. Depending on the spectral range of radiation it sometimes is more convenient to express the wavelength dependence in units of $[\cdot/\mu\mathrm{m}]$ ($1\,\mu\mathrm{m} = 10^{-6}\,\mathrm{m}$) or $[\cdot/\mathrm{nm}]$ ($1\,\mathrm{nm} = 10^{-9}\,\mathrm{m}$). Integrated quantities over a specific wavelength range $[\lambda_1, \lambda_2]$ can be derived from

spectral distributions by

$$Q_{\lambda_1}^{\lambda_2} = \int_{\lambda_1}^{\lambda_2} Q_\lambda d\lambda \tag{2.30}$$

with $\lambda_1 = 0$ and $\lambda_2 = \infty$ as a special case. All definitions and relations derived in Sections 2.3.3 and 2.3.4 can be used for both spectral distributions of radiometric quantities and total quantities, integrated over the spectral distribution.

2.4 Fundamental concepts of photometry

Photometry relates radiometric quantities to the brightness sensation of the human eye. Historically, the naked eye was the first device to measure light and visual perception is still important for designing illumination systems and computing the apparent brightness of sources and illuminated surfaces.

While radiometry deals with electromagnetic radiation of all wavelengths, photometry deals only with the visible portion of the electromagnetic spectrum. The human eye is sensitive to radiation between 380 and 780 nm and only radiation within this visible portion of the spectrum is called "light."

2.4.1 Spectral response of the human eye

Light is perceived by stimulating the retina after passing the preretinal optics of the eye. The retina consists of two different types of receptors: rods and cones. At high levels of irradiance the cones are used to detect light and to produce the sensation of colors (*photopic vision*). Rods are used mainly for night vision at low illumination levels (*scotopic vision*). Both types of receptors have different sensitivities to light at different wavelengths.

The response of the "standard" light-adapted eye is defined by the normalized *photopic spectral luminous efficiency function* $V(\lambda)$ (Fig. 2.9). It accounts for eye response variation as relates to wavelength and shows the effectiveness of each wavelength in evoking a brightness sensation. Correspondingly, the *scotopic luminous efficiency function* $V'(\lambda)$ defines the spectral response of a dark-adapted human eye (Fig. 2.9). These curves were formally adopted as standards by the International Lighting Commission (CIE) in 1924 and 1951, respectively. Tabulated values can be found in [1, 2, 3, 4, 5]. Both curves are similar in shape. The peak of the relative spectral luminous efficiency curve for scotopic vision is shifted to 507 nm compared to the peak at 555 nm for photopic vision. The two efficiency functions can be thought of as the transfer

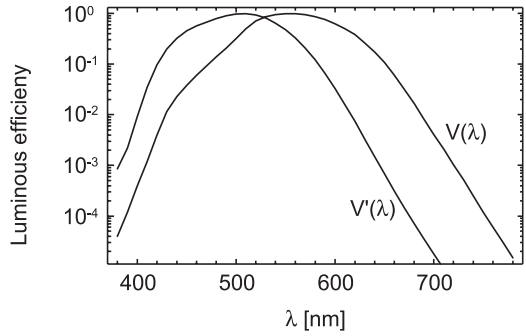


Figure 2.9: *Spectral luminous efficiency function of the "standard" light-adapted eye for photopic vision $V(\lambda)$ and scotopic vision $V'(\lambda)$, respectively.*

function of a filter, which approximates the behavior of the human eye under good and bad lighting conditions, respectively.

As the response of the human eye to radiation depends on a variety of physiological parameters, differing for individual human observers, the spectral luminous efficiency function can correspond only to an average normalized observer. Additional uncertainty arises from the fact that at intermediate illumination levels both photopic and scotopic vision are involved. This range is called *mesopic vision*.

2.4.2 Definition of photometric quantities

In order to convert radiometric quantities to their photometric counterparts, absolute values of the spectral luminous efficiency function are needed instead of relative functions. The relative spectral luminous efficiency functions for photopic and scotopic vision are normalized to their peak values, which constitute the quantitative conversion factors. These values have been repeatedly revised and currently (since 1980) are assigned the values 683 lm W^{-1} (lumen/watt) at 555 nm for photopic vision, and 1754 lm W^{-1} at 507 nm for scotopic vision, respectively.

The absolute values of the conversion factors are arbitrary numbers based on the definition of the unit *candela* (or international standard candle) as one of the seven base units of the metric system (SI). The name of this unit still reflects the historical illumination standard: a candle at a distance of 1 mile observed by the human eye. It is obvious that this corresponds to the definition of light intensity: a point source emitting light into a solid angle defined by the aperture of an average human eye and the squared distance. The current definition of candela is the *luminous intensity* of a source emitting monochromatic radiation of frequency 5.4×10^{14} Hz with a *radiant intensity* of 1/683 W sr^{-1} [2]. A practical calibration standard is the *primary standard of light* adopted

in 1918. It defines the candela as luminous intensity in the perpendicular direction of a surface of 1/60 cm^2 of a blackbody (Section 2.5.1) at the temperature of freezing platinum under a pressure of 1013.25 mbar [6, 7].

The conversion from photometric to radiometric quantities reduces to one simple equation. Given the conversion factors for photopic and scotopic vision, any (energy-derived) radiometric quantity $Q_{e,\lambda}$ can be converted into its photometric counterpart Q_v by

$$Q_v = 683\,\mathrm{lm\,W^{-1}} \int_{380}^{780} Q_{e,\lambda} V(\lambda)\,\mathrm{d}\lambda \tag{2.31}$$

for photopic vision and

$$Q_v = 1754\,\mathrm{lm\,W^{-1}} \int_{380}^{780} Q_{e,\lambda} V'(\lambda)\,\mathrm{d}\lambda \tag{2.32}$$

for scotopic vision, respectively. From this definition it can be concluded that photometric quantities can be derived only from known spectral distributions of the corresponding radiometric quantities. For invisible sources emitting radiation below 380 nm or above 780 nm all photometric quantities are null.

Table 2.2 on page 15 summarizes all basic photometric quantities together with their definition and units.

Luminous energy and luminous flux. The *luminous energy* can be thought of as the portion of radiant energy causing a visual sensation at the human retina. Radiant energy beyond the visible portion of the spectrum can also be absorbed by the retina, eventually causing severe damage to the tissue, but without being visible to the human eye. The *luminous flux* defines the total luminous energy per unit time interval ("luminous power") emitted from a source or received by a detector. The units for luminous flux and luminous energy are lm (lumen) and lm s, respectively.

Luminous exitance and illuminance. Corresponding to radiant exitance and irradiance, the photometric quantities *luminous exitance* and *illuminance* define the luminous flux per unit surface area leaving a surface or incident on a surface, respectively. As with the radiometric quantities, they are integrated over the angular distribution of light. The units of both luminous exitance and illuminance are lm m^{-2} or lux.

Luminous intensity. *Luminous intensity* defines the total luminous flux emitted into unit solid angle under a specified direction. As with its

radiometric counterpart, radiant intensity, it is used mainly to describe point sources and rays of light. Luminous intensity has the unit lm sr^{-1} or candela (cd). For a monochromatic radiation source with $I_\lambda = I_0\,\delta(\lambda - 555\,\text{nm})$ and $I_0 = 1/683\,\text{W sr}^{-1}$, Eq. (2.31) yields $I_\nu = 1\,\text{cd}$ in correspondence to the definition of candela.

Luminance. *Luminance* describes the subjective perception of "brightness" because the output of a photometer is proportional to the luminance of the measured radiation (Chapter 5). It is defined as luminant flux per unit solid angle per unit projected surface area perpendicular to the specified direction, corresponding to radiance, its radiometric equivalent. Luminance is the most versatile photometric quantity, as all other quantities can be derived by integrating the luminance over solid angles or surface areas. Luminance has the unit cd m^{-2}.

2.4.3 Luminous efficacy

Luminous efficacy is used to determine the effectiveness of radiative or electrical power in producing visible light. The term "efficacy" must not be confused with "efficiency". Efficiency is a dimensionless constant describing the ratio of some energy input to energy output. Luminous efficacy is not dimensionless and defines the fraction of luminous energy output able to stimulate the human visual system with respect to incoming radiation or electrical power. It is an important quantity for the design of illumination systems.

Radiation luminous efficacy. *Radiation luminous efficacy* K_r is a measure of the effectiveness of incident radiation in stimulating the perception of light in the human eye. It is defined as the ratio of any photometric quantity Q_ν to the radiometric counterpart Q_e integrated over the entire spectrum of electromagnetic radiation:

$$K_r = \frac{Q_\nu}{Q_e}\,[\text{lm W}^{-1}], \quad \text{where} \quad Q_e = \int_0^\infty Q_{e,\lambda}\,\mathrm{d}\lambda \tag{2.33}$$

It is important to note that Eq. (2.33) can be evaluated for any radiometric quantity with the same result for K_r. Substituting Q_ν in Eq. (2.33) by Eq. (2.31) and replacing $Q_{e,\lambda}$ by monochromatic radiation at 555 nm, that is, $Q_{e,\lambda} = Q_0\,\delta(\lambda - 555\,\text{nm})$, K_r reaches the value $683\,\text{lm W}^{-1}$. It can be easily verified that this is the theoretical maximum luminous efficacy a beam can have. Any invisible radiation, such as infrared or ultraviolet radiation, has zero luminous efficacy.

Lighting system luminous efficacy. The *lighting system luminous efficacy* K_s of a light source is defined as the ratio of perceptible luminous

flux Φ_v to the total power P_e supplied to the light source:

$$K_s = \frac{\Phi_v}{P_e} \, [\mathrm{lm\,W^{-1}}] \tag{2.34}$$

With the *radiant efficiency* $\tilde{\eta} = \Phi_e / P_e$ defining the ratio of total radiative flux output of an illumination source to the supply power, Eq. (2.34) can be expressed by the radiation luminous efficacy, K_r:

$$K_s = \frac{\Phi_v}{\Phi_e} \frac{\Phi_e}{P_e} = K_r \tilde{\eta} \tag{2.35}$$

Because the radiant efficiency of an illumination source is always smaller than 1, the lighting system luminous efficacy is always smaller than the radiation luminous efficacy. An extreme example is monochromatic laser light at a wavelength of 555 nm. Although K_r reaches the maximum value of $683\,\mathrm{lm\,W^{-1}}$, K_s might be as low as $1\,\mathrm{lm\,W^{-1}}$ due to the low efficiency of laser radiation.

2.5 Thermal emission of radiation

All objects at temperatures above absolute zero emit electromagnetic radiation. This *thermal radiation* is produced by accelerated electrical charges within the molecular structure of objects. Any accelerated charged particle is subject to emission of electromagnetic radiation according to the Maxwell equations of electromagnetism. A rise in temperature causes an increase in molecular excitation within the material accelerating electrical charge carriers. Therefore, radiant exitance of thermally emitting surfaces increases with the temperature of the body.

2.5.1 Blackbody radiation

In order to formulate the laws of thermal radiation quantitatively, an idealized perfect steady-state emitter has been specified. A *blackbody* is defined as an ideal body absorbing all radiation incident on it regardless of wavelength or angle of incidence. No radiation is reflected from the surface or passing through the blackbody. Such a body is a perfect absorber. *Kirchhoff* demonstrated in 1860 that a good absorber is a good emitter and, consequently, a perfect absorber is a perfect emitter. A blackbody, therefore, would emit the maximum possible radiative flux that any body can radiate at a given kinetic temperature, unless it contains fluorescent or radioactive materials.

Due to the complex internal structure of matter thermal radiation is made up of a broad range of wavelengths. However, thermal radiation

emitted from incandescent objects obeys the same laws as thermal radiation emitted from cold objects at room temperature and below. In 1900, *Max Planck* theoretically derived the fundamental relationship between the spectral distribution of thermal radiation and temperature [8]. He found that the spectral radiance of a perfect emitter at absolute temperature T is given by

$$L_{e,\lambda}(T) = \frac{2hc^2}{\lambda^5}\left[\exp\left(\frac{ch}{k_B\lambda T}\right) - 1\right]^{-1} \tag{2.36}$$

$$L_{p,\lambda}(\lambda, T) = \frac{2c}{\lambda^4}\left[\exp\left(\frac{ch}{k_B\lambda T}\right) - 1\right]^{-1} \tag{2.37}$$

with

$$\begin{array}{lcll} h & = & 6.6256 \times 10^{-34}\,\mathrm{J\,s} & \text{Planck's constant} \\ k_B & = & 1.3805 \times 10^{-23}\,\mathrm{J\,K^{-1}} & \text{Boltzmann constant} \\ c & = & 2.9979 \times 10^{8}\,\mathrm{m\,s^{-1}} & \text{speed of light in vacuum} \end{array} \tag{2.38}$$

The photon-related radiance of a blackbody $L_{p,\lambda}(T)$ is obtained by dividing the energy related radiance $L_{e,\lambda}(T)$ by the photon energy e_p as given by Eq. (2.2). Detailed derivations of Planck's law can be found in [7, 9, 10].

Although the assumption of a perfect emitter seems to restrict the practical usage, Planck's law proves useful to describe a broad range of thermally emitting objects. Sources like the sun, incandescent lamps, or—at much lower temperatures—water and human skin have blackbody-like emission spectra. The exact analytical form of blackbody radiation is an invaluable prerequisite for absolute radiometric calibration standards.

Figure 2.10 shows several Planck distributions for different temperatures. As already pointed out at the beginning of this chapter, the shapes of energy-derived and photon-derived quantities deviate from each other due to the conversion from photon energy into photon number. It is also of interest to note that a single generalized blackbody radiation curve may be drawn for the combined parameter λT, which can be used for determining spectral exitance at any wavelength and temperature. Figure 2.11a shows this curve as fractional exitance relative to the peak value, plotted as a function of λT. The fraction of the total exitance lying below any given value of λT is also shown. An interesting feature of Planck's curve is the fact that exactly one-fourth of the exitance is radiated below the peak value.

In Fig. 2.11b the solar irradiance above the earth's atmosphere is plotted together with the exitance of a blackbody at $T = 6000\,\mathrm{K}$, which corresponds to the temperature of the solar surface (Section 6.2.1).

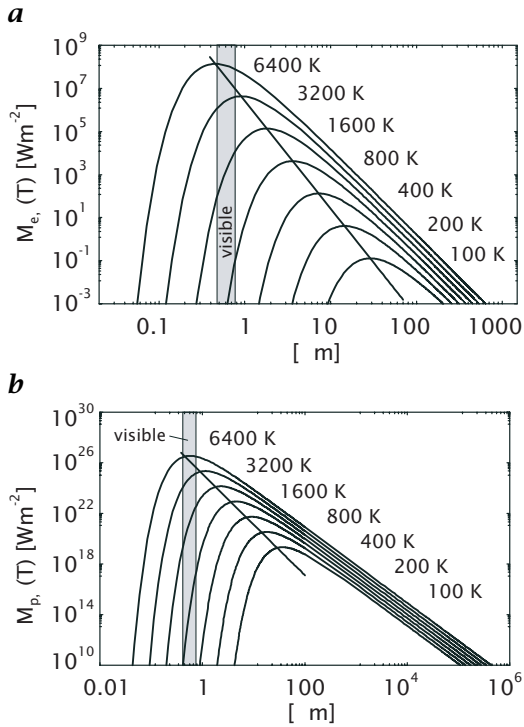


Figure 2.10: ***a*** *Spectral energy-derived exitance of a blackbody vs wavelength at temperatures from 100 K-6400 K.* ***b*** *Spectral photon-derived exitance of a blackbody at the same temperatures.*

2.5.2 Properties of Planck's distribution

Angular Distribution. A blackbody, by definition, radiates uniformly in angle. The radiance of a blackbody surface is independent of view angle, that is, $L_\lambda(T, \theta, \phi) = L_\lambda(T)$. This surface property is called Lambertian (Section 2.3.4). Therefore, blackbody radiation is fully specified by the surface temperature T. All radiometric quantities can be derived from the spectral radiance distributions, Eq. (2.36) or Eq. (2.37), as outlined in Section 2.3.4. An important example is the spectral radiant exitance of a blackbody $M_\lambda(T)$, which is simply given by $\pi L_\lambda(T)$ because a blackbody, by definition, has a Lambertian surface:

$$M_{e,\lambda}(T) = \frac{2\pi hc^2}{\lambda^5}\left[\exp\left(\frac{ch}{k_B\lambda T}\right) - 1\right]^{-1} \tag{2.39}$$

$$M_{p,\lambda}(T) = \frac{2\pi c}{\lambda^4}\left[\exp\left(\frac{ch}{k_B\lambda T}\right) - 1\right]^{-1} \tag{2.40}$$

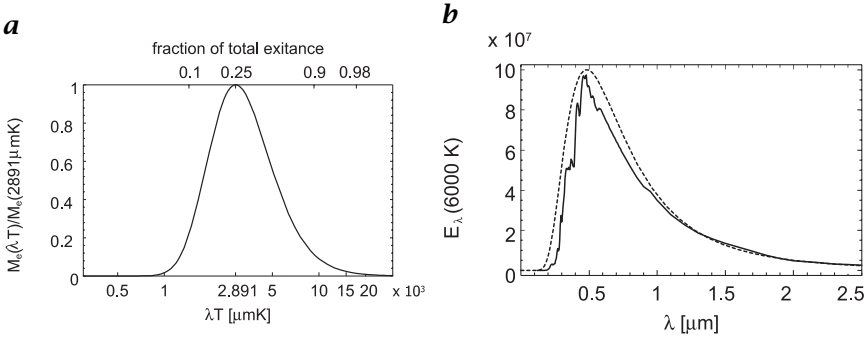


Figure 2.11: ***a*** *Generalized blackbody exitance for any combination of* λ *and* T. ***b*** *Solar irradiance above the earth's atmosphere compared to the exitance of a blackbody at a temperature of* $T = 6000\,K$ *(dashed line).*

Stefan-Boltzmann law. Integrating the spectral radiant exitance $M_\lambda(T)$ over all wavelengths yields the total radiant exitance $M(T)$:

$$M_e(T) = \int_0^\infty M_{e,\lambda}(T)d\lambda = \frac{2}{15}\frac{k_B^4\pi^5}{c^2h^3}T^4 = \sigma T^4 \tag{2.41}$$

where $\sigma = 5.668 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$ is the *Stefan-Boltzmann constant*. The total radiant flux emitted by a blackbody per unit surface element increases with the fourth power of the temperature, which is known as the *Stefan-Boltzmann law*. This relation was originally postulated by *Josef Stefan* in 1879 and verified by *Ludwig Boltzmann* in 1884 by thermodynamic considerations, before the Planck relation was derived.

Similarly, the total photon exitance over all wavelengths can be derived by integrating Eq. (2.40) over the entire spectrum:

$$M_p(T) = \int_0^\infty M_{p,\lambda}(T)\,\mathrm{d}\lambda = \sigma_p T^3 \tag{2.42}$$

where σ_p is approximately $\sigma_p = 1.52 \times 10^{15}\,\mathrm{photon\,s^{-1}\,m^{-2}\,K^{-3}}$. Note that the total photon exitance only increases with the third power of the temperature.

Not only does the total radiant exitance increase with temperature but also the spectral radiant exitance for any wavelength. This means that the Planck curve for a temperature T_2 lies completely above the Planck curve for any temperature T_1 with $T_1 < T_2$ (Fig. 2.10). This property allows a quantitative temperature measurement of a blackbody surface from the radiant exitance within any spectral subregion of Planck's distribution (Chapter 5). For overlapping curves this could not be done unambiguously.

Wien's displacement law. The wavelength of maximum radiant exitance of a blackbody is given by the zero crossing of the partial derivative of $M_{e,\lambda}(T)$ with respect to λ:

$$\frac{dM_{e,\lambda}(T)}{d\lambda} = 0 \quad \rightsquigarrow \quad \left(1 - \frac{ch}{5k_b\lambda T}\right) \exp\left(\frac{ch}{k_b\lambda T}\right) = 1 \tag{2.43}$$

Solving Eq. (2.43) for λ yields *Wien's displacement law*

$$\lambda_{m,e}\, T = 2.891 \times 10^{-3}\,\mathrm{m\,K} \tag{2.44}$$

quantifying the decrease in the wavelength of peak energy exitance of a blackbody $\lambda_{m,e}$ to be inversely proportional to the temperature T. If the integral in Eq. (2.41) is split into two parts for wavelengths $0 < \lambda < \lambda_{m,e}$ and $\lambda_{m,e} < \lambda < \infty$, it can be verified that exactly 25 % of the total radiant exitance is emitted below $\lambda_{m,e}$ and the remaining 75 % above $\lambda_{m,e}$. Typical incandescent lamps with a temperature of approximately 1000 K have a peak exitance at $\lambda_{m,e} \approx 3\,\mu$m. Therefore, only a small portion well below 25 % of the total exitance is emitted in the visible spectral range (Section 2.5.4).

Similarly, the corresponding Wien's displacement law for maximum wavelength for photon-related radiant exitance can be found to be

$$\lambda_{m,p}\, T = 3.662 \times 10^{-3}\,\mathrm{m\,K} \tag{2.45}$$

Therefore, the peak photon exitance (Fig. 2.10b) is shifted towards longer wavelengths compared to the peak energy-related exitance of a blackbody (Fig. 2.10a).

2.5.3 Approximations of Planck's distribution

For small and large values of λT the Planck curve can be approximated by two distributions that historically were known as partial solutions before Planck's law.

Wien's radiation law. If λT is sufficiently small, that is, $\lambda T \ll hc/k_B$, then $\exp(hc/\lambda k_B T) \gg 1$ and Eq. (2.36) reduces to

$$L_{e,\lambda}(T) = \frac{2hc^2}{\lambda^5} \exp\left(-\frac{ch}{k_B\lambda T}\right) \tag{2.46}$$

This relation is known as *Wien's radiation law*. It predicts the existence of a peak exitance but deviates for large values of λT from the Planck distribution (Fig. 2.12).

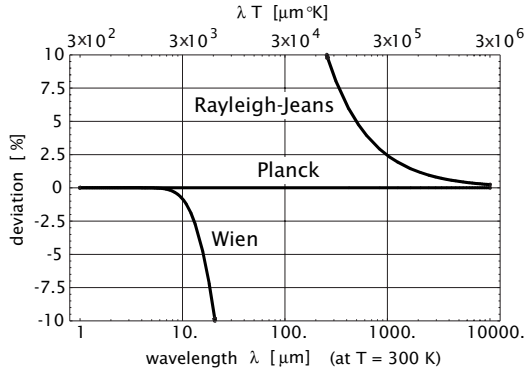


Figure 2.12: *Deviation of Wien's radiation law and Rayleigh-Jeans law from the exact Planck distribution.*

Rayleigh-Jeans law. For large values of $\lambda T \gg hc/k_B$ an approximate solution can be found by expanding the exponential factor of Eq. (2.36) in a Taylor series

$$L_{e,\lambda}(T) = \frac{2hc^2}{\lambda^5}\left[\frac{ch}{k_B\lambda T} + \frac{1}{2}\left(\frac{ch}{k_B\lambda T}\right)^2 + \cdots\right]^{-1} \tag{2.47}$$

Disregarding all terms of second and higher order in Eq. (2.47) yields the *Rayleigh-Jeans law*

$$L_{e,\lambda}(T) = \frac{2ck_B}{\lambda^4}T \tag{2.48}$$

This law is a good approximation of the decrease of $L_{e,\lambda}(T)$ at large wavelengths. At small wavelengths the predicted exitance approaches infinity, which is known as the UV catastrophe (Fig. 2.12).

2.5.4 Luminous efficacy of blackbody radiation

An important quantity of an incandescent object used as illumination source is the radiation luminous efficacy K_r. Replacing Q_ν in Eq. (2.33) by the blackbody luminous exitance $M_\nu(T)$ computed from Eq. (2.31) with Eq. (2.39) and using the Stefan-Boltzmann law Eq. (2.41) yields

$$K_r(T) = \frac{683}{\sigma T^4}\int_{380}^{780} M_\lambda(T)V(\lambda)d\lambda \;[\mathrm{lm\,W^{-1}}] \tag{2.49}$$

Figure 2.13 shows K_r for a temperature range from 2000 K to 40,000 K. For temperatures up to 2000 K the radiant luminous efficacy lies well

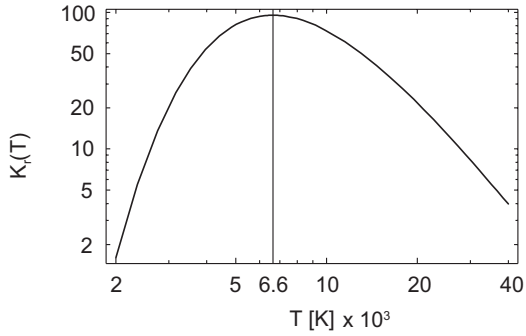


Figure 2.13: *Radiation luminous efficacy of a blackbody vs temperature T.*

below 1 lm W^{-1}. This shows that typical incandescent lamps with temperatures below 2000 K are very inefficient illumination sources. Most of the energy is emitted in the IR region. The peak of the radiation luminous efficacy of blackbody radiation lies at 6600 K which is close to the surface temperature of the sun. This demonstrates how the human visual system has adapted to the solar spectrum by evolution.

2.6 Acoustic waves

Although it does not belong to electromagnetic radiation, *ultrasound* is gaining increasing importance in *acoustic imaging* applications such as medical imaging. With improved detector performance resolutions of less than 1 mm can be achieved. The major advantage of ultrasound is its performance in penetrating opaque objects, rigid bodies, as well as fluid systems, in a nondestructive way. Prominent examples are material research and medical diagnostics

Ultrasound consists of acoustic waves with frequencies between 15 kHz and 10 GHz (10^{10} Hz). It is generated by electroacoustical transducers such as piezoelectric crystals at resonant frequencies. The lowest eigenfrequency of a Piezo quartz plate of thickness l is given by

$$\nu_0 = \sqrt{c_q/2l} \tag{2.50}$$

where $c_q = 5.6 \times 10^5$ cm s^{-1} is the speed of sound in quartz. The spectrum of emitted frequencies consists of integer multiples of ν_0.

In contrast to electromagnetic waves, acoustic waves need a carrier. They travel with the speed of sound in the carrier medium, which is given by

$$c_m = (\rho_0 \beta_{ad})^{-1/2} \tag{2.51}$$

where ρ_0 is the static density and β_{ad} the *adiabatic compressibility*:

$$\beta_{ad} = -\frac{1}{V}\frac{\partial V}{\partial P} \tag{2.52}$$

It is given as the relative volume change caused by a uniform pressure without heat exchange. As the speed of acoustic waves c_m depends only on the elastic properties of the medium, acoustic waves of all frequencies travel with the same speed. Thus, acoustic waves show no dispersion. This important feature is used in acoustic imaging techniques to measure the density of the medium by run length measurements of ultrasonic reflexes.

Equation (2.51) is only valid for *longitudinal waves* caused by isotropic pressure with deformation in the direction of propagation. Due to the internal structure of solids the propagation of sound waves is no longer isotropic and shear forces give rise to transversal acoustic waves.

2.7 References

[1] Oriel Corporation, (1994). *Light Sources, Monochromators & Spectrographs, Detectors & Detection Systems, Fiber Optics*, Vol. II. Stratford, CT: Oriel Corporation.

[2] CIE, (1983). *The Basis of Physical Photometry*. Technical Report.

[3] Kaufman, J. E. (ed.), (1984). *IES Lighting Handbook—Reference Volume*. New York: Illuminating Engineering Society of North America.

[4] Laurin Publishing, (1998). *The Photonics Design and Applications Handbook*, 44th edition. Pittsfield, MA: Laurin Publishing CO.

[5] McCluney, W. R., (1994). *Introduction to Radiometry and Photometry*. Boston: Artech House.

[6] Walsh, J. W. T. (ed.), (1965). *Photometry*, 3rd edition. New York: Dover.

[7] Wolfe, W. L. and Zissis, G. J. (eds.), (1989). *The Infrared Handbook*, 3rd edition. Michigan: The Infrared Information Analysis (IRIA) Center, Environmental Research Institute of Michigan.

[8] Planck, M., (1901). *Ann. Phys.*, **4(3)**:p. 553.

[9] Dereniak, E. L. and Boreman, G. D., (1996). *Infrared Detectors and Systems*. New York: John Wiley & Sons, Inc.

[10] Planck, M., (1991). *The Theory of Heat Radiation*. New York: Dover.

3 Interaction of Radiation with Matter

Horst Haußecker

Interdisziplinäres Zentrum für Wissenschaftliches Rechnen (IWR)
Universität Heidelberg, Germany

3.1 Introduction

This chapter provides the fundamentals of interaction of radiation with objects and matter. It should help those who want to search for an appropriate spectral range to visualize object features rather than taking illumination as the given. Quantitative visualization in computer vision requires knowledge of both the physical properties of the objects

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

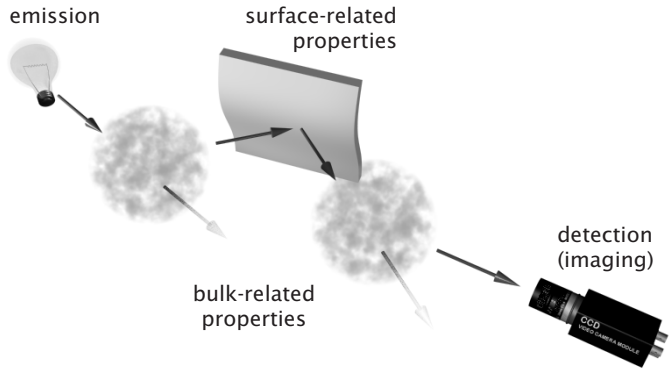


Figure 3.1: *Illustration of the radiometric chain of image formation. (By C. Garbe, University of Heidelberg.)*

of interest in terms of interaction with radiation as well as the optical properties of the imaging system. In addition to the performance of the detector, the performance and availability of optical components are essential factors for quality and computer vision system costs.

Physical quantities such as penetration depth or surface reflectivity are essential to probe the internal structures of objects, scene geometry, and surface-related properties. Physical object properties, therefore, not only can be encoded in the geometrical distribution of emitted radiation but also in the portion of radiation being emitted, scattered, absorbed, or reflected and finally reaching the imaging system. Most of these processes are sensitive to certain wavelengths and additional information might be hidden in the spectral distribution of radiation. Using different types of radiation allows images from different depths or object properties to be attained. As an example, infrared radiation of between 3 and 5 μm is absorbed in human skin within a depth of less than 1 mm, while x-rays pass through the whole body without major attenuation. Therefore, totally different properties of the human body (such as skin temperature as well as skeletal structures) can be revealed for medical diagnosis.

Standard scenes usually contain more than one single object in a uniform enclosure. Radiation has to pass a series of events, called the *radiometric chain*, before it reaches the imaging system. Figure 3.1 illustrates how incident radiation is influenced by all objects and matter along the optical path. In this chapter, the basic mechanisms influencing the emission of radiation and its propagation in matter will be detailed.

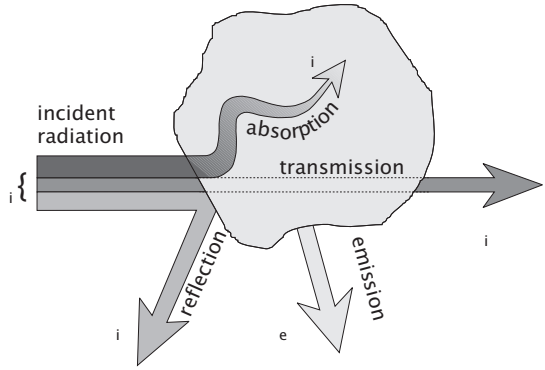


Figure 3.2: *Radiative flux, Φ_i incident on an object is partially reflected (fraction $\tilde{\rho}$) and absorbed (fraction $\tilde{\alpha}$). For nonopaque objects a fraction $\tilde{\tau}$ is passing the body. The radiative flux $\tilde{\epsilon}\Phi_e$ is emitted to maintain or reach thermodynamic equilibrium.*

3.2 Basic definitions and terminology

3.2.1 Definition of optical properties

Radiation incident on or passing through objects is subject to various processes changing the direction of propagation, attenuating or amplifying the radiant intensity, and changing the spectral distribution or polarization of radiation. Without going into the details of the complex physical processes governing the interaction of radiation with the molecular structure of objects, the macroscopic properties of objects relevant for radiometry are detailed in this section.

In order to quantify the optical properties of surfaces and objects the following dimensionless quantities are defined (Fig. 3.2):

Reflectivity *Reflectivity* or *reflectance* $\tilde{\rho}$ defines the ratio of the reflected radiative flux Φ_r to the incident radiative flux Φ_i,

$$\tilde{\rho} = \frac{\Phi_r}{\Phi_i} \tag{3.1}$$

Absorptivity *Absorptivity* or *absorptance* $\tilde{\alpha}$ defines the ratio of the absorbed radiative flux Φ_a to the incident radiative flux Φ_i,

$$\tilde{\alpha} = \frac{\Phi_a}{\Phi_i} \tag{3.2}$$

Transmissivity *Transmissivity* or *transmittance* $\tilde{\tau}$ defines the ratio of the radiative flux Φ_t transmitting the object to the incident radiative flux Φ_i,

$$\tilde{\tau} = \frac{\Phi_t}{\Phi_i} \tag{3.3}$$

Emissivity The forementioned quantities $\tilde{\rho}$, $\tilde{\alpha}$, and $\tilde{\tau}$ define the property of *passive* receivers in modifying incident radiative flux. The *emissivity* or *emittance* $\tilde{\epsilon}$ quantifies the performance of an *actively* radiating object compared to a blackbody, which provides the upper limit of the spectral exitance of a source. It is defined by the ratio of the exitances,

$$\tilde{\epsilon} = \frac{M_s(T)}{M_b(T)} \tag{3.4}$$

where M_s and M_b denote the exitance of the emitting source, and the exitance of the blackbody at the temperature T, respectively. As a blackbody has the maximum possible exitance of an object at the given temperature, $\tilde{\epsilon}$ is always smaller than 1.

3.2.2 Spectral and directional dependencies

All of the foregoing introduced quantities can have strong variations with direction, wavelength, and polarization state that have to be specified in order to measure the optical properties of an object. The emissivity of surfaces usually only slightly decreases for angles of up to 50° and rapidly falls off for angles larger than 60°; it approaches zero for 90° [1]. The reflectivity shows the inverse behavior.

To account for these dependencies, we define the spectral *directional* emissivity $\tilde{\epsilon}(\lambda, \theta, \phi)$ as ratio of the source spectral radiance $L_{\lambda,s}$ to the spectral radiance of a blackbody $L_{\lambda,b}$ at the same temperature T:

$$\tilde{\epsilon}(\lambda, \theta, \phi) = \frac{L_{\lambda,s}(\theta, \phi, T)}{L_{\lambda,b}(\theta, \phi, T)} \tag{3.5}$$

The spectral *hemispherical* emissivity $\tilde{\epsilon}(\lambda)$ is similarly given by the radiant exitance of the source and a blackbody at the same temperature, T:

$$\tilde{\epsilon}(\lambda) = \frac{M_{\lambda,s}(T)}{M_{\lambda,b}(T)} \tag{3.6}$$

Correspondingly, we can define the spectral directional reflectivity, the spectral directional absorptivity, and the spectral directional transmissivity as functions of direction and wavelength. In order to simplify notation, the symbols are restricted to $\tilde{\rho}$, $\tilde{\alpha}$, $\tilde{\tau}$ and $\tilde{\epsilon}$ without further indices. Spectral and/or directional dependencies will be indicated by the variables and are mentioned in the text.

3.2.3 Terminology conventions

Emission, transmission, reflection, and absorption of radiation either refer to surfaces and interfaces between objects or to the net effect of extended objects of finite thickness. In accordance with Siegel and Howell [2] and McCluney [3] we assign the suffix -ivity to surface-related (*intrinsic*) material properties and the suffix -ance to volume-related (*extrinsic*) object properties. To reduce the number of equations we exclusively use the symbols $\tilde{\epsilon}$, $\tilde{\alpha}$, $\tilde{\rho}$ and $\tilde{\tau}$ for both types. If not further specified, surface- and volume-related properties can be differentiated by the suffixes -ivity and -ance, respectively. More detailed definitions can be found in the *CIE International Lighting Vocabulary* [4].

3.2.4 Spectral selectivity

For most applications the spectral optical properties have to be related to the spectral sensitivity of the detector system or the spectral distribution of the radiation source. Let $\tilde{p}(\lambda)$ be any of the following material properties: $\tilde{\alpha}$, $\tilde{\rho}$, $\tilde{\tau}$, or $\tilde{\epsilon}$. The *spectral selective* optical properties $\tilde{p}_s$ can be defined by integrating the corresponding spectral optical property $\tilde{p}(\lambda)$ over the entire spectrum, weighted by a spectral window function $w(\lambda)$:

$$\tilde{p}_s = \frac{\int_0^\infty w(\lambda)\tilde{p}(\lambda)d\lambda}{\int_0^\infty w(\lambda)\,\mathrm{d}\lambda} \tag{3.7}$$

Examples of spectral selective quantities include the *photopic luminous transmittance* or *reflectance* for $w(\lambda) = V(\lambda)$ (Chapter 2), the *solar transmittance*, *reflectance*, or *absorptance* for $w(\lambda) = E_{\lambda,s}$ (solar irradiance), and the *emittance* of an object at temperature T for $w(\lambda) = E_{\lambda,b}(T)$ (blackbody irradiance). The *total* quantities $\tilde{p}$ can be obtained by integrating $\tilde{p}(\lambda)$ over all wavelengths without weighting.

3.2.5 Kirchhoff's law

Consider a body that is in thermodynamic equilibrium with its surrounding environment. Conservation of energy requires $\Phi_i = \Phi_a + \Phi_r + \Phi_t$ and, therefore,

$$\tilde{\alpha} + \tilde{\rho} + \tilde{\tau} = 1 \tag{3.8}$$

Table 3.1: *Basic (idealized) object and surface types*

Object	Properties	Description
Opaque body	$\tilde{\epsilon}(\lambda) + \tilde{\rho}(\lambda) = 1$, $\tilde{\tau}(\lambda) = 0$	Cannot be penetrated by radiation. All exitant radiation is either reflected or emitted.
AR coating	$\tilde{\epsilon}(\lambda) + \tilde{\tau}(\lambda) = 1$, $\tilde{\rho}(\lambda) = 0$	No radiation is reflected at the surface. All exitant radiation is transmitted or emitted.
Ideal window	$\tilde{\epsilon}(\lambda) = \tilde{\rho}(\lambda) = 0$, $\tilde{\tau}(\lambda) = 1$	All radiation passes without attenuation. The temperature is not accessible by IR thermography because no thermal emission takes place.
Mirror	$\tilde{\epsilon}(\lambda) = \tilde{\tau}(\lambda) = 0$, $\tilde{\rho}(\lambda) = 1$	All incident radiation is reflected. The temperature is not accessible by IR thermography because no thermal emission takes place.
Blackbody	$\tilde{\tau}(\lambda) = \tilde{\rho}(\lambda) = 0$, $\tilde{\epsilon}(\lambda) = \tilde{\epsilon} = 1$	All incident radiation is absorbed. It has the maximum possible exitance of all objects.
Graybody	$\tilde{\epsilon}(\lambda) = \tilde{\epsilon} < 1$, $\tilde{\rho}(\lambda) = 1 - \tilde{\epsilon}$, $\tilde{\tau}(\lambda) = 0$	Opaque object with wavelength independent emissivity. Same spectral radiance as a blackbody but reduced by the factor $\tilde{\epsilon}$.

In order to maintain equilibrium, the emitted flux must equal the absorbed flux at each wavelength and in each direction. Thus

$$\tilde{\alpha}(\lambda, \theta, \phi) = \tilde{\epsilon}(\lambda, \theta, \phi) \tag{3.9}$$

This relation is known as *Kirchhoff's law* [5]. It also holds for the integrated quantities $\tilde{\epsilon}(\lambda)$ and $\tilde{\epsilon}$. Kirchoff's law does not hold for active optical effects shifting energy between wavelengths, such as fluorescence, or if thermodynamic equilibrium is not reached. Kirchhoff's law also does not apply generally for two different components of polarization [6, 7].

Table 3.1 summarizes basic idealized object and surface types in terms of the optical properties defined in this section. Real objects and surfaces can be considered a mixture of these types. Although the ideal cases usually do not exist for the entire spectrum, they can be realized for selective wavelengths. Surface coatings, such as, for example, anti-reflection (AR) coatings, can be technically produced with high precision for a narrow spectral region.

Figure 3.3 shows how radiometric measurements are influenced by the optical properties of objects. In order to measure the emitted flux Φ_1 (e. g., to estimate the temperature of the object), the remaining seven quantities $\tilde{\epsilon}_1$, $\tilde{\epsilon}_2$, $\tilde{\epsilon}_3$, $\tilde{\rho}_1$, $\tilde{\tau}_1$, Φ_2, and Φ_3 have to be known. Only for a blackbody is the total received flux the flux emitted from the object of interest.

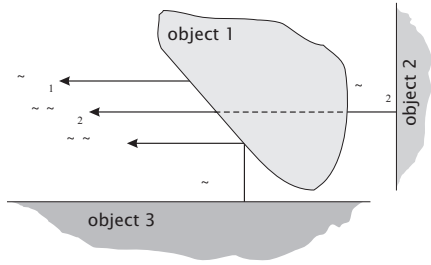


Figure 3.3: *Radiometric measurements of object 1 are biased by the radiation of the environment emitted from objects 2 and 3.*

3.2.6 Index of refraction

Solving the Maxwell equations for electromagnetic radiation in matter yields the *complex index of refraction*, N:

$$N(\lambda) = n(\lambda) + \mathrm{i}k(\lambda) \tag{3.10}$$

with the real part n and the imaginary part k.

The real part n constitutes the well-known index of refraction of geometric optics (Section 3.3.2, Chapter 4). From the complex part k other important optical properties of materials, such as *reflection*, and *absorption* can be derived (Sections 3.3 and 3.4).

3.3 Properties related to interfaces and surfaces

In this section properties of interfaces between two different materials are detailed. In this context an interface is defined as a discontinuity in optical properties over a distance that is much smaller than the wavelength of the radiation.

3.3.1 Surface emission

Thermal emission of radiation from object surfaces is characterized by the surface emissivity $\tilde{\epsilon}$. The spectral radiance of a real source $L_{\lambda,s}$ is given by

$$L_{\lambda,s}(\theta,\phi,T) = \tilde{\epsilon}(\lambda,\theta,\phi)L_{\lambda,b}(\theta,\phi,T) \tag{3.11}$$

where $L_{\lambda,b}$ denotes the radiance of a blackbody. A blackbody source will have $\tilde{\epsilon}(\lambda,\theta,\phi) = \tilde{\epsilon} = 1$.

A surface is called *graybody* if the emissivity is independent from wavelength and angle, $\tilde{\epsilon}(\lambda,\theta,\phi) = \tilde{\epsilon} < 1$. Graybodies are by definition Lambertian radiators. Radiometric quantities of graybodies have

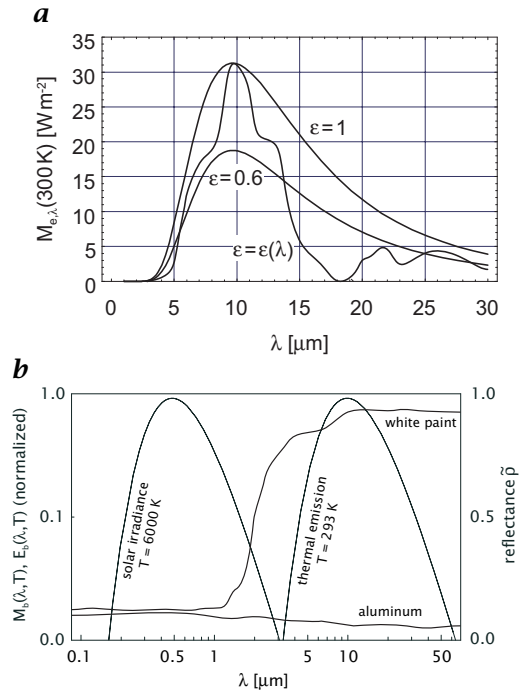


Figure 3.4: ***a*** *Spectral exitance of a blackbody, a graybody, and a selective emitter at the same temperature.* ***b*** *Spectral solar irradiance and spectral thermal exitance of a blackbody at ambient temperature vs spectral emissivity of aluminum and white paint, respectively (schematic).*

the same spectral shape as the same radiometric quantity of blackbodies, multiplied by the constant factor $\tilde{\epsilon}$ (Fig. 3.4a). Graybodies do not necessarily have to be gray. They appear to have the same color as a blackbody at the same temperature but have a lower total exitance:

$$M_{\lambda,g}(T) = \tilde{\epsilon}\sigma T^4 \tag{3.12}$$

A surface is called *nonblackbody* if the emissivity varies with wavelength. Such a surface is the general case and is also called *selective emitter* (Fig. 3.4a). Tabulated values of $\tilde{\epsilon}$ for common surface materials can be found in [3, 7].

Example 3.1: Infrared thermography

The temperature T of objects can be measured remotely by infrared thermography (Section 2.5, and Volume 3, Chapter 35). As already pointed out in Section 3.2, the fraction $(1 - \tilde{\epsilon})$ of the total exitance originates from the environment biasing the temperature measurement. The measured total exitance is interpreted to originate from

a blackbody at the apparent temperature T'. Assuming an isothermal environment at blackbody temperature T_e, the temperatures are related by the Stefan-Boltzmann law Eq. (2.41):

$$\sigma T'^4 = \tilde{\epsilon}\sigma T^4 + (1-\tilde{\epsilon})\sigma T_e^4 \tag{3.13}$$

In the limit of small temperature differences between environment and the body of interest ($T_e - T \ll T$), Eq. (3.13) can be approximated by [8]

$$T' \approx \tilde{\epsilon}T + (1-\tilde{\epsilon})T_e \quad \text{or} \quad T' - T = (1-\tilde{\epsilon})(T_e - T) \tag{3.14}$$

This simplified estimation gives a rule of thumb for errors associated with low emissivity. A 1 % deviation of ϵ from unity results in a 0.01 K temperature error per 1 K difference of object and ambient temperature. Although this is a simplified computation, it can be used to estimate the influence of ambient temperature on thermography of nonblackbodies. If the ambient temperature and the emissivity of the object are known, this error can be corrected according to Eq. (3.13). In this context it has to be pointed out that radiation from the environment can also originate from the cooled CCD detector of an IR camera itself being reflected from the object of interest. As IR detectors usually operate at liquid nitrogen temperature (75 K), errors in the temperature measurement in the order of 2 K can occur even for a very high emissivity of $\tilde{\epsilon} = 0.99$! Uncooled infrared imagers can reduce this type of error.

Example 3.2: Solar absorbers

A *solar absorber* has to be designed in such a way that as much solar irradiance as possible is collected without emitting the collected energy by thermal radiation. The absorber has to be covered with a coating that has a high absorptivity and, correspondingly, a high emissivity over the solar spectrum and a low emissivity over the longwave IR portion of the spectrum.

Example 3.3: Solar emitters

An aircraft painting needs to be a *solar emitter*. In order to reduce thermal heating and relieve air conditioning requirements during ground-based operations, the solar irradiance has to be reflected as much as possible. The absorptivity over the solar spectrum, therefore, has to be as low as possible. According to Fig. 3.4b this can be achieved by either white paint (TiO_2, $\tilde{\alpha}(0.5\,\mu\text{m}) = 0.19$ [7]) or polished aluminum ($\tilde{\alpha}(0.5\,\mu\text{m}) = 0.19$ [7]). Because an aircraft is made from aluminum, the surfaces used to be finished by the blank aluminum. Aluminum, however, remains at low emissivity over the entire IR portion of the spectrum ($\tilde{\epsilon}(10\,\mu\text{m}) = 0.05$ [3]; refer to Fig. 3.4b). Any solar energy that is not reflected heats up the plane and has to be emitted in the IR with maximum emissive power near 10 μm. White paint has a much higher emissivity in this portion of the spectrum (TiO_2, $\tilde{\epsilon}(10\,\mu\text{m}) = 0.94$ [9]), so white-painted surfaces remain up to

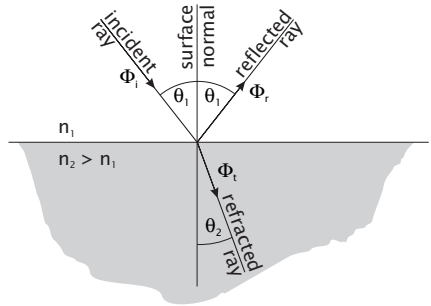


Figure 3.5: *Refraction and specular reflection at interfaces.*

19 K cooler under direct sunlight exposure than aluminum surfaces [10, 11]. Airline operators paint fuselage tops white today, rather than leaving their aluminum surface shiny.

3.3.2 Refraction

The real part $n(\lambda)$ of the complex index of refraction N Eq. (3.10) constitutes the index of refraction of geometric optics, that is, the ratio of the speed of light in a vacuum to the speed of light in a medium under consideration. It determines the change in the direction of propagation of radiation passing the interface of two materials with different dielectric properties. According to *Snell's law*, the angles of incidence θ_1 and refraction θ_2 are related by (Fig. 3.5)

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{n_2}{n_1} \tag{3.15}$$

where n_1 and n_2 are the indices of refraction of the two materials. It is the basis for transparent optical elements, such as lenses and prisms (Chapter 4). While prisms make use of the wavelength dependence of refraction to separate radiation of different wavelengths, lenses suffer from this effect (chromatic aberration).

3.3.3 Specular reflection

The direction of incident ray, reflected ray, and the surface normal vector span the plane of incidence perpendicular to the surface of reflection (Fig. 3.5). At smooth interfaces between two materials with different dielectric properties specular reflection occurs. The angles of incidence and reflection are equal (Fig. 3.6a).

The reflectivity, $\tilde{\rho}$, of a surface is defined as the ratio between incident and reflected flux. It depends on the indices of refraction of the two materials, the angle of incidence, and the polarization of the

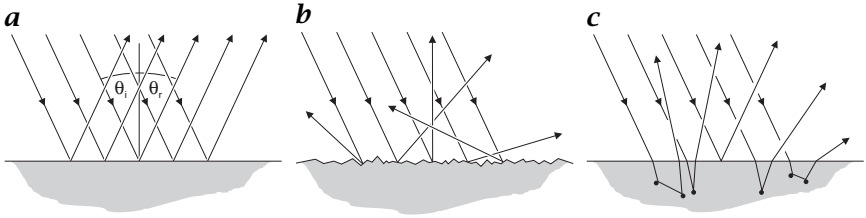


Figure 3.6: ***a*** *Specular,* ***b*** *diffuse,* ***c*** *and subsurface reflection at interfaces.*

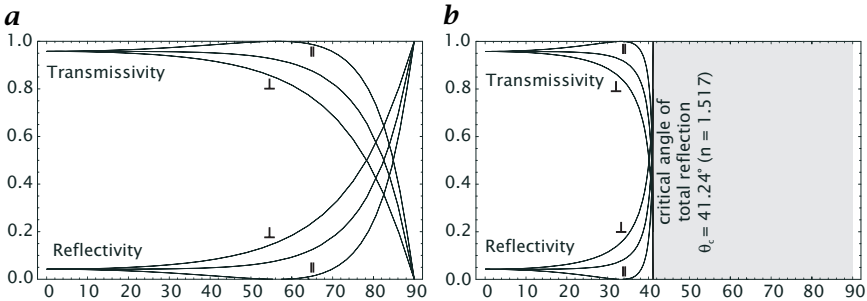


Figure 3.7: *Reflectivities and transmissivities vs angle of incidence for parallel (∥) and perpendicular (⊥) polarized light at the interface between air ($n_1 = 1.0$) and BK7 glass ($n_2 = 1.517$).* ***a*** *Transition air to glass.* ***b*** *Transition glass to air. The shaded area shows angles beyond the critical angle of total internal reflection.*

radiation. The specular reflectivities of the polarization components parallel (∥) and perpendicular (⊥) to the plane of incidence are given by *Fresnel's equations* [12]:

$$\tilde{\rho}_{\parallel} = \frac{\tan^2(\theta_1 - \theta_2)}{\tan^2(\theta_1 + \theta_2)}, \quad \tilde{\rho}_{\perp} = \frac{\sin^2(\theta_1 - \theta_2)}{\sin^2(\theta_1 + \theta_2)}, \quad \text{and} \quad \tilde{\rho} = \frac{\tilde{\rho}_{\parallel} + \tilde{\rho}_{\perp}}{2} \tag{3.16}$$

where the total reflectivity for unpolarized radiation $\tilde{\rho}$ is the average (arithmetic mean) of the two polarization components. The angles θ_1 and θ_2 are the angles of incidence and refraction in the medium, which are related by Snell's law, Eq. (3.15). Figure 3.7 shows the angular dependence of Eq. (3.16) for the transition from *BK7* glass to air and vice versa.

From Fresnel's equations three important properties of specular reflection at object interfaces can be inferred (Fig. 3.7):

1. Parallel polarized light is not reflected at all at a certain angle, called the *polarizing* or *Brewster angle* θ_b. At this angle the reflected and

refracted rays are perpendicular to each other [12]:

$$\theta_b = \arcsin \frac{1}{\sqrt{1 + n_1^2/n_2^2}} \tag{3.17}$$

2. At the transition from the medium with higher refractive index to the medium with lower refractive index, there is a *critical angle* θ_c

$$\theta_c = \arcsin \frac{n_1}{n_2}, \quad \text{with} \quad n_1 < n_2 \tag{3.18}$$

beyond which all light is reflected back into the medium of origin. At this angle Snell's law would produce an angle of refraction of 90°. The reflectivity is unity for all angles of incidence greater than θ_c, which is known as *total internal reflection* and used in light conductors and fiber optics.

3. At large (grazing) angles, object surfaces have a high reflectivity, independent from n. Therefore, objects usually deviate from an ideal Lambertian reflector for large angles of incidence.

At normal incidence ($\theta = 0$) there is no difference between perpendicular and parallel polarization and

$$\tilde{\rho} = \frac{(n_1 - n_2)^2}{(n_1 + n_2)^2} = \frac{(n-1)^2}{(n+1)^2}, \quad \text{with} \quad n = \frac{n_1}{n_2} \tag{3.19}$$

Note that Eqs. (3.16) and (3.19) are only exact solutions for transparent dielectric objects (Section 3.4) with small imaginary parts, k, of the complex refractive index N, Eq. (3.10): $k \ll 1$. For non-negligible imaginary parts the normal reflectivity Eq. (3.19) has to be modified:

$$\tilde{\rho} = \frac{(n_1 - n_2)^2 + k^2}{(n_1 + n_2)^2 + k^2} \tag{3.20}$$

The wavelength dependence of the refractive index can change the spectral composition of radiation by reflection. Silver (Ag) has a high reflectivity above 0.9 over the entire visible spectrum. The reflectivity of Gold (Au) also lies above 0.9 for wavelengths beyond 600 nm, but shows a sudden decrease to 0.4 for wavelengths below 500 nm. This increased absorption of blue light compared to red light is responsible for the reddish appearance of gold surfaces in contrast to the white metallic glare of silver surfaces.

3.3.4 Diffuse reflection

Very few materials have pure specular surface reflectivity. Most surfaces show a mixture of matte and specular reflection. As soon as surface microroughness has the same scale as the wavelength of radiation,

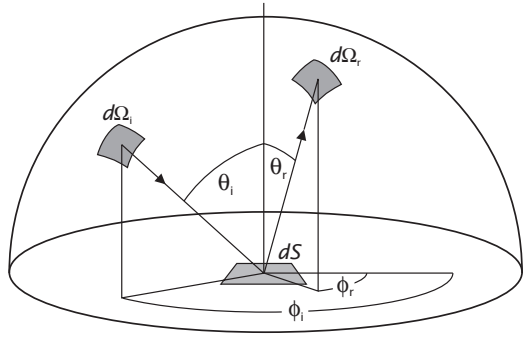


Figure 3.8: *Illustration of the angles used in the definition of the bidirectional reflectivity distribution function (BRDF).*

diffraction at the microstructures occurs. At larger scales, microfacets with randomly distributed slopes relative to the surface normal are reflecting incident light in various directions (Fig. 3.6b). Depending on the size and slope distribution of the microroughness, these surfaces have a great variety of reflectivity distributions ranging from isotropic (Lambertian) to strong forward reflection, where the main direction is still the angle of specular reflection. An excellent introduction into light scattering and surface roughness is provided by Bennet and Mattsson [13].

A mixture of specular and diffuse reflection can also be caused by subsurface scattering of radiation, which is no longer a pure surface-related property. Radiation penetrating a partially transparent object can be scattered at optical inhomogeneities (Section 3.4) and leave the object to cause diffuse reflection (Fig. 3.6c). Reflected light from below the surface is subject to bulk related interactions of radiation with matter that can change the spectral composition of radiation before it is re-emitted. For this reason, diffusely scattered light shows the colors of objects while highlights of specular reflections usually show the color of the incident light, which is white for ambient daylight.

In order to describe quantitatively the angular reflectivity distribution of arbitrary objects, the *bidirectional reflectivity distribution function* (BRDF), f, is used (Fig. 3.8). It is a function of the spherical angles of incidence (θ_i, ϕ_i) and reflection (θ_r, ϕ_r), and defines the ratio of reflected radiance L_r to the incident irradiance E_i of the reflecting surface [7]:

$$f(\theta_i, \phi_i, \theta_r, \phi_r) = \frac{L_r(\theta_r, \phi_r)}{E_i(\theta_i, \phi_i)} \tag{3.21}$$

This definition accounts for the fact that an optical system measures the radiance leaving a surface while distribution of incident radiation

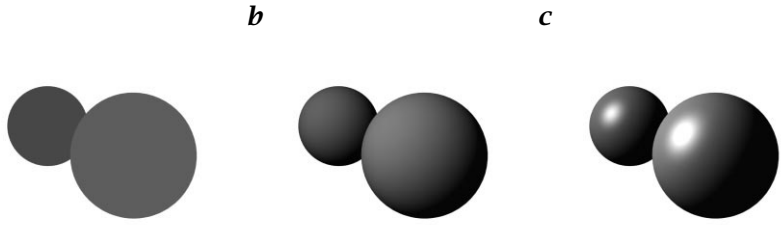


Figure 3.9: *Spheres shaded using the Phong illumination model:* ***a*** *ambient reflection,* ***b*** *diffuse reflection, and* ***c*** *specular reflection. (By C. Garbe, University of Heidelberg.)*

is quantified by the surface irradiance. The two extreme cases are specular and Lambertian surfaces. A purely specular surface has a nonzero value only for $\theta_i = \theta_r$ and $\phi_i = \phi_r$ so that $f = \tilde{\rho}\delta(\theta_i - \theta_r)\delta(\phi_i - \phi_r)$. A Lambertian surface has no dependence on angle, and a flat surface therefore has $f = \tilde{\rho}\pi^{-1}$. The hemispherical reflectivity in each case is $\tilde{\rho}$.

3.3.5 Reflection models in computer graphics

A major task of computer graphics is the realistic visualization of object surfaces incorporating material properties. A number of illumination models, called *lighting models* or *shading models*, have been developed for photorealistic rendering. Graphics researchers have often approximated the underlying rules of radiation theory either to simplify computation or because more accurate models were not known in the graphics community [14].

A physically motivated model has been introduced by Cook and Torrance [15], incorporating the surface roughness by microfacets with a certain probability distribution around the normal of the macroscopic surface. The internal complexity, however, prevents this approach from common usage in real-time computer graphics.

For practical usage an illumination model has become standard in computer graphics, assuming reflection to be a mixture of *ambient*, *diffuse* (Lambertian), and *specular* reflection. It can be implemented very efficiently and allows adaptation to most natural surface properties with good agreement to physical models[1].

[1] To stay consistent with radiometric notation, we replace the computer graphics symbols for reflectivity k_x by $\tilde{\rho}_x$ and replace the color coefficient $O_{x,\lambda}$ by using a spectral reflectivity $\tilde{\rho}_{x,\lambda}$. The subscript x denotes one of the indices a, d, and s for ambient, diffuse, and specular reflection. It also has to be pointed out that the term *intensity* is frequently used for the apparent brightness of a surface in computer graphics. As the brightness of a surface corresponds to the radiometric term *radiance* (Section 2.3.3) we use the term radiance exclusively.

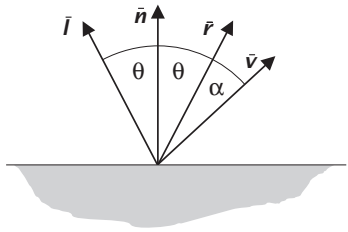


Figure 3.10: *Reflection at surfaces: Direction to light source $\bar{\boldsymbol{l}}$ surface normal vector $\bar{\boldsymbol{n}}$, direction of specular reflection $\bar{\boldsymbol{r}}$ direction to the viewer, $\bar{\boldsymbol{v}}$.*

Ambient reflection. The most simple approach assumes *ambient light*, with a spectral intensity $I_{a\lambda}$, impinging equally on all surfaces from all directions. The reflected spectral radiance $L_{a\lambda}$ of such a surface will be independent from viewing direction:

$$L_{a\lambda} = I_{a\lambda}\tilde{\rho}_{a\lambda} \tag{3.22}$$

where $\tilde{\rho}_{a\lambda}$ is the spectral *ambient reflection coefficient*. It is a material property that does not necessarily correspond to the physical reflectivity of the material. A surface rendered according to Eq. (3.22) will appear flat with a homogeneous brightness if $\tilde{\rho}_{a\lambda}$ remains constant over the object surface (Fig. 3.9a).

Diffuse reflection. For a perfectly diffuse (Lambertian) surface the reflected radiance $L_{d\lambda}$ does not depend on the angle of reflection. If a Lambertian surface is illuminated by a point light source with intensity $I_{p\lambda}$, the surface irradiance will vary with the cosine of the angle of incidence θ, which can be replaced by the inner vector product $\bar{\boldsymbol{n}}^T\bar{\boldsymbol{l}}$ of the surface normal $\bar{\boldsymbol{n}}$ and the normalized direction of incidence $\bar{\boldsymbol{l}}$ (Fig. 3.10). Thus,

$$L_{d\lambda} = f_p I_{p\lambda}\tilde{\rho}_{d\lambda}\cos\theta = f_p I_{p\lambda}\tilde{\rho}_{d\lambda}\bar{\boldsymbol{n}}^T\bar{\boldsymbol{l}} \tag{3.23}$$

where $\tilde{\rho}_{d\lambda}$ is the *diffuse reflection coefficient* and f_p defines the *light source attenuation factor* accounting for the distance d of the point source. A common practice is to set $f_p = 1/d^2$ according to the inverse square law Eq. (2.26). Refined models use an inverse second-order polynomial [14]. Objects rendered according to Eq. (3.23) appear to have been illuminated by a flashlight in a dark room (Fig. 3.9b).

Specular reflection. A popular illumination model for nonperfect reflectors was developed by Phong [16]. The *Phong illumination model* assumes that maximum reflectance occurs when the angle α between the direction of specular reflection $\bar{\boldsymbol{r}}$ and the viewing direction $\bar{\boldsymbol{v}}$ is

zero and falls off sharply with increasing α (Fig. 3.10). The falloff is approximated by $\cos^n \alpha$ with the *specular reflection exponent* n. This complies with the fact that the BRDF f of Eq. (3.21) can be approximated by a power of cosine for most surfaces. For a point light source with intensity $I_{p\lambda}$, the reflected radiance $L_{s\lambda}$ in this model is given by

$$L_{s\lambda} = f_p I_{p\lambda} \tilde{\rho}_{s\lambda}(\theta) \cos^n \alpha = f_p I_{p\lambda} \tilde{\rho}_{s\lambda} \left(\bar{\boldsymbol{r}}^T \bar{\boldsymbol{n}}\right)^n \tag{3.24}$$

where the *specular reflection coefficient* $\tilde{\rho}_{s\lambda}$ depends on the angular reflectivity distribution of specular reflection. It is, however, typically set to a constant. For a perfect mirror, n would be infinite; for a Lambertian surface it would be zero. Figure 3.9c shows a sphere illuminated by the Phong illumination model with $n = 10$.

Combined model. Combining all three different contributions gives the total reflected radiance

$$L_\lambda = I_{a\lambda} \tilde{\rho}_{a\lambda} + f_p I_{p\lambda} \left[\tilde{\rho}_{d\lambda} \bar{\boldsymbol{n}}^T \bar{\boldsymbol{l}} + \tilde{\rho}_{s\lambda} \left(\bar{\boldsymbol{r}}^T \bar{\boldsymbol{n}}\right)^n \right] \tag{3.25}$$

Instead of the accurate wavelength dependence, a simplified solution can be obtained, replacing Eq. (3.25) by three separate equations L_R, L_G, and L_B for the red, green, and blue components of the light source intensity and the reflection coefficients, respectively.

Refined surface illumination models can be found in [14]. Visualization of volume data will be detailed in Volume 2, Chapter 28.

3.4 Bulk-related properties of objects

This section deals with the various processes influencing the propagation of radiation within optical materials. The basic processes are attenuation by absorption or scattering, changes in polarization, and frequency shifts. For active emitters, radiation emitted from partially transparent sources can originate from subsurface volumes, which changes the radiance compared to plain surface emission.

3.4.1 Attenuation of radiation

Only a few optical materials have a transmissivity of unity, which allows radiation to penetrate without attenuation. The best example is ideal crystals with homogeneous regular grid structure. Most materials are either opaque or attenuate transmitted radiation to a certain degree. Let z be the direction of propagation along the optical path. Consider the medium being made up from a number of infinitesimal layers of thickness $\mathrm{d}z$ (Fig. 3.11). The fraction of radiance $\mathrm{d}L_\lambda = L_\lambda(z) - L_\lambda(z +$

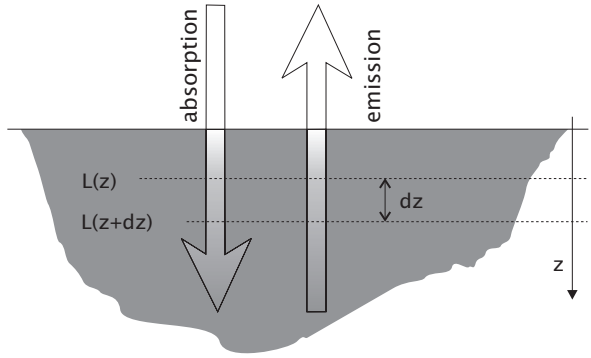


Figure 3.11: *Depth dependence of the volumetric absorption and emission of radiation.*

$\mathrm{d}z$) removed within the layer will be proportional to both the thickness $\mathrm{d}z$ and the radiance $L_\lambda(z)$ incident on the layer at z:

$$\mathrm{d}L_\lambda(z) = -\kappa(\lambda, z) L_\lambda(z)\,\mathrm{d}z \tag{3.26}$$

with the *extinction coefficient* or *attenuation coefficient* κ of the material (in environmental sciences, κ is sometimes referred to as *turbidity*). The unit of κ is a reciprocal length, such as m^{-1}. Solving Eq. (3.26) for L and integrating over z yields:

$$L_\lambda(z) = L_\lambda(0) \exp\left(-\int_0^z \kappa(\lambda, z')\,\mathrm{d}z'\right) \tag{3.27}$$

If the medium shows homogeneous attenuation, that is, $\kappa(\lambda, z) = \kappa(\lambda)$, Eq. (3.27) reduces to

$$L_\lambda(z) = L_\lambda(0) \exp\left(-\kappa(\lambda) z\right) \tag{3.28}$$

which is known as *Lambert Beer's* or *Bouguer's law* of attenuation. It has to be pointed out that Bouguer's law holds only for first-order (linear) processes Eq. (3.26), where $\mathrm{d}L$ is proportional to L. This is true for a wide range of practical applications, but breaks down for very high intensities, such as laser radiation, or if multiscatter processes play a dominant role.

So far there has not been a discussion as to which processes are responsible for attenuation of radiation. The two basic processes are *absorption* and *scattering*. Separating the total amount $\mathrm{d}L$ of radiation that is lost into the parts $\mathrm{d}L_a$ (absorption) and $\mathrm{d}L_s$ (scattering), $\mathrm{d}L = \mathrm{d}L_a + \mathrm{d}L_s$, the attenuation coefficient κ splits into the *absorption*

coefficient α and the *scattering coefficient* β:

$$\kappa = -\frac{1}{L}\frac{\mathrm{d}L}{\mathrm{d}z} = -\frac{1}{L}\frac{\mathrm{d}L_a}{\mathrm{d}z} - \frac{1}{L}\frac{\mathrm{d}L_s}{\mathrm{d}z} = \alpha + \beta \tag{3.29}$$

Both coefficients have the dimension of a reciprocal length (m^{-1}) and are intrinsic material properties.

In order to separate the effect of absorption and scattering on attenuation, both the transmitted as well as the scattered radiation in all directions has to be measured. For the transmitted beam, only the net effect of both processes can be measured if no further knowledge on the material properties is available.

The *transmittance*[2] of a layer of thickness z can be computed from Eq. (3.28) as

$$\tilde{\tau}(\lambda) = \frac{L_\lambda(z)}{L_\lambda(0)} = \exp\left(-\kappa(\lambda)z\right) \tag{3.30}$$

Therefore, a layer of thickness $\kappa^{-1}(\lambda)$ has a transmittance of e^{-1}. This distance is called *penetration depth* of the radiation at the specific wavelength. A variety of materials do not exhibit scattering. In these cases $\kappa = \alpha$.

Another frequently used term (mainly in spectroscopy) is the *optical depth* $\tau(z_1, z_2)$ of a medium. It is defined as integral over the attenuation coefficient:

$$\tau(z_1, z_2) = \int_{z_1}^{z_2} \kappa(z)\,\mathrm{d}z \tag{3.31}$$

Taking the logarithm of the radiance, Lambert Beer's law (see Eq. (3.27)) reduces to a sum over the optical depths of all M layers of material:

$$\ln L_\lambda(z) - \ln L_\lambda(0) = \sum_{m=0}^{M} \tau(z_m, z_{m+1}) \tag{3.32}$$

Again, for nonscattering media κ has to be replaced by α.

Absorption. The *absorption coefficient* α of a material can be computed from the imaginary part k of the complex index of refraction (Eq. (3.10)):

$$\alpha(\lambda) = \frac{4\pi k(\lambda)}{\lambda} \tag{3.33}$$

[2] As mentioned in Section 3.2.1, the *transmittance* of a layer of finite thickness must not be confused with the *transmissivity* of an interface.

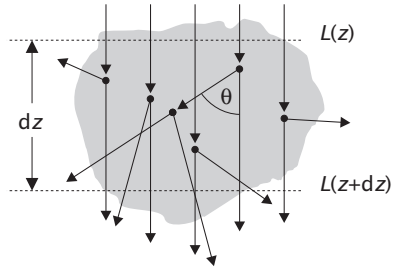


Figure 3.12: *Single and multiple scatter of radiation in materials with local inhomogeneities.*

Tabulated values of absorption coefficients for a variety of optical materials can be found in [7, 9, 17, 18].

The absorption coefficient of a medium is the basis for quantitative spectroscopy. With an imaging spectrometer, the distribution of a substance can be quantitatively measured, provided there is appropriate illumination (Volume 3, Chapter 37). The measured spectral absorption coefficient of a substance depends on the amount of material along the optical path and, therefore, is proportional to the concentration of the substance:

$$\alpha = \epsilon c \tag{3.34}$$

where c is the concentration in units mol l^{-1} and ϵ denotes the molar absorption coefficient with unit l mol^{-1} m^{-1}).

Scattering. Scatter of radiation is caused by variations of the refractive index as light passes through a material [18]. Causes include foreign particles or voids, gradual changes of composition, second phases at grain boundaries, and strains in the material. If radiation traverses a perfectly homogeneous medium, it is not scattered. Although any material medium has inhomogeneities as it consists of molecules, each of which can act as a scattering center, whether the scattering will be effective depends on the size and arrangement of these molecules. In a perfect crystal at zero temperature the molecules are arranged in a very regular way and the waves scattered by each molecule interfere in such a way as to cause no scattering at all but just a change in the velocity of propagation, given by the index of refraction (Section 3.3.2).

The net effect of scattering on incident radiation can be described in analogy to absorption Eq. (3.26) with the *scattering coefficient* $\beta(\lambda, z)$ defining the proportionality between incident radiance $L_\lambda(z)$ and the amount dL_λ removed by scattering along the layer of thickness dz (Fig. 3.12).

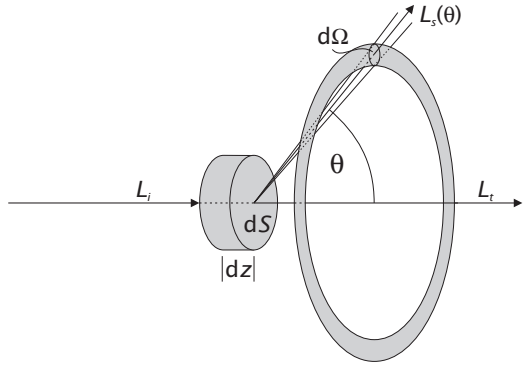


Figure 3.13: *Geometry for the definition of the volume scattering function f_{VSF}.*

The basic assumption for applying Eq. (3.26) to scattering is that the effect of a volume containing M scattering particles is M times that scattered by a single particle. This simple proportionality to the number of particles holds only, if the radiation to which each particle is exposed is essentially radiation of the initial beam. For high particle densities and, correspondingly, high scattering coefficients, multiple scattering occurs (Fig. 3.12) and the simple proportionality does not exist. In this case the theory becomes very complex. A means of testing the proportionality is to measure the optical depth τ Eq. (3.31) of the sample. As a rule of thumb, single scattering prevails for $\tau < 0.1$. For $0.1 < \tau < 0.3$ a correction for double scatter may become necessary. For values of $\tau > 0.3$ the full complexity of multiple scattering becomes a factor [19]. Examples of multiple scatter media are white clouds. Although each droplet may be considered an independent scatterer, no direct solar radiation can penetrate the cloud. All droplets only diffuse light that has been scattered by other drops.

So far only the net attenuation of the transmitted beam due to scattering has been considered. A quantity accounting for the angular distribution of scattered radiation is the *spectral volume scattering function*, f_{VSF}:

$$f_{VSF}(\theta) = \frac{\mathrm{d}^2\Phi_s(\theta)}{E_i\,\mathrm{d}\Omega\,\mathrm{d}V} = \frac{\mathrm{d}^2L_s(\theta)}{L_i\,\mathrm{d}\Omega\,\mathrm{d}z} \tag{3.35}$$

where $\mathrm{d}V = \mathrm{d}S\,\mathrm{d}z$ defines a volume element with a cross section of $\mathrm{d}S$ and an extension of $\mathrm{d}z$ along the optical path (Fig. 3.13). The indices i and s denote incident and scattered quantities, respectively. The volume scattering function considers scatter to depend only on the angle θ with axial symmetry and defines the fraction of incident radiance being scattered into a ring-shaped element of solid angle (Fig. 3.13).

From the volume scattering function, the total scattering coefficient β can be obtained by integrating f_{VSF} over a full spherical solid angle:

$$\beta(\lambda) = \int_0^{2\pi}\int_0^{\pi} f_{VSF}(\lambda,\theta)\,\mathrm{d}\theta\,\mathrm{d}\Phi = 2\pi\int_0^{\pi}\sin\theta f_{VSF}(\lambda,\theta)\,\mathrm{d}\theta \tag{3.36}$$

Calculations of f_{VSF} require explicit solutions of Maxwell's equations in matter. A detailed theoretical derivation of scattering is given in [19]. Three major theories can be distinguished by the radius r of the scattering particles compared to the wavelength λ of radiation being scattered, which can be quantified by the dimensionless ratio $q = 2\pi r/\lambda$.

$q \ll 1$: If the dimension of scattering centers is small compared to the wavelength of the radiation, *Rayleigh theory* can be applied. It predicts a volume scattering function with a strong wavelength dependence and a relatively weak angular dependence [3]:

$$f_{VSF}(\lambda,\theta) = \frac{\pi^2(n^2-1)^2}{2N\lambda^4}(1+\cos^2\theta) \tag{3.37}$$

depending on the index of refraction n of the medium and the density N of scattering particles.
It is due to this λ^{-4} dependence of the scattering that the sky appears to be blue, compared to direct solar illumination, since short wavelengths (blue) are scattered more efficiently than the long wave (red) part of the solar spectrum. For the same reason the sun appears to be red at sunset and sunrise as the blue wavelengths have been scattered away along the optical path through the atmosphere at low angles.

$q \approx 1$: For scattering centers with sizes about the wavelength of the radiation, *Mie scatter* is the dominant process. Particles of this size act as diffractive apertures. The composite effect of all scattering particles is a complicated diffraction and interference pattern. Approximating the scattering particles by spheres, the solutions of Mie's theory are series of associated Legendre polynomials $P_l^m(\cos\theta)$, where θ is the scattering angle with respect to the initial direction of propagation. They show strong variations with the scattering angle with maximum scatter in a forward direction. The wavelength dependence is much weaker than that of Rayleigh scatter.

$q \gg 1$: Particles that can be considered macroscopic compared to the wavelength act as apertures in terms of geometric optics (Chapter 4). A particle either blocks the light if it completely reflects the radiation or it has partial transparency.

3.4.2 Volume emission

For partially transparent sources the emission of radiation is no longer a plain surface property. Volume emission cannot be separated from absorption as all radiation emitted from subsurface volume elements is subject to reabsorption along the optical path within the medium. Likewise, all subsurface layers contribute to the net radiance by their local radiance and emissivity.

Assuming that no reflection occurs at interfaces between adjacent layers within the same medium, the transmittance of a layer of thickness z (Fig. 3.11a) is given by Eq. (3.30) as $\tilde{\tau}(z) = \exp(-\alpha(\lambda)z)$. If $\tilde{\rho}(z) = 0$ the emissivity of the same layer is $\tilde{\epsilon}(z) = 1 - \tilde{\tau}(z) = 1 - \exp(-\alpha(\lambda)z)$. With

$$\frac{\mathrm{d}\tilde{\epsilon}(z)}{\mathrm{d}z} = \alpha(\lambda)\exp(-\alpha(\lambda)z) \tag{3.38}$$

the infinitesimal emissivity of a layer with thickness $\mathrm{d}z$ at depth z is given as:

$$\mathrm{d}\tilde{\epsilon} = \alpha(\lambda)\exp(-\alpha(\lambda)z)\,\mathrm{d}z \tag{3.39}$$

With this result, the net radiance leaving the surface of the medium can be computed by integrating the local radiance along the optical path, weighted by the local emissivity Eq. (3.39). For emission perpendicular to the surface the integration can be carried out along the z-direction:

$$L_\lambda = \int_0^{D_z} L_\lambda(z)\,\mathrm{d}\tilde{\epsilon} = \alpha(\lambda)\int_0^{D_z} L_\lambda(z)\exp(-\alpha(\lambda)z)\;\mathrm{d}z \tag{3.40}$$

with the diameter D_z of the object along the optical path. For $D_z \gg \alpha^{-1}$ the exponential factor approaches zero long before the upper integration limit is reached and the integration can be carried out from zero to infinity. At the surface the radiance will be partially reflected according to Eq. (3.19) and the net radiance leaving the object will be additionally reduced by the factor $\tilde{\rho}$.

Although Eq. (3.40) depends on the depth distribution of the radiance (e. g., the temperature profile in infrared thermography), two simple cases will demonstrate the basic properties of volume emission. Generally $L_\lambda(z)$ is not known *a priori* and Eq. (3.40) constitutes an ill-posed problem that is referred to in mathematics as the *inverse problem*. The depth profile of an object cannot be inferred simply from measuring its net radiance.

Example 3.4: Homogeneous radiance

For $L_\lambda(z) = L_\lambda(0)$ the integral Eq. (3.40) has the simple solution

$$L_\lambda = L_\lambda(0)\alpha(\lambda)\int_0^{D_z} \exp(-\alpha(\lambda)z)\,\mathrm{d}z = L_\lambda(0)\exp(-\alpha(\lambda)D_z) \tag{3.41}$$

For a medium with infinite thickness $D_z \gg \alpha^{-1}$ with homogeneous radiance, the net emitted radiance is the same as the radiance emitted from a surface with the radiance $L_\lambda(0)$. For a thick body with homogeneous temperature, the temperature measured by IR thermography equals the surface temperature. Thin sources ($D_z \ll \alpha^{-1}$) with homogeneous radiance behave like surface emitters with an emissivity given by the exponential factor in Eq. (3.41). For IR thermography, the absorption constant α has to be known to account for transmitted thermal radiation that does not originate from the temperature of the body (Fig. 3.3).

Example 3.5: Linear radiance profile

For a linear radiance profile, $L_\lambda(z) = L_\lambda(0) + az$, $a = \mathrm{d}L_\lambda/\mathrm{d}z$, the integral Eq. (3.40) yields

$$\begin{aligned} L_\lambda &= \alpha(\lambda)\int_0^{\infty} (L_\lambda(0) + az)\exp(-\alpha(\lambda)z)\,\mathrm{d}z \\ &= L_\lambda(0) + \frac{a}{\alpha(\lambda)} = L_\lambda(\alpha^{-1}(\lambda)) \end{aligned} \tag{3.42}$$

For a medium with infinite thickness $D_z \gg \alpha^{-1}$ with a linear radiance profile, the net emitted radiance equals the radiance emitted from a subsurface element at depth $z = \alpha^{-1}$. For infrared thermography, the measured temperature is not the surface temperature but the temperature in a depth corresponding to the penetration depth of the radiation. As the absorption coefficient α can exhibit strong variability over some orders of magnitude within the spectral region of a thermography system, the measured radiation originates from a mixture of depth layers. An application example is IR thermography to measure the temperature gradient at the ocean surface (detailed in Volume 3, Chapter 35 and [20]).

3.4.3 Luminescence

Luminescence describes the emission of radiation from materials by radiative transition between an excited state and a lower state. In a complex molecule, a variety of possible transitions between states exist and not all are optical active. Some have longer lifetimes than others, leading to a delayed energy transfer. Two main cases of luminescence are classified by the time constant of the process.

Fluorescence. *Fluorescence*, by definition, constitutes the emission of electromagnetic radiation, especially of visible light, stimulated in a substance by the absorption of incident radiation and persisting only as long as the stimulating radiation is continued. It has short lifetimes, that is, the radiative emission occurs within 1–200 ns after the excitation.

Phosphorescence. *Phosphorescence* defines a delayed luminescence, occurring milliseconds to minutes after the excitation. Prominent examples of such materials are watch displays or light switches that glow in the dark. The intensity decreases as the time from the last exposure to light increases.

There are a variety of physical and chemical processes leading to a transition between molecular states. A further classification of luminescence accounts for the processes that lead to excitation:

- *Photoluminescence*: Excitation by absorption of radiation (photons);
- *Electroluminescence*: Excitation by electric current (in solids and solutions) or electrical discharge (in gases);
- *Thermoluminescence*: Thermal *stimulation* of the emission of already excited states;
- *Radioluminescence*: Excitation by absorption of ionizing radiation or particle radiation;
- *Chemoluminescence*: Excitation by chemical reactions; and
- *Bioluminescence*: Chemoluminescence in living organisms; prominent examples include fireflies and marine organisms.

For practical usage in computer vision applications, we have to consider how luminescence can be used to visualize the processes or objects of interest. It is important to note that fluorescent intensity depends on both the concentration of the fluorescent material as well as on the mechanism that leads to excitation. Thus, fluorescence allows us to visualize *concentrations* and *processes quantitatively*.

The most straightforward application can be found in biology. Many biological processes are subject to low-level bioluminescence. Using appropriate cameras, such as amplified intensity cameras (Chapter 5), these processes can be directly visualized (Chapter 12). An application example is the imaging of Ca^{2+} concentration in muscle fibers, as will be outlined in (Volume 3, Chapter 34).

Other biochemical applications make use of fluorescent markers. They use different types of fluorescent dyes to mark individual parts of chromosomes or gene sequences. The resulting image data are multispectral confocal microscopic images (Volume 3, Chapters 40 and 41) encoding different territories within the chromosomes).

Fluorescent dyes can also be used as tracers in fluid dynamics to visualize flow patterns. In combination with appropriate chemical tracers, the fluorescence intensity can be changed according to the relative concentration of the tracer. Some types of molecules, such as oxygen, are very efficient in deactivating excited states during collision without radiative transfer—a process referred to as *fluorescence quenching*. Thus, fluorescence is reduced proportional to the concentration of the quenching molecules. In addition to the flow field, a quantitative analysis of the fluorescence intensity within such images allows direct measurement of trace gas concentrations (Volume 3, Chapter 30).

3.5 References

[1] Gaussorgues, G., (1994). *Infrared Thermography*. London: Chapmann & Hall.

[2] Siegel, R. and Howell, J. R. (eds.), (1981). *Thermal Radiation Heat Transfer*, 2nd edition. New York: McGraw-Hill Book, Co.

[3] McCluney, W. R., (1994). *Introduction to Radiometry and Photometry*. Boston: Artech House.

[4] CIE, (1987). *CIE International Lighting Vocabulary*. Technical Report.

[5] Kirchhoff, G., (1860). *Philosophical Magazine and Journal of Science*, **20(130)**.

[6] Nicodemus, F. E., (1965). Directional reflectance and emissivity of an opaque surface. *Applied Optics*, **4**:767.

[7] Wolfe, W. L. and Zissis, G. J. (eds.), (1989). *The Infrared Handbook*, 3rd edition. Michigan: The Infrared Information Analysis (IRIA) Center, Environmental Research Institute of Michigan.

[8] Jähne, B., (1997). *Handbook of Digital Image Processing for Scientific Applications*. Boca Raton, FL: CRC Press.

[9] Dereniak, E. L. and Boreman, G. D., (1996). *Infrared Detectors and Systems*. New York: John Wiley & Sons, Inc.

[10] Arney, C. M. and Evans, C. L., Jr., (1953). *Effect of Solar Radiation on the Temperatures in Metal Plates with Various Surface Finishes*. Technical Report.

[11] Merrit, T. P. and Hall, F. F., (1959). Blackbody radiation. *Proc. IRE*, **47(2)**: 1435–1441.

[12] Hecht, E. and Zajac, A., (1977). *Optics*, 2nd edition. Addison-Wesley World Student Series. Reading, MA: Addison-Wesley Publishing.

[13] Bennet, J. M. and Mattsson, L. (eds.), (1989). *Introduction to Surface Roughness and Scattering*. Washington, DC: Optical Society of America.

[14] Foley, J. D., van Dam, A., Feiner, S. K., and Hughes, J. F., (1990). *Computer Graphics, Principles and Practice*, 2nd edition. Reading, MA: Addison-Wesley.

[15] Cook, R. and Torrance, K., (1982). A reflectance model for computer graphics. *ACM TOG*, **1(1)**:7–24.

[16] Phong, B.-T., (1975). Illumination for computer generated pictures. *CACM*, **6**:311–317.

[17] Bass, M., Van Stryland, E. W., Williams, D. R., and Wolfe, W. L. (eds.), (1995). *Handbook of Optics. Fundamentals, Techniques, and Design*, 2nd edition, Vol. 1. New York: McGraw-Hill.

[18] Harris, D. C., (1994). *Infrared Window and Dome Materials*. Bellingham, WA: SPIE Optical Engineering Press.

[19] van de Hulst, H. C., (1981). *Light Scattering by Small Particles*. New York: Dover Publications.

[20] Haussecker, H., (1996). *Messung und Simulation von kleinskaligen Austauschvorgängen an der Ozeanoberfläche mittels Thermographie*. Dissertation, Universität Heidelberg.

4 Imaging Optics

Peter Geißler

Interdisziplinäres Zentrum für Wissenschaftliches Rechnen (IWR)
Universität Heidelberg, Germany

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

4.1 Introduction

Computer vision and image processing always start with image acquisition, mostly done by illuminating the scene with natural or artificial light in the visible range and taking images with a photographic lens. The importance of proper image acquisition is ignored in many applications, at the expense of an increased effort in the processing of the images. In addition to the fact that appropriate visualization can enhance image quality in such a manner that image processing requires fewer processing steps, becomes much faster, or is even for the first time possible, image degradations caused by unsuitable imaging may seriously complicate image analysis or even be uncorrectable afterwards. Although most of today's camera lenses are of very good quality, they are always optimized for a particular purpose and may fail if used in other setups. In addition, in some applications an optics setup from one or two simple lenses may provide better image quality than stock lenses because the setup can be optimized exactly for that imaging problem. For these reasons, this chapter will provide the reader with the essential concepts of optical imaging, focusing on the geometric ray approximation that will be sufficient for most applications besides microscopic imaging. Special emphasis is placed on the description of nonparaxial optics (the main reason for image distortions).

4.2 Basic concepts of geometric optics

Basic to geometric optics are light rays, which can be seen as an approximation of a parallel wavefront of zero cross section. Therefore, rays are always perpendicular to the wavefront, as can be seen in Fig. 4.1 for the fundamental cases of spherical and planar wavefronts. In a homogeneous dielectric medium, a ray travels with the local speed of light and is denoted by c/n; c denotes the vacuum light speed, and n is the refractive index of the dielectric medium and depends on the medium and the wavelength. These figures illustrate another commonly used technique in ray optics—the representation of light intensity by the density of the rays. Of course, rays represent an abstraction from wave optics that neglects diffraction effects.

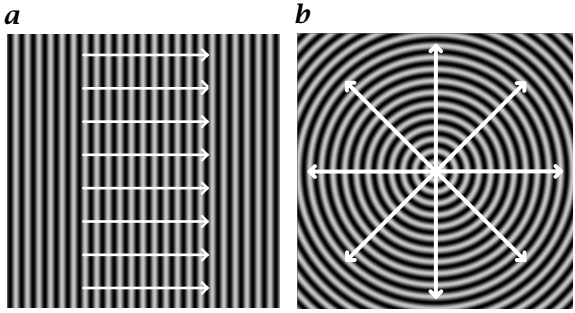


Figure 4.1: ***a*** *Planar wavefront and its ray representation;* ***b*** *circular wavefront and its ray representation.*

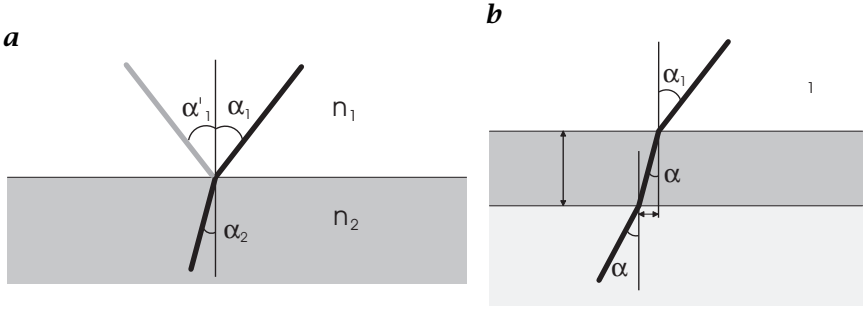


Figure 4.2: ***a*** *Snellius' law of refraction;* ***b*** *refraction at a three-media transition.*

4.2.1 Reflection and refraction

Within a medium of constant index of refraction, a ray travels as a straight line without any changes in its direction. A ray passing through the boundary surface of two media of different index of refraction is bent by an angle described by the law of Snellius (Eq. (4.1)). It relates the ratio of the incoming and outgoing deviation angles to the ratio of the refractive indices.

$$n_1 \sin \alpha_1 = n_2 \sin \alpha_2 \tag{4.1}$$

Besides refraction into the adjacent medium, reflection of the incoming ray occurs. In this case the simple relation $\alpha_1 = \alpha_2$ applies.

It is useful in many cases to express both refraction and reflection as vector equations. We specify the direction of the incoming ray by the unit vector $\bar{\boldsymbol{r}}$, the direction of the outgoing ray again by the unit vector $\bar{\boldsymbol{r}}'$, and the vector normal to the surface dividing the two media by the unit vector $\bar{\boldsymbol{n}}$. Then reflection can be written as

$$\bar{\boldsymbol{r}}' = \bar{\boldsymbol{r}} - 2(\bar{\boldsymbol{n}}\bar{\boldsymbol{r}})\bar{\boldsymbol{n}} \tag{4.2}$$

whereas refraction reads

$$\bar{\boldsymbol{r}}' = \frac{1}{n_a/n_e}\bar{\boldsymbol{r}} - \left[\frac{\bar{\boldsymbol{n}}\bar{\boldsymbol{r}}}{n_a/n_e} + \sqrt{1 - \frac{\left(1 + (\bar{\boldsymbol{n}}\bar{\boldsymbol{r}})^2\right)}{(n_a/n_e)^2}}\right]\bar{\boldsymbol{n}} \tag{4.3}$$

4.2.2 Multimedia refraction

Often not only does a single change of the refractive index have to be taken into account, but also a sequence of consecutive phase transitions. This is the case, for example, in any underwater optics, where a glass plate protects the optics from the aqueous medium. This situation is illustrated in Fig. 4.2b. Fortunately, Snellius' law remains valid between the media n_1 and n_3

$$\frac{\sin\alpha_1}{\sin\alpha_3} = \frac{\sin\alpha_1}{\sin\alpha_2}\frac{\sin\alpha_2}{\sin\alpha 3} = \frac{n_2}{n_1}\frac{n_3}{n_1} = \frac{n_3}{n_1} \tag{4.4}$$

Because of the optical path length within the medium n_2, the ray is shifted in parallel by

$$d = D\tan\alpha_2 \tag{4.5}$$

4.2.3 Paraxial optics

From the Taylor series of the trigonometric functions, their corresponding small angle approximation is found to be

$$\sin(\alpha) = \alpha - \frac{\alpha^3}{3!} + \frac{\alpha^5}{5!}... \approx \alpha \tag{4.6}$$

$$\cos(\alpha) = 1 - \frac{\alpha^2}{2!} + \frac{\alpha^4}{4!}... \approx 1 \tag{4.7}$$

$$\tan(\alpha) = \alpha + \frac{\alpha^3}{3} + \frac{2\alpha^5}{15} + ... \approx \alpha \tag{4.8}$$

These rays form the *paraxial domain*, where the approximations in Eq. (4.8) can be applied with acceptable deviations. It is important to notice that there is no clear definition of the paraxial domain as its boundaries depend on the maximum error that is tolerated. Figure 4.3 shows the relative angular error of the paraxial approximation.

In paraxial approximation, Snellius simplifies to

$$n_1\alpha_1 = n_2\alpha_2 \tag{4.9}$$

This linear equation is much easier than the correct Eq. (4.1), which contains the trigonometric terms. Unless indicated otherwise, all calculations of geometric optics in this chapter are done using the paraxial approximation. Its power will be shown first in the description of

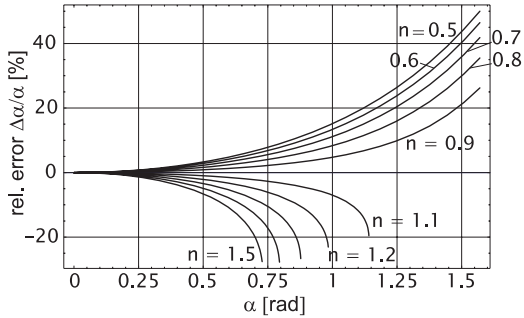


Figure 4.3: *Relative angular error of the paraxial approximation for various values of the ratio of refractive indices* $n = n_1/n_2$.

lenses, from spherical lenses to the approximation of thin, paraxial lenses, which is sufficient in most cases. Deviations from the paraxial domain will be discussed with the lens aberrations in Section 4.5.

4.3 Lenses

All imaging optics uses lenses as central imaging elements. Therefore it is important to examine the optical properties of these fundamental elements. We start with spherical lenses, which have only one kind of glass. Despite the fact that spherical lenses do not best approximate the ideal paraxial lens, they are the most common kind of lenses used. This is due to the fact that it is easier to manufacture spherical surfaces than it is to polish aspherical surfaces. Therefore, it is more economical in most cases to use systems of spherical surfaces and lenses in order to correct lens aberrations than to use aspherical lenses. Nevertheless, new technologies in the pressing of plastic lenses have made the production of aspherical lenses inexpensive.

4.3.1 Definitions

Lenses can be described by means of a set of cardinal points and surfaces. This method also works for systems of lenses and other refracting surfaces, that is, it is commonly used to describe any optical system. The basic terms and definitions are as follows:

Optical Axis The optical axis is the main axis of the optics, usually denoted as z-direction. For a typical system of centered and axial symmetric elements, the optical axis is the axis of symmetry of the optics. Usually it coincides with the main direction of light propagation. Points located on the optical axis and elements centered

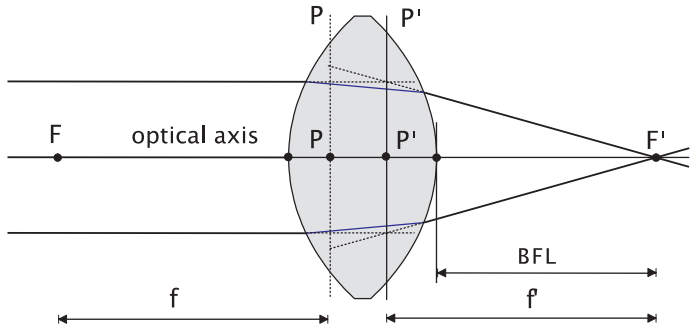


Figure 4.4: *Fundamental terms of the paraxial description of lenses.*

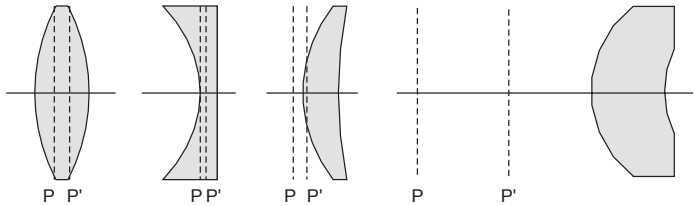


Figure 4.5: *Position of principal planes for different lens types.*

around it are called *on-axis*, otherwise denoted as *off-axis*. Mirrors can fold the linear axis into a set of piecewise linear sections.

Cardinal Planes Refraction on the lens surfaces can be described by the concept of the *principal planes*, without having to take into account the exact radius of curvature. Extended towards the lens interior, the incoming and the outgoing rays intersect at a point on the *principal surface*. The projection of the intersection point onto the optical axis is called the corresponding *principal point*. In paraxial approximation the generally bent principal surface becomes flat, forming the *principal plane*. All principal points than merge into a single one. The principal planes allow for the graphical construction of ray paths, as will be explained in detail in Section 4.3.5.
It is important to note that the principal planes are not necessarily located within the lens itself (Fig. 4.5). This is often used to extend the optical length of compact telephoto lenses.

Focal Length Within the paraxial domain, all incident rays entering parallel to the optical axis intersect at an on-axis point behind the lens, the *back focal point* (BFP) F'. Due to the reversibility of the ray paths, rays emerging from the *front focal point* (FFP) F run parallel to the axis after passing the lens. Rays emerging from off-axis points on the *focal plane* still form a parallel ray bundle, but are

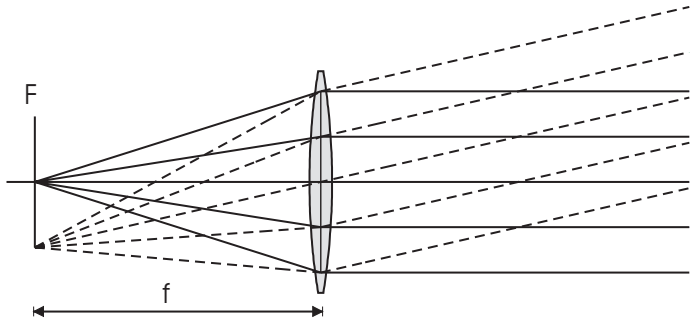


Figure 4.6: *Bundles of parallel rays emerging from object points on the focal plane.*

now nonparallel to the optical axis. The distance from the FFP to the front principal plane gives the *effective focal length* (EFL) of the lens. The front EFL equals the back EFL, provided that there is no change in refractive index. A change in refractive index from n_1 in front of the lens to n_2 behind the lens changes the back EFL f' to $n_2/n_1 f$. Therefore, the EFL in air is often referred to as the *focal length* of the lens. Additionally, the distance between the focal points and the lens vertices are called the *front focal length* (FFL) and *back focal length* (BFL) , respectively; they equal each other only for symmetric lenses.

4.3.2 Spherical lenses

A spherical lens can be seen as two spherical surfaces with a medium of a constant index of refraction between them. To understand the behavior of these lenses, it is sufficient to analyze one of the surfaces. As illustrated in Fig. 4.7, a ray emerging from an on-axis object point O_1 intersects the optical axis at a point O_2 behind the spherical surface. Within the paraxial domain, all rays emerging from an object point intersect in one point in the image space. Thus, we say the object point is imaged onto its *optical conjugate* image point. The distances d_1 and d_2 of object and image points are correlated with the radius of curvature R of the surface and the indices of refraction n_1 and n_2 by Eq. (4.10).

$$\frac{n_2}{d_2} - \frac{n_1}{d_1} = \frac{n_2 - n_1}{R} \tag{4.10}$$

Written in an alternative form

$$n_1 \left(\frac{1}{R} - \frac{1}{d_1} \right) = n_2 \left(\frac{1}{R} - \frac{1}{d_2} \right) \tag{4.11}$$

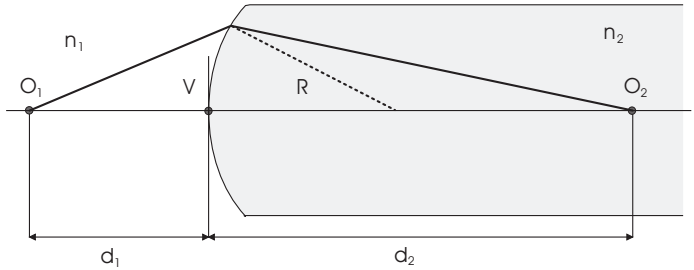


Figure 4.7: *Path of rays at a single spherical surface.*

Equation (4.10) separates object and image space. Equation (4.11) is known as *Abbe's invariant*.

A single surface separating regions of different refractive index is therefore sufficient to form an imaging optics, and can therefore be seen as the simplest possible lens. For every lens, focal length and principal planes can be used in order to describe paraxial properties. Setting either of the distances d_1 or d_2 to infinity yields both focal lengths

$$f_1 = R\frac{n_2}{n_2 - n_1} \qquad f_2 = -R\frac{n_1}{n_2 - n_1} \tag{4.12}$$

and

$$f_1 + f_2 = R \qquad n_1 f_1 = -n_2 f_2 \tag{4.13}$$

Both principal planes coincide at the location of the vertex V.

At present, a lens consists of two spherical surfaces, thereby enclosing the lens material. Using ray calculations similar to those for a single surface, without giving details of the calculations, the paraxial properties of the lens are obtained. We restrict ourselves to the commonly used case of a lens in air, thus the refractive indices of the surrounding medium become $n_1 = n_2 = 1$. With $D = V_1V_2$ denoting the thickness of the lens, n_l its refractive index, and R_1 and R_2 the radii of curvature of its surfaces, the lens data calculates to

$$f = \frac{1}{n_l - 1}\frac{n_l R_1 R_2}{(n_l - 1)d + n_l(R_1 + R_2)} \tag{4.14}$$

$$\nu_1 = -\frac{R_2 D}{(n_l - 1)d + n_l(R_1 + R_2)} \tag{4.15}$$

$$\nu_2 = -\frac{R_1 D}{(n_l - 1)d + n_l(R_1 + R_2)} \tag{4.16}$$

$$h = D(1 - \frac{R_2 - R_1}{(n_l - 1)d + n_l(R_1 + R_2)}) \tag{4.17}$$

where $h = P_1P_2$ denotes the distance between the principal planes, and $\nu_i = V_iP_i$ is the distance to the corresponding vertices. Because of the assumption of an identical refractive index on both sides of the lens, the front and back focal lengths of the lens coincide with the focal length f.

4.3.3 Aspherical lenses

Although they are the most popular lens type, spherical lenses are subject to certain limitations. For example, focusing of parallel ray bundles onto the focal point only works within the narrow paraxial domain. Non-spherically shaped surfaces allow lenses to be customized for specific purposes, for example, for optimal focusing, without the restriction to the paraxial domain. Typically, there are three types of aspherical surfaces:

Rotational symmetric surface. This type of surface is still rotationally symmetric to an axis, which usually coincides with the optical axis. Aspherical lenses are the most common type used for the correction of ray aberrations, which cannot be avoided. This type of surface can be described in terms of a curvature $C = 1/R$ and the *conic constant* K

$$z = \frac{Cx^2}{1 + \sqrt{1 - (K+1)C^2x^2}} + \sum_{i=1}^{\infty} \alpha_{2i} x^{2i} \tag{4.18}$$

wherein the first term describes conic sections, and the second term higher-order deformations. As illustrated in Table 4.1, the conic constant controls the shape of the surface.

Aspherical lenses with conic surfaces are often used to extend ideal ray paths beyond the paraxial domain. These lenses do not satisfy the paraxial equations in any case, but have to be designed for the exact purpose for which they are intended. As an example, Fig. 4.8 shows a hyperbolic lens, which is designed for perfect focusing. If used for imaging with noninfinite distances, strong aberrations occur.

Toroidal lenses. Toroidal surfaces are spherical in two principal sections, which are perpendicular to each other. The radii of curvature differ between the two sections. The particular case of one of the curvatures is infinity, which results in *cylindrical lenses*. As an example of the use of toroidal lenses, two crossed cylindrical lenses of different focal length can be used to achieve different magnifications in sagittal and meridional sections. This *anamorphic imaging* is illustrated in Fig. 4.9.

Freeform surfaces. Arbitrarily formed surfaces are used only for special applications and shall not be discussed herein.

Table 4.1: *Conic surfaces*

Conic constant	Surface type	Illustration
$K < -1$	Hyperboloid	
$K = -1$	Paraboloid	
$-1 < K < 0$	Ellipsoid	
$K = 0$	Sphere	
$K > 0$	Ellipsoid	

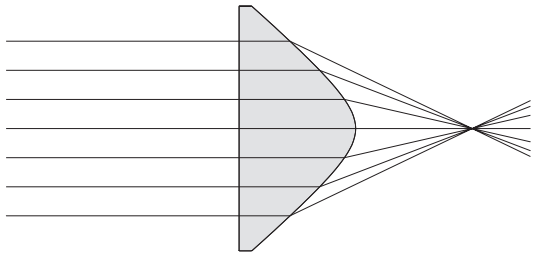


Figure 4.8: *Perfect focusing outside the paraxial domain by an aspherical condenser lens.*

4.3.4 Paraxial lenses

If the distance between the lens vertices (the lens thickness) can be neglected, the principal planes and the nodal planes converge onto a single plane, located at the lens position. Further restricting the rays to the paraxial domain, the lens can be described by a single parameter, its focal length. This is called the *thin paraxial lens*, which is used widely in order to gain first-order approximations of the behavior of the optics. Above all, paraxial lens equations are most powerful in the first step of optics design, where its constraints can be established without the details of physical lenses. In many cases, paraxial lenses can describe the optics adequately. Especially, optimal distances of object and image, depth-of-field and other basic properties of an imaging optics can be estimated with this approach.

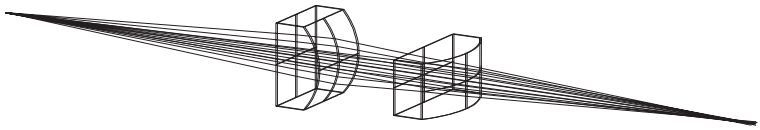


Figure 4.9: *Principle of anamorphic imaging.*

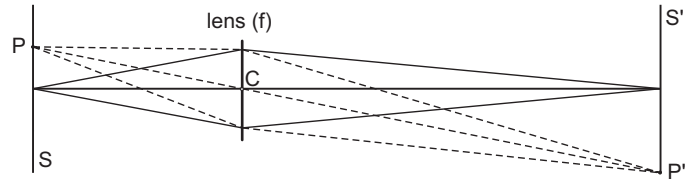


Figure 4.10: *Optical conjugates of a paraxial lens.*

With a thin paraxial lens, all rays emerging from a point P intersect at its *conjugate point* P' behind the lens. Because all rays meet at exactly the same point, the lens is *aberration-free*. Furthermore, because of the restriction to the paraxial domain, a plane S perpendicular to the optical axis is also imaged into a plane S'. Again, S' is called the optical conjugate of S. If the object point is at infinity distance to the lens, its conjugate is located on the focal plane. Therefore, rays intersecting the lens center C are not changed.

In most optical systems, compared to a single lens, several lenses are used to improve the image quality. First, we introduce the extension of the thin paraxial lens toward the thick paraxial lens, where the lens thickness is taken into account. It can be shown that this lens can equivalently be seen as the combination of two thin paraxial lenses. This will lead to a general method to describe arbitrary paraxial systems by a single paraxial lens.

4.3.5 Thick lenses

If the thickness of a lens cannot be neglected, the concept of the paraxial lens has to be extended towards *thick paraxial lenses*. In this case, the two principal planes no longer converge to a single plane, but are separated by an equivalent distance, the *nodal space*. As a general rule, for lenses in air the nodal space is approximately one-third of the lens thickness [1]. As illustrated in Fig. 4.11, rays can be constructed by elongation of the unrefracted ray towards the first principal plane P, traversing the ray parallel to the optical axis to the second principal plane, and continuing to the conjugate point P'. For geometric construction of ray paths, rays in between the principal planes are always parallel to the axis. As a consequence, axis-parallel rays are deviated

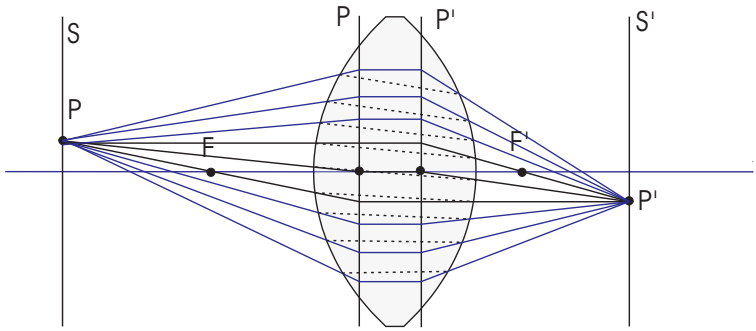


Figure 4.11: *Ray paths for a thick paraxial lens. Dashed lines show the physical ray paths; solid lines show the virtual rays used for construction of ray paths.*

at the principal plane near the corresponding focal point, and rays intersecting a principal point emerge from the conjugate principal point, maintaining the same angle to the axis as the incident ray. In other words, the nodal points coincide with the principal points.

4.3.6 Systems of lenses

A complex optical system consists of several thick lenses. A pair of thick lenses, described by the set of four principal planes and two focal points, can be converted into a new equivalent lens, with two principal planes and one focal length. Applying this recursively to the lens system, the complete setup can be condensed into one thick lens. Within the paraxial domain, this powerful approach facilitates dealing with optics of high complexity. Figure 4.12 illustrates the equivalent principal planes of the two-lens system; P_{11} and P_{12} are the principal planes of the first lens, and P_{21} and P_{22} are the principal planes of the second lens.

The position p_i of the principal planes and the effective focal length of the equivalent system, provided the lenses are used in air (n=1), are given by

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 f_2} \tag{4.19}$$

$$p_1 = \overline{P_{11}P_1} = \frac{fd}{f_2} \tag{4.20}$$

$$p_2 = \overline{P_{22}P_2} = -\frac{fd}{f_1} \tag{4.21}$$

$$p = \overline{P_1P_2} = -\frac{fd^2}{f_1 f_2} \tag{4.22}$$

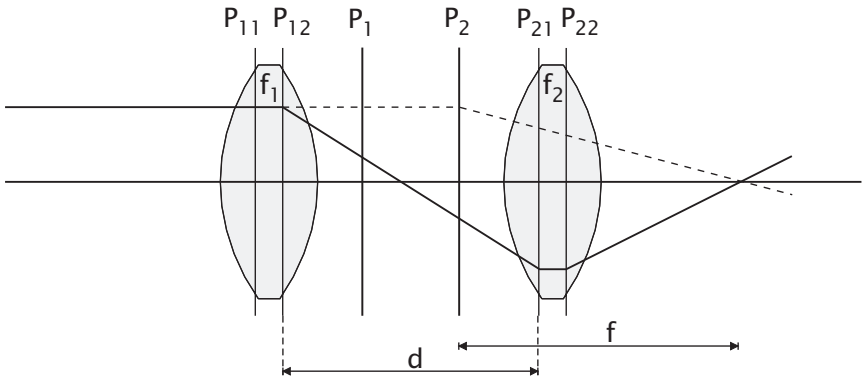


Figure 4.12: *A system of thick lenses and its equivalent thick lens.*

Table 4.2: *Overview of the most important parameters of the combined lens and the order of the cardinal planes in case of $d, f_1, f_2 > 0$; L_i indicates the position of lens i*

Focal length	$d < f_1 + f_2$	$d > f_1 + f_2$
	$f > 0$	$f < 0$
p_1	$p_1 > 0$	$p_1 < 0$
p_2	$p_2 < 0$	$p_2 > 0$
	$\|\nu_1\| + \|\nu_2\| > d$	$\|\nu_1\| + \|\nu_2\| < d$
Relative position	P_1 is behind P_2	P_1 is in front of P_2
Order	$f_1 \le d, f_2 \le d \rightarrow P_2 L_1 L_2 P_1$	
of	$f_1 \le d, f_2 \ge d \rightarrow P_2 L_1 P_1 L_2$	$P_1\ L_1\ L_2\ P_2$
cardinal	$f_1 \ge d, f_2 \le d \rightarrow L_1 P_2 L_2 P_1$	
planes	$f_1 \ge d, f_2 \ge d \rightarrow L_1 P_2 P_1 L_2$	

The cardinal planes can occur in any order, for example, it is common that the order of the principal planes P_1 and P_2 becomes reversed with lenses located closely together. Table 4.2 gives an overview of the order of the cardinal planes of a system of two lenses of positive focal length.

4.3.7 Matrix optics

Tracing rays through an optical system allows for in-depth analysis of the optics, taking into account all surfaces and materials. An elegant method to describe ray propagation between the surfaces of the system has been introduced by T. Smith [2]. Within the paraxial domain, it is possible to describe both refraction and ray propagation by simple matrix operations. The ray tracing can be achieved by matrix

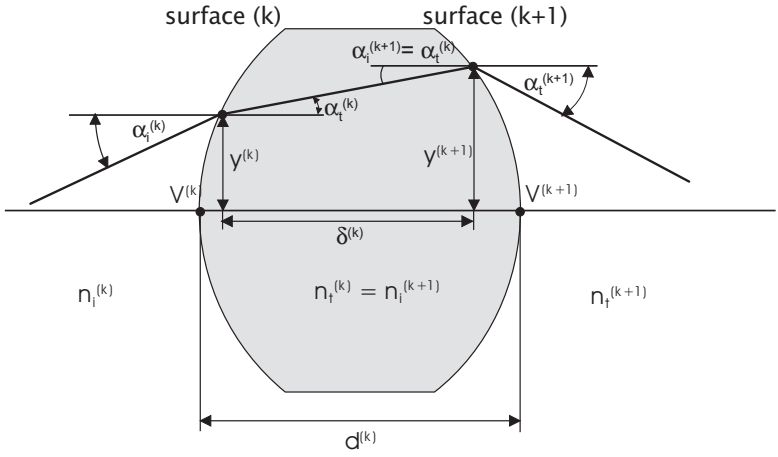


Figure 4.13: *Notation used for the matrix optic calculations.*

multiplication of the matrices describing the optical elements and their distances. In order to describe this method, all surfaces are numbered consecutively from left to right and are denoted by superscripts. Rays incoming to a surface are denoted by i; outgoing rays are denoted by t. The notation is illustrated in Fig. 4.13.

Vector notation for rays. A ray of angle α and distance y with respect to the optical axis is denoted by the vector

$$\boldsymbol{r} = \begin{pmatrix} n\alpha \\ y \end{pmatrix} \tag{4.23}$$

Refraction at a single surface. Refraction of a ray of incident angle n_i and distance y_i to the optical axis can be written using the power $\mathcal{D}$ of a single surface

$$n_t^{(k)}\alpha_t^{(k)} = n_i^{(k)}\alpha_i^{(k)} - \mathcal{D}^{(k)} y_i^{(k)} \tag{4.24}$$

$$y_t^{(k)} = y_i^{(k)} \tag{4.25}$$

$$\mathcal{D}^{(k)} = \frac{n_t^{(k)} - n_i^{(k)}}{R^{(k)}} \tag{4.26}$$

Equation (4.27) can be rewritten as a matrix equation

$$\boldsymbol{r}_t^{(k)} = \begin{pmatrix} n_t^{(k)}\alpha_t^{(k)} \\ y_t^{(k)} \end{pmatrix} = \begin{pmatrix} 1 & -\mathcal{D}^{(k)} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} n_i^{(k)}\alpha_i^{(k)} \\ y_i^{(k)} \end{pmatrix} =: \mathcal{R}^{(k)} \boldsymbol{r}_i^{(k)} \tag{4.27}$$

whereas the matrix $\mathcal{R}^{(k)}$ denotes the *refraction matrix* of the surface (k).

Ray propagation. The propagation of a ray between two consecutive surfaces (k) and $(k+1)$ is linear due to the fact that no change in the refractive index can occur. Therefore replacing the true distance $\delta^{(k)}$ by its paraxial approximation $d^{(k)}$ yields $y_i^{(k+1)} = d^{(k)}\alpha_t^{(k)} + y_t^{(k)}$, and thus ray propagation towards the next surface can be expressed by the *transfer matrix* $\mathcal{T}^{()}$

$$\boldsymbol{r}_i^{(k+1)} = \begin{pmatrix} n_i^{(k+1)}\alpha_i^{(k+1)} \\ y_i^{(k+1)} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{d^{(k)}}{n_t^{(k)}} & 1 \end{pmatrix} \begin{pmatrix} n_t^{(k)}\alpha_t^{(k)} \\ y_t^{(k)} \end{pmatrix} =: \mathcal{T}^{(k)}\boldsymbol{r}_t^{(k)} \tag{4.28}$$

System matrix. Now refraction at single surfaces (Eq. (4.27)) is combined with ray propagation between two surfaces (Eq. (4.28)) to grasp the behavior of a lens consisting of two surfaces. A ray emerging from the second lens surface can be calculated from the incident ray by applying the refraction matrix of the first surface, the transfer matrix between the surfaces, and finally the refraction matrix of the second surface. This is done by simple matrix multiplication:

$$\boldsymbol{r}_t^{(k+1)} = \mathcal{R}^{(k+1)}\mathcal{T}^{(k)}\mathcal{R}^{(k)}\boldsymbol{r}_i^{(k)} \tag{4.29}$$

The system matrix of the optical element is defined as

$$\mathcal{S}^{(k+1,k)} = \mathcal{R}^{(k+1)}\mathcal{T}^{(k)}\mathcal{R}^{(k)} \tag{4.30}$$

It transforms an incident ray at the first surface (k) to an emerging ray at the next surface $(k+1)$. In general, any optical element with an arbitrary number of surfaces is described by a single system matrix. Assuming N surfaces, the system matrix is denoted $\mathcal{S}^{(N,1)}$ in order to indicate the number of surfaces. It is given by

$$\mathcal{S}^{(N,1)} = \mathcal{R}^{(N)}\mathcal{T}^{(N-1)}\mathcal{R}^{(N-1)}\ldots\mathcal{T}^{(1)}\mathcal{R}^{(1)} = \mathcal{R}^{(N)}\prod_{k=1}^{N-1}\mathcal{T}^{(k)}\mathcal{R}^{(k)} \tag{4.31}$$

Equation (4.31) can be split at any surface (k) between the first and the last and rewritten as

$$\mathcal{S}^{(N,1)} = \mathcal{S}^{(N,k)}\mathcal{T}^{(k-1)}\mathcal{S}^{(k-1,1)} \qquad \text{with} \quad 1 < k < N \tag{4.32}$$

Equation (4.32) makes it easy to combine optical elements into more and more complex optical systems by reusing the known system matrices of the simpler elements.

Table of system matrices. The system matrix is the fundamental description of optical elements, and therefore is the basis of matrix optics calculation. Table 4.3 provides an overview of the most important

Table 4.3: *System matrices for various optical elements*

Optics	System Matrix
Straight section	$\begin{pmatrix} 1 & 0 \\ \frac{d}{n} & 1 \end{pmatrix}$
Dielectric interface	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
Spherical dielectric interface	$\begin{pmatrix} 1 & -\mathcal{D} \\ 0 & 1 \end{pmatrix}$
Spherical lens	$\begin{pmatrix} 1-\frac{d}{n}\mathcal{D}^{(2)} & \frac{d}{n}\mathcal{D}^{(1)}\mathcal{D}^{(2)} - (\mathcal{D}^{(1)}+\mathcal{D}^{(2)}) \\ \frac{d}{n} & 1-\frac{d}{n}\mathcal{D}^{(2)} \end{pmatrix}$
Plate in air	$\begin{pmatrix} 1 & 0 \\ \frac{d}{n} & 1 \end{pmatrix}$
Thin lens in air	$\begin{pmatrix} 1 & -1/f \\ 0 & 1 \end{pmatrix}$
Thick lens in air	$\begin{pmatrix} 1-\frac{p_1}{f} & -\frac{1}{f} \\ \frac{p_1 p_2}{f}+p_1-p_2 & 1+\frac{p_2}{f} \end{pmatrix}$
Two thin lenses in air	$\begin{pmatrix} 1-d/f_2 & 1/f \\ d & 1-d/f_1 \end{pmatrix}$
Spherical mirror	$\begin{pmatrix} 1 & -\frac{2}{R} \\ 0 & 1 \end{pmatrix}$

system matrices of simple optical elements consisting of two surfaces. Elements of higher complexity can be calculated according to Eq. (4.32). To simplify notation, the index of refraction of the lens material is denoted by n, and the thickness of the lens is denoted by d.

4.4 Optical properties of glasses and other materials

4.4.1 Dispersion

Glasses and other material are characterized mainly by two properties: refractive index and dispersion. Dispersion means that the refractive index depends on the wavelength of the light. Therefore, in order to describe the refractive properties of any material, the dispersion curve $n(\lambda)$ has to be given. In practice, the refractive index is given only for a number of standardized wavelengths. These wavelengths correspond to spectral lines of specific chemical elements in which wavelengths are known with great precision. A table of the widely used wavelengths,

Table 4.4: *Most important Fraunhofer spectral lines*

Symbol	Wavelength [nm]	Color	Element
i	365.0	UV	Hg
h	404.7	violet	Hg
g	435.8	blue	Hg
F'	480.0	blue	Cd
F	486.1	blue/green	H
e	546.1	yellow/green	Hg
d or D_3	587.6	orange	He
D_2	589.0	orange	Na
D	589.3	orange	Na
D_1	589.6	orange	Na
C'	643.8	orange	Cd
C	656.3	red	H
r	706.5	red	He
A'	768.2	red	K

together with their international symbol and the chemical element from which they arise, are given in Table 4.4.

For any other wavelengths in the visible, near UV and in the near IR range, the refractive index can be calculated by several common interpolation formulas. The most widely used are summarized in Table 4.5. The coefficients needed for the formulas are available in the glass catalogs of all major glass manufacturers, such as Schott [3]. It is often recommended to check the exact definitions of the formulas used before inserting coefficients from glass catalogs. This is because the formulas are often slightly modified by the manufacturers.

4.4.2 Glasses and plastics

In many cases, it is not necessary to know the complete dispersion relation $n(\lambda)$. Instead, a usable and short characterization of the glass is more useful. Usually, the *main refractive index* is employed as a characterization of the glass. It is defined as the refractive index at the wavelength λ_d or λ_e according to Table 4.4. As a code for the dispersion, *Abbe number* is widely used. Two definitions according to the use of either n_e or n_d as the main refractive index are common:

$$V_d = \frac{n_d - 1}{n_F - n_C} \qquad V_e = \frac{n_e - 1}{n_{F'} - n_{C'}} \tag{4.33}$$

Table 4.5: *Dispersion formulas for glasses*

Name	Formula
Schott[1]	$n(\lambda) = a_0 + a_1\lambda^2 + a_2\lambda^{-2} + a_3\lambda^{-4} + a_4\lambda^{-6} + a_5\lambda^{-8}$
Sellmeier 1	$n^2(\lambda) = 1 + \frac{K_1\lambda^2}{\lambda^2 - L_1} + \frac{K_2\lambda^2}{\lambda^2 - L_2} + \frac{K_3\lambda^3}{\lambda^3 - L_3}$
Sellmeier 2	$n^2(\lambda) = 1 + A + \frac{B_1\lambda^2}{\lambda^2 - \lambda_1^2} + \frac{B_2\lambda^2}{\lambda^2 - \lambda_2^2}$
Herzberger[2]	$n(\lambda) = A + BL(\lambda) + CL^2(\lambda) + D\lambda^2 + E\lambda^4 + F\lambda^4$ with $L(\lambda) = \frac{1}{\lambda^2 - 0.028)}$
Conrady[3]	$n(\lambda) = n_0 + \frac{A}{\lambda} + \frac{B}{\lambda^{3.5}}$

[1] Schott no longer uses this formula, but it is still widely used.
[2] Mainly used in the infrared.
[3] Mainly used for fitting of sparse data.

Table 4.6: *Properties of the most important optical glasses*

Glass	MIL	n_d	V_d
BaK1	573576	1.572500	57.549999
BaK2	540597	1.539960	59.709999
BaK4	569561	1.568830	56.130001
BK1		1.510090	63.4600
BK7	517642	1.516800	64.169998
F1	626357	1.625880	35.700001
F2	620364	1.620040	36.369999
F4	617366	1.616590	36.630001
K5	522595	1.522490	59.480000
K7	511604	1.511120	60.410000
LASFN9	850322	1.850250	32.169998
SF2	648339	1.647690	33.849998

Main refractive index and the Abbe number are combined in order to form a six-digit number, the so-called *MIL number*. The first three digits of the MIL number are the d-light refractive index minus one, without the decimal place. The last three digits are the Abbe number V_d times 10. Table 4.6 lists the most important glasses used for lenses and their main data.

Table 4.7: *Optical properties of the most important plastics*

Material	MIL	n_d	V_d
Polystyrol		1.590	30.8
Polycarbonat		1.585	30.0
PMMA (Perspex)		1.491	57.2
CR 39		1.499	57.8

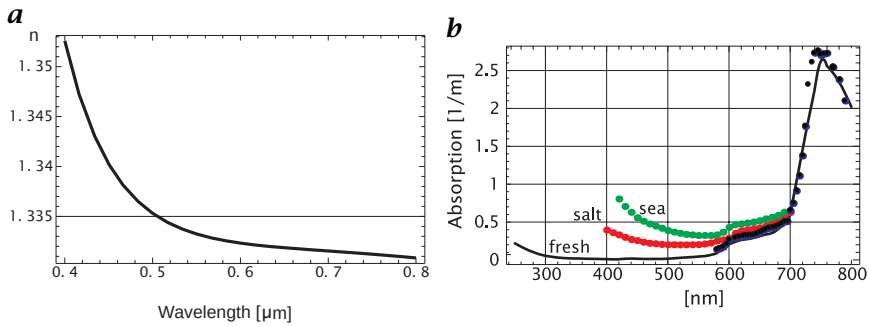


Figure 4.14: ***a*** *Refractive index n of fresh water;* ***b*** *absorption coefficients of fresh water, salt water and sea water.*

In addition to optical glasses, some plastics are used for optical components as well. Mainly Polystyrol and Perspex (Polymethylmethacrylic, PMMA) are used. Because of the limited variety of refractive indices and Abbe numbers, plastics are less flexible than glasses in optical design. However, they are very suitable for the inexpensive production of aspherical and free-from elements by injection molding. Moreover, they may be preferable because of their light weight.

4.4.3 Other materials

Similar to the glasses, fluids are characterized mainly by their refractive index, Abbe number and absorption coefficient. Figure 4.14 shows the *refractive index* $n(\lambda)$ and the *absorption coefficient* of water.

4.5 Aberrations

So far, lenses have been described by the paraxial approximation. Within the limits of this approximation, perfect image quality is achieved. In practice, an optics never reaches this ideal behavior, but shows degradations of image quality caused by *aberrations* of the optics. These are divided into two main classes according to their cause. The change of

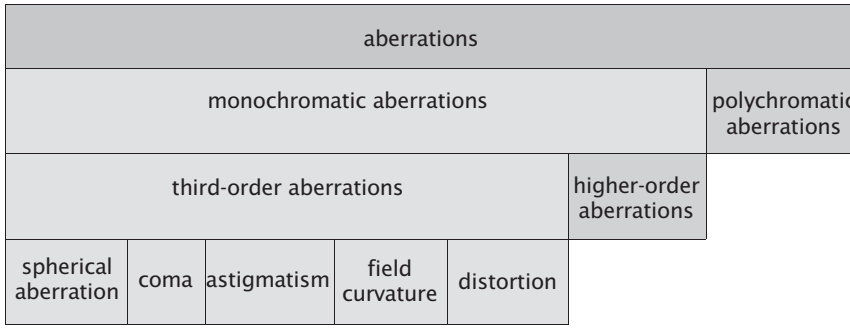


Figure 4.15: *Classification of aberrations.*

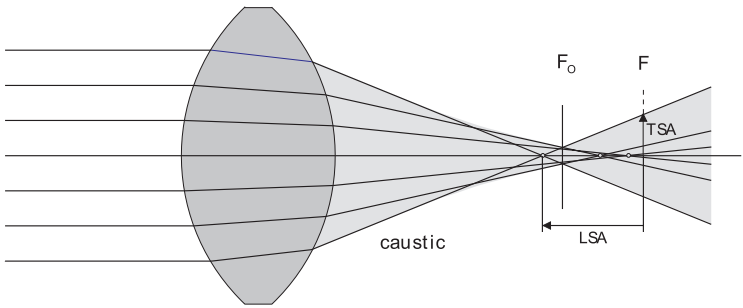


Figure 4.16: *Spherical aberration of a convex lens. To obtain the best image quality, the image plane has to be moved from the paraxial focal plane F to the optimal position F_o. The caustic is the envelope of the outgoing ray bundle.*

refractive index with wavelength causes *polychromatic aberrations* that even exist in paraxial optics. Nonparaxial rays, which appear in any real optics, are the cause of *monochromatic aberrations*. The latter can be described by taking into account the higher-order terms in the series expansion equation (Eq. (4.8)). The third-order aberrations are divided into the five *primary aberrations* (see Fig. 4.15), also known as *Seidel aberrations*. Three of them, namely, spherical aberration, coma and astigmatism, cause image degradations by blurring, while field curvature and distortion deform the image. Understanding aberrations helps to achieve the best possible image quality, and leads to the suppression of aberrations by corrected optics.

4.5.1 Spherical aberrations

Outside the paraxial domain, a spherical surface no longer focuses parallel ray bundles onto a single point. On the contrary, rays hitting the surface at a greater distance to the axis are focused on a point closer to

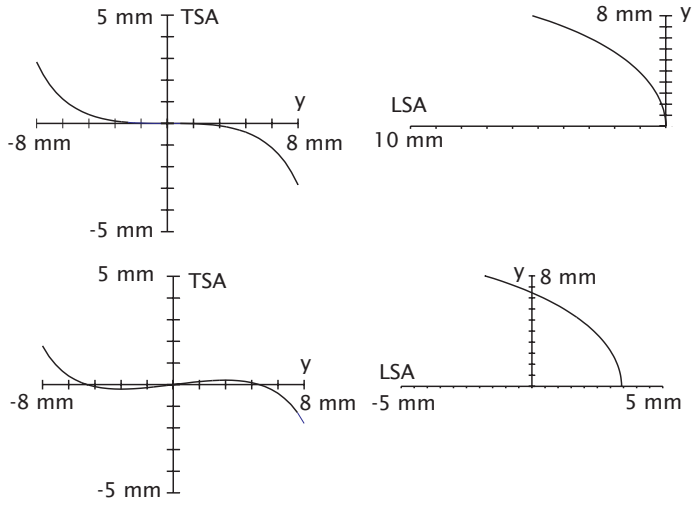


Figure 4.17: *Longitudinal and transversal spherical aberration for the lens from Fig. 4.16. Top row: TSA and LSA at the paraxial focal plane. Bottom row: TSA and LSA at the optimized location. Only TSA can be reduced by relocating the image plane.*

the surface than rays nearer to the axis. The focal length then depends on the radial distance y of the ray to the optical axis.

To describe the strength of a spherical aberration, the axial distance from the true focal point to the paraxial focal point is used; this is called the *longitudinal spherical aberration* (LSA). The sign of the LSA equals the sign of the focal length of the lens. Thus a convex lens with positive focal length bends nonparaxial rays too much, so they intersect the axis in front of the paraxial focus. Diverging lenses with negative focal length focus tend to focus behind the paraxial focus.

To represent the influence of spherical aberrations on image quality, the *transversal spherical aberration* (TSA) can be used. It is defined as the radial distance of the intersection of the outgoing ray with the rear paraxial focal plane, as illustrated in Fig. 4.16. Due to the aberration, exact focusing become impossible.

For practical purposes, it is necessary to minimize the influence of the aberration. This can be done by several methods:

- **Low aperture.** Choosing a larger f-number reduces SA, but causes an unavoidable loss of brightness. Nevertheless, because LSA $\sim y^2$ and TSA $\sim y^3$, this is a very effective way to suppress SA.
- **Image plane shift.** To minimize blur while persevering the aperture setting, it is optimal to move the image plane to the position I_o where the diameter of the caustic is minimal. The minimal but unavoidable

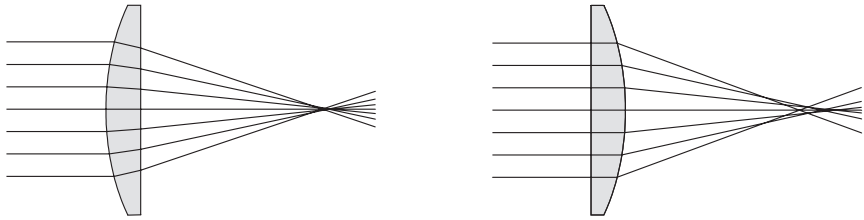


Figure 4.18: *SA of a planoconvex lens (left: correct lens orientation; right: incorrect lens orientation). Turning the lens to the correct orientation strongly reduces SA.*

blur circle is called the circle of least confusion. The suppression of spherical aberration is illustrated in Fig. 4.16. It is important to note that the location of the image plane I_o depends on the imaging conditions, in particular on object distance and f-number.

- **Optimal lens arranging.** Reducing spherical aberration can also be achieved by arranging the surfaces of the system in such a manner that the angles of the rays to the surfaces are as small as possible. This is because SA is caused by the violation of the small angle approximation. The refraction should be evenly distributed among the various surfaces.
  As a general rule, a single lens should always be used with its flat side pointing towards the rays with the higher angles of incidence. When imaging distant objects, a plano-convex lens with an almost flat rear side will produce the best results. For close range imaging a more symmetric lens is more preferable. The reduction of SA by simply turning the lens is illustrated in Fig. 4.18.

4.5.2 Coma

Coma is an aberration associated with off-axis object points. Even a small distance from the axis can cause visible coma in the image. Because of its asymmetric shape, coma is often considered the worst of all aberrations. It is caused by the dependence of the transversal magnification M_T on the ray height. Even in the absence of spherical aberration, this inhibits a focusing of the object point onto a single image point (Fig. 4.19). Coma is considered positive if the magnification increases with increasing ray height h. The image of a point source formed by a lens flawed with coma only shows a comet tail like shape. The pattern can be seen as a series of nonconcentric circles, whereby each circle is formed from the rays passing the lens at the same radial distance h (Fig. 4.20). The centers of the circles are shifted according to the change of M_T with h. Notice that as the rays go around the aperture circle on

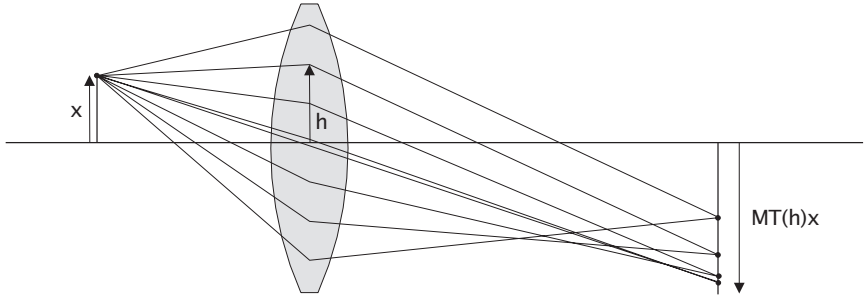


Figure 4.19: *Illustration of negative coma. The transversal magnification decreases with ray height h.*

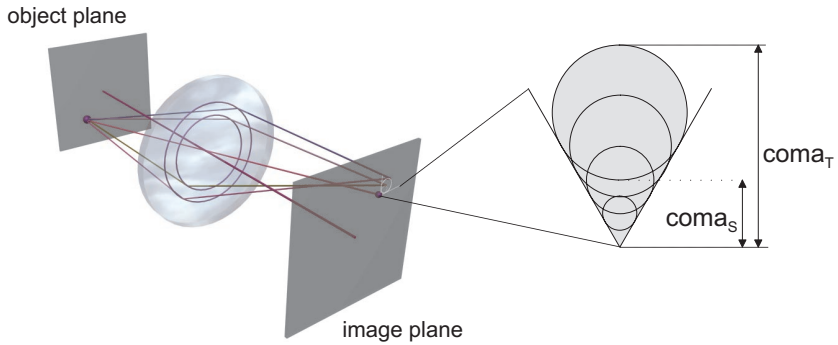


Figure 4.20: *Positive coma of a single point source. The larger the ring on the lens is, the larger is the diameter of the circles in the image. This is reversed with negative coma.*

the lens once, they go around the circle in the coma patch twice. This is why both the tangential as well as the sagittal ray fan form a radial line in the patch. Consequently, the length of both lines is used in order to describe the amount of coma, denoted as sagittal and tangential coma (see Fig. 4.20).

4.5.3 Astigmatism

Astimatism is associated with nonskew ray bundles emerging from nonaxial source points. It is convenient to look at two planar ray bundles in the meridional and in the sagittal plane. The meridional plane is defined as the plane containing the optical axis and the chief ray, while the sagittal plane contains the chief ray and is perpendicular to the meridional plane. Both planes change with the source point of the rays. In addition, the saggital plane changes with each surface, while

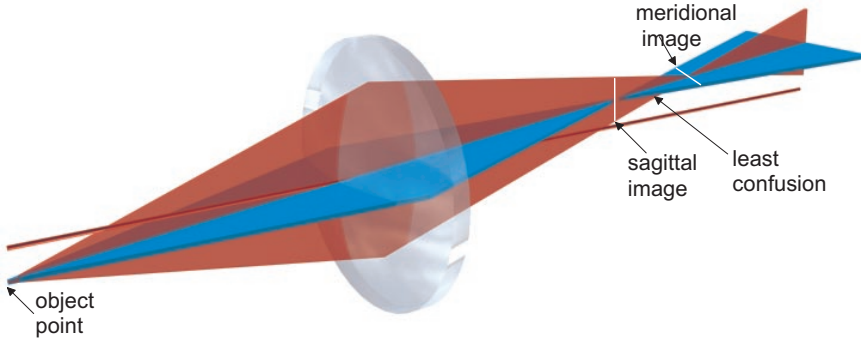


Figure 4.21: *Astigmatism. The focal length differs for the sagittal and the meridional plane.*

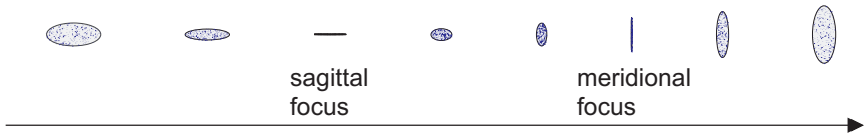


Figure 4.22: *Spot diagrams showing the change of the cross section of the caustic with increasing distance from the lens. The circle of least confusion is located between the two foci.*

the meridional plane remains the same within the optical system. Assuming an optical element of axial symmetry, for an on-axis point there is no difference between the saggital and the meridional plane. An off-axis point will show the lens under different angles, causing the effective focal lengths in the two planes to be different. The difference of the focal length increases with the paraxial focal length of the lens and the skew angle of the rays. The shape of the caustic of the outgoing ray bundle changes from circular shape near the lens to a line in the meridional plane at the meridional image distance. The shape changes further to a perpendicular line at the sagittal image (see Fig. 4.21 and Fig. 4.22). Of course, astigmatism is present for on-axis object points in systems without axial symmetry such as optics containing cylindrical lenses.

4.5.4 Field curvature

With an optical system otherwise free of aberrations, the fact that the cardinal planes are not truly plane causes a primary aberration called the *Petzval field curvature*. Because of the absence of other aberrations the image of a point source is again a point. Within the paraxial domain, all points on the object plane would be imaged exactly to points on

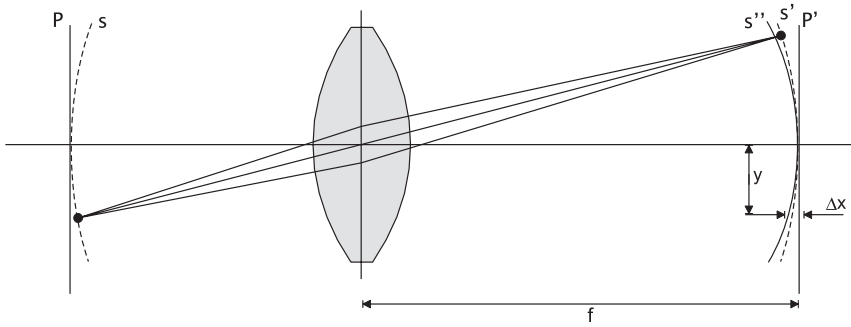


Figure 4.23: *Effect of field curvature. Instead of the plans P and P′ being conjugated, the spheres S and S′ are conjugated. Thus, the parabolic Petzval surface S″ is conjugated to the object plane P.*

the image plane. Because of the cardinal planes being spheres outside the paraxial domain, the conjugate planes turn into conjugate spheres (Fig. 4.23). Consequently, forcing the source points on a plane surface deforms the image surface to a parabolic surface, the *Petzval surface*. A lens with positive focal length bends the Petzval surface towards the lens while a negative lens bends the Petzval surface away from it. Combining lenses with positive and negative focal length can therefore eliminiate field curvature by flattening the Petzval surface to a plane. It can be shown that the horizontal distanze Δz of the Petzval surface from a plane surface is given by

$$\Delta z = \frac{1}{2} y^2 \sum_i \frac{1}{n_i f_i} \tag{4.34}$$

in a system consisting of thin lenses with focal length f_i and refractive indices n_i between the lenses. A system of two thin lenses of focal lengths f_1 and f_2 fulfilling the Petzval condition

$$n_1 f_1 + n_2 f_2 = 0 \tag{4.35}$$

is therefore free of any field curvature. In air ($n_1 = n_2 = 1$), Eq. (4.35) can be written as

$$f_2 = -f_1 \qquad f = \frac{f_1^2}{d} \tag{4.36}$$

Thus a field-corrected lens system in air always has a positive focal length f. Field curvature can also be corrected by moving the stop. Such methods are often combined by using an additional meniscus lens according to Eq. (4.35) and a stop near that lens.

Often lenses are corrected for field curvature by a stop near a meniscus lens.

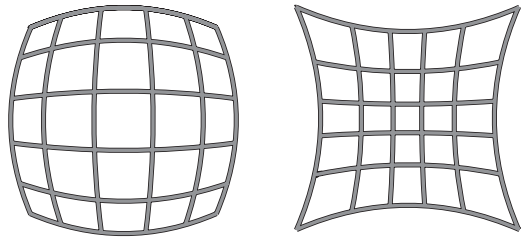


Figure 4.24: *Distortion illustrated by imaging a rectangular grid. Positive distortion causes a pincushion-like shape (right), negative distortion a barrel-shaped image (left).*

In the presence of astigmatism, the Petzval surface splits into a saggital and a meridional surface, corresponding to the location of all saggital and all meridional images. The distance of the meridional image to the Petzval surface is always three times the corresponding distance of the saggital image. Furthermore, both surfaces are located at the same side of the Petzval surface.

4.5.5 Distortions

Displacement of image points with respect to their paraxial locations causes distortions of the image geometry without degrading sharpness. Usually, the displacement increases with the object height as the rays become more inclined. For an optical system of rotational symmetry, the shift of the image points is purely radial and distortion can also be seen as a dependence of the transversal magnification of the distance of the object to the axis. Figure 4.24 illustrates this by imaging a rectangular grid with a complex wide angle lens. As always typical for a wide angle lens, it is flawed with heavy radial distortion. It is important to note that reversing the lens elements causes the distortion change from barrel to pincushion or vice versa. This can be used to eliminate distortion in slides by using the same lens for imaging and for projection. Distortion is influenced by the thickness of the lens and the position of the aperture stop. However, stopping down the aperture does not reduce distortion but it reduces the other aberrations. Therefore, positioning the stop at an appropriate position is often done to correct for distortion.

Table 4.8: *Distortion caused by stop position*

Focal length	Stop in front of lens	Stop behind lens
Positive	Negative distortion (barrel)	Positive distortion (pincushion)
Negative	Positive distortion (pincushion)	Negative distortion (barrel)

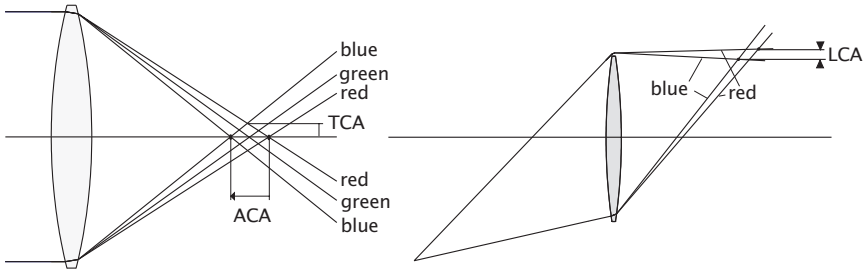


Figure 4.25: *Axial, transverse and longitudinal chromatic aberrations. Different rays correspond to different wavelengths.*

A complex lens system consisting of several lenses or lens groups tends to show distortions because the front lens group acts as an aperture stop in front of the rear lens group. Telephoto lenses typically consist of a positive front group and a negative rear group that can be moved against each other in order to focus or change focal length. Distortion can therefore change with the focal length, even from positive to negative distortion.

4.5.6 Chromatic aberrations

So far, we have only considered monochromatic aberrations caused by the nonlinearity of the law of refraction. The dependence of the refractive index of almost all materials on the wavelength of the light introduces a new type of aberration, because rays of different colors travel on different paths through the optics. Therefore, the images of a point source are different for light of different wavelengths. In particular, the focal length of a lens varies with wavelength.

The effects of chromatic aberration are similar to those of spherical aberration (SA) and in analogy to SA described as axial (ACA) and transverse (TCA) chromatic aberration. As shown in Fig. 4.25, ACA is defined as the axial distance of the focal points corresponding to

two different wavelengths. ACA is called positive if the focal length increases with wavelength, otherwise it is denoted as negative. A positive lens generally shows positive ACA because of the positive Abbe number of all glasses. As then expected, negative lenses cause negative ACA. The radius of the blur circle caused by the different focal lengths is called the transverse chromatic aberration TCA. In addition, CA causes the transversal magnification to become wavelength dependent. This is described by the lateral chromatic aberration (LCA), defined as the axial distance of the different image points.

4.5.7 Reducing aberrations

In the previous sections the primary aberrations have been explained in detail. It is obvious that the image degradation caused by the aberrations has to be suppressed as much as possible in order to achieve a good image quality. This in normally done during the design process of an optics, where ray tracing techniques are used in order to calculate the aberrations and to optimize the system for its desired purpose. Besides these inner parameters of the optics, the strength of aberration is influenced by outer parameters such as f-number or field angle. Image quality can therefore be improved by paying attention to some basic design rules. First of all, aberrations can be influenced by the aperture size h, which is the radial height of the ray hitting the aperture stop, and the radial distance of the object source point from the axis, the field height y. Table 4.9 summarizes the dependence of the Seidel and chromatic aberration from these two parameters. Thus it can be seen that distortion is the only primary aberration that cannot be suppressed by stopping down the aperture. Spherical aberration does not depend on the field height and is therefore the only monochromatic aberration that occurs for on-axis points. In order to estimate the strength of image blur, the radial column of Table 4.9 can be used. For example, if the f-number is increased one step, the aperture size is decreased by a factor of $\sqrt{2}$, meaning that blur circle according to SA is decreased by nearly a factor of three.

4.6 Optical image formation

4.6.1 Geometry of image formation

This section summarizes the most important lens equations used in order to calculate image position and size for imaging optics using the paraxial approximation. The terms used in the following formulas are illustrated in Fig. 4.26. The distance d of the object point P from the front principal plane and its conjugate distance d' of the image point P'

Table 4.9: *Summary of the strength of primary aberrations by field height h and aperture y*

Aberration	Radial (blur)	Axial (focal shift)
Spherical aberration	y^3	y^2
Coma	$y^2 h$	
Astigmatism	$y h^2$	h^2
Field curvature	$y h^2$	h^2
Distortion	h^3	
Axial chromatic aberration		y
Lateral chromatic aberration	h	

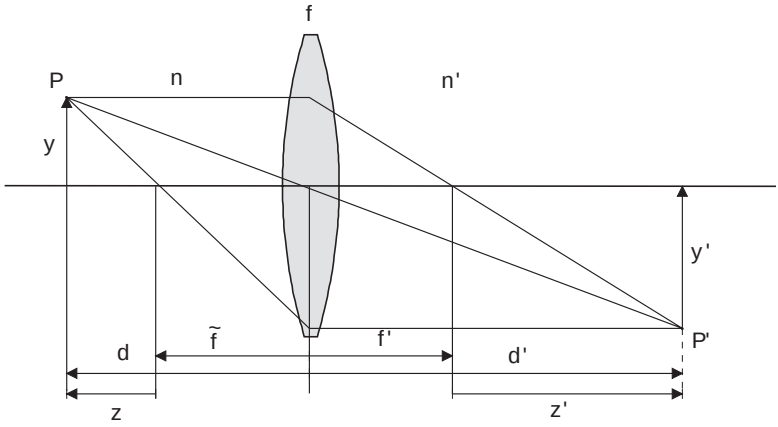


Figure 4.26: *Terms used for the lens equations.*

from the back principal plane both have positive sign in the particular direction away from the lens. The radial distance of image and source point are denoted by y' and y, respectively. As the refractive index of the medium can change from n to n' at the lens, its vacuum focal length f changes to $f' = n'f$ or $\tilde{f} = nf$. Because rays can be thought of as being axis-parallel between the two principal planes, these have been collapsed into a single one for simplicity in the drawing.

The lens equations are commonly expressed either in terms of distances related to the principal planes (d, d') or related to the focal points (z, z'), defined as $z = d - \tilde{f}$ and $z' = d' - f'$. The basic lens equation relates the object and source distances with the focal length:

Distances related to principal planes	Distances related to focal planes
$\frac{f'}{d'} + \frac{\tilde{f}}{d} = 1$ or $\frac{1}{f} = \frac{n}{d} + \frac{n'}{d'}$	$zz' = \tilde{f} f'$

Besides the distances, the image and source heights are related by the *transversal magnification* M_T, defined as the ratio of image to source height; M_T is therefore given by

Distances related to principal planes	Distances related to focal planes
$M_T = \frac{y'}{y} = -\frac{d'n}{dn'}$	$M_T = -\sqrt{\frac{z'n}{zn'}}$

It is sometimes convenient to express image space quantities only in object space terms and vice versa.

Distances related to principal planes		Distances related to focal planes	
Image space	Object space	Image space	Object space
$d' = \frac{n'fd}{d - nf}$	$d = \frac{nfd'}{d' - n'f}$		
$d' = f'(1 - M_T)$	$d = \tilde{f}\left(1 - \frac{1}{M_T}\right)$	$z' = -f'M_T$	$z = -\frac{\tilde{f}}{M_T}$
$M_T = \frac{n'f}{d - nf}$	$M_T = -\frac{d' - n'f}{nf}$		

Imaging an object extended in all three dimensions results in a three-dimensional image filling the image space. In addition to the transversal magnification therefore, the axial extent of the image has to be related to the axial extent of the object. This is done by the *longitudinal magnification*

$$M_L := \frac{\partial d'}{\partial d} = M_T^2 \tag{4.37}$$

which is the square of the transversal magnification.

Figure 4.27 gives an overview of the image distance and the magnification with respect to the object distance. It can be seen that depending on the object distance, the image distance can have positive or negative values. A positive image distance corresponds to a *real image* at which position the rays are focused to from an image.

A *virtual image*, associated with negative image distances, means that the rays in the image space behave as if they would emerge from a point in the object space. There is no point where the rays physically intersect each other, meaning that a virtual image cannot be recorded directly. This is summarized in Table 4.10.

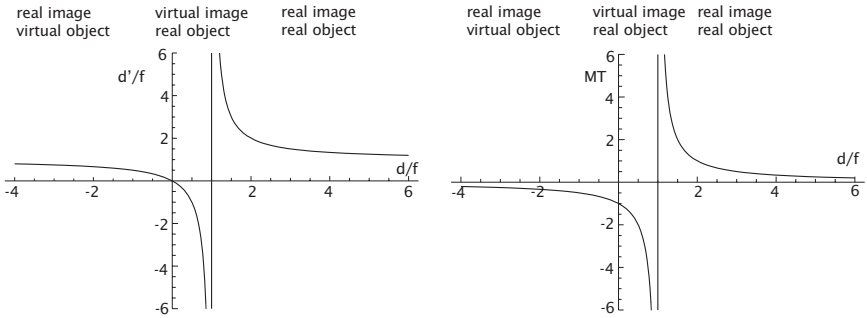


Figure 4.27: *Dependence of the image distance and the transversal magnification with object location. Note that all axes are drawn in units of the focal length of the lens. Their signs will be reversed if a negative lens is considered.*

Table 4.10:

Object location	Image location	Image type	Image orientation	M_T
Convex lens ($f > 0$)				
$\infty > d > 2f$	$f < d' < 2f$	real	inverted	$-1 < M_T < 0$
$d = 2f$	$d' = 2f$	real	inverted	$M_T = -1$
$f < d < 2f$	$\infty > d' > 2f$	real	inverted	$M_T < -1$
$d = f$	$d' = \infty$			
$d < f$	$d' > d$	virtual	erected	$M_T > 1$
Concave lens ($f < 0$)				
$0 < d \leq \infty$	$\lvert d' \rvert < \min(\lvert f \rvert, d)$	virtual	erected	$0 < M_T < 1$

4.6.2 Depth-of-field and focus

A paraxial lens of focal length f focuses all rays emerging from a point P onto its corresponding point P' in image space according to the basic lens equation

$$\frac{1}{f} = \frac{1}{d} + \frac{1}{d'} \tag{4.38}$$

Therefore only objects located at a given distance d are well focused onto the image plane at the fixed position d', whereas objects at other distances $\tilde{d}$ appear blurred (see Fig. 4.28). The distance range in which the blur does not exceed a certain value is called the depth-of-field. A good value to characterize the depth-of-field is f-number $f/2R$, which gives the ratio of the focal length to the diameter of the lens. At a zero

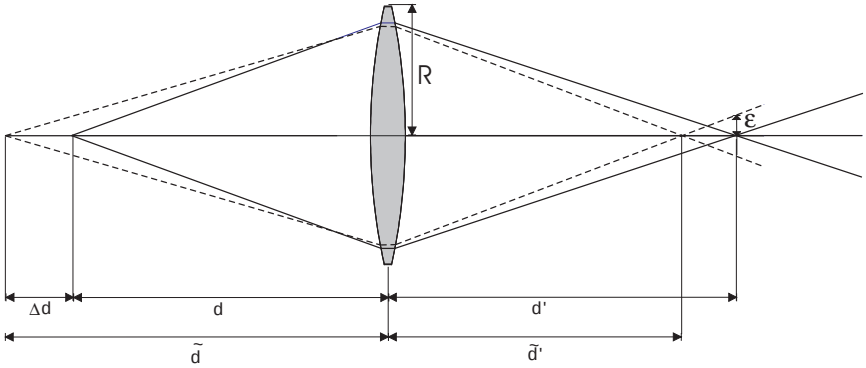


Figure 4.28: *Geometry of image formation for depth-of-field calculations.*

order approximation, blurring is described by the radius ϵ of the blur circle for an object point at $\tilde{d} = d + \Delta d$, which is controlled by the ratio of the image distances

$$\frac{\epsilon}{R} = \frac{d'}{\tilde{d}'} - 1 = d' \frac{\Delta d}{d\tilde{d}} \tag{4.39}$$

The depth-of-field is now determined by the choice of a maximal radius of the blur circle, the so-called circle of confusion. If ϵ_c denotes the circle of confusion, the depth-of-field can be expressed in terms of the magnification $M = b/g$, the f-number $O = f/2R$, and the object distances:

$$\Delta d = \frac{2O}{M_T f} \tilde{d} \epsilon_c = \frac{d}{\frac{M_T f}{2O\epsilon_c} - 1} \tag{4.40}$$

In Eqs. (4.39) and (4.40) we combined the two distinct cases of Δd being positive or negative by understanding ϵ having the same sign as Δd. Distinguishing between positive and negative signs shows the inherent asymmetry for the depth-of-field, caused by the nonlinearity of Eq. (4.38)

$$|\Delta d| = \frac{2O}{M_T f} \tilde{d} |\epsilon_c| = \frac{d}{1 \mp \frac{M_T f}{2O\epsilon_c}} \tag{4.41}$$

Therefore it is a common practice to assume $M_T R \gg \epsilon_c$, leading to the approximation of $\tilde{d} \approx d$ in Eq. (4.40) and removing the asymmetry. For the implications of Eq. (4.40) we consider three special cases, distinguished by the object distance:

Far-field Imaging ($d \gg f$) This case is well known from standard photography using lenses with focal length in the range of more than

some 10 mm. The object is located at a distance large enough to approximate it with infinity, so that $d \gg f$ and therefore $d' \approx f$. The depth-of-field is $\Delta d \approx 2O\epsilon_c/M_T^2$.

Close-up Imaging ($d \approx b$) Close-up or macrophotography indicates the transition from far-field imaging of the microscopic range by using moderate magnifications. Macrophotography is commonly defined as the range of magnifications from 1 to 50, and close-up photography from 0.1 to 1. The depth-of-field is $\Delta d \approx 2O\epsilon_c(1 + M_T)/M_T$.

Microscopic Imaging ($d \approx f$) Optical microscopy works with object distances similar to the focal length of the lens, thus imaging to infinity. The depth-of-field is $\Delta d \approx 2O\epsilon_c 1/M_T$.

Moving the image plane instead of moving the object plane also causes a defocused image. Equivalent to the depth-of-field in object space the term depth of focus in image space denotes the maximal dislocation of the image plane with respect to a given circle of confusion. Again, with the approximation of the circle of confusion being small compared to the lens radius, the depth of focus is given by

$$\Delta d' = \frac{2O}{f} d' \epsilon_c \tag{4.42}$$

The relation between depth of focus and depth-of-field is given by the longitudinal magnification M_T^2.

$$\Delta d = M_T^2 \Delta d' = M_L \Delta d' \tag{4.43}$$

For far-field imaging, M_T is small and therefore a small depth-of-field causes a small depth of focus. In contrast, close-up or microscopic imaging with large magnifications show a large depth of focus and a small depth-of-field at the same time. Finding the position of best focus may be difficult in this particular situation.

4.6.3 Telecentric optics

With this setup, the aperture stop is located at the rear focal point of the respective optics. The effect is that all principal rays in object space are parallel to the optical axis (Fig. 4.29). Only narrow and axis-parallel ray bundles contribute to image formation. This is often used in precision measuring, where an object is viewed by a screen or camera at a fixed position. If the object is moved slightly away from the optimal position, its image becomes blurred, but also the transversal magnification changes so that a different object size is obtained. A telecentric setup corrects this by making the principal ray independent of the object position, therefore preserving the magnification. Obviously only an object smaller than the lens diameter can be viewed. Therefore the use

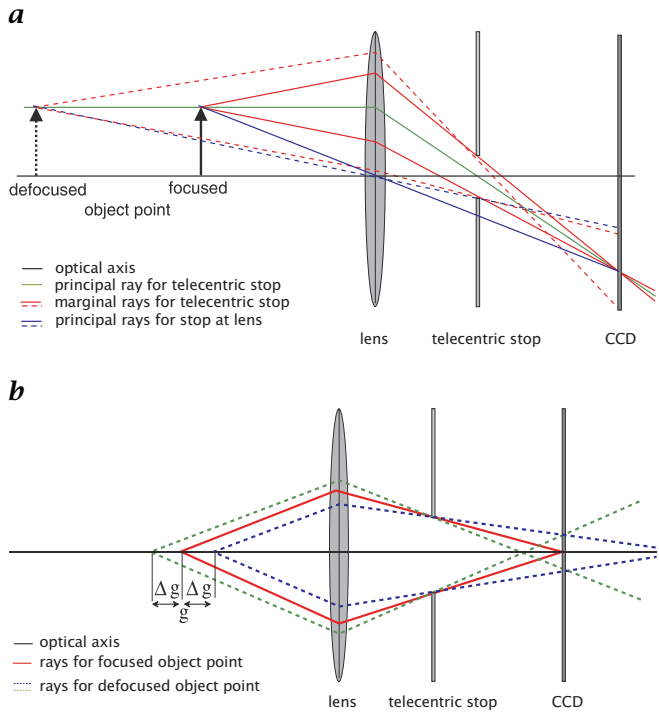


Figure 4.29: ***a*** *As the principal ray is independent of the object position blur does not cause size changes;* ***b*** *independence of the radius of the blur circle from the location.*

of telecentric optics is normally restricted to close-range imaging. To archive the best results, the illumination system should be telecentric as well, and the aperture of illumination and imaging system should be the same.

4.7 Wave and Fourier optics

Pure geometric optics, as we have considered so far, is limited to the calculation of the paths of bundles of light rays through an optical system and the parameters that can be extracted from these. Intensity of these bundles is especially important for imaging optics but is not readily quantified with geometric optics. The depth-of-field calculations explained in Section 4.6 clearly demonstrate this drawback, and while it is possible to obtain the size of the blur circle, the intensity distribution of the image of a blurred spot cannot be calculated exactly. Fourier optics provide a better means of understanding the behavior of

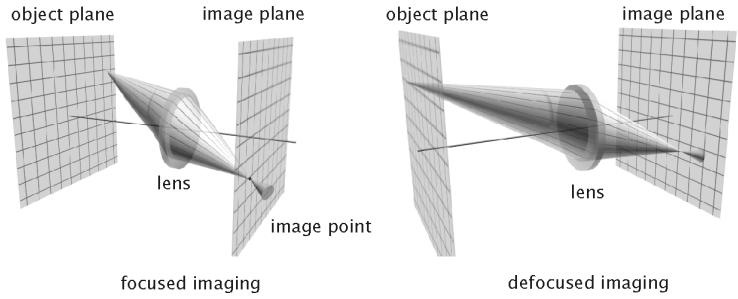


Figure 4.30: *Focused and defocused imaging of an object point onto the image plane.*

an optical system without the need to go deep into the details of wave optics.

4.7.1 Linear optical systems

Point spread function. The point spread function is one of the central concepts used in Fourier optics because it allows the description of a complex optical system as a linear superposition of images of single spot sources. This concept allows the handling of different imaging problems such as quantitative description of image blurring, depth-from-focus reconstruction, and 3-D imaging of non-opaque volume objects as it occurs with light or confocal microscopy, using the same mathematical description. The image of an object is the superposition of the images of all object points. Figure 4.30 illustrates the situation for a well-focused and an ill-focused setup. An ideal aberration-free optics would image every object point onto its conjugate point in the image plane. In the case of defocus the rays emerging from the object point no longer intersect at the image plane but at the plane conjugate to the actual object plane. The image of the object point is therefore an intensity distribution at the image plane, which is is called the *point spread function* (PSF) of the lens.

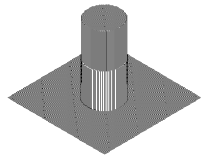
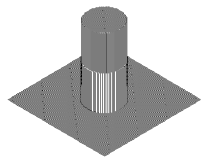
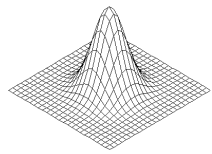
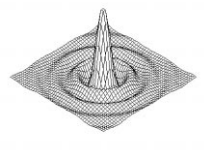
Assuming that the PSF does not change for various object points, the effect of blurring can be described as a convolution of the well-focused image, as it would be achieved by a pinhole camera, with the PSF:

$$g(\boldsymbol{x}') = \int f(\boldsymbol{x}(\vec{\xi}'))PSF(\vec{\xi}' - \boldsymbol{x})d^2\xi' = f(\boldsymbol{x}(\boldsymbol{x}')) * PSF(\boldsymbol{x}') \qquad (4.44)$$

It is important to note that the description by a convolution is only valid in case of a linear, shift-invariant system.

Shape of the PSF. In many cases, we can assume that the shape of the PSF remains unchanged for every object point, independent of its

Table 4.11: *Standard functions for point spread functions of optical systems*

Function	PSF	Used for
Box $\frac{1}{\pi\sigma^2}\Pi\left(\frac{\lvert\boldsymbol{x}\rvert}{2\sigma}\right)$		Optical systems with circular aperture stop that are not dominated by wave optics.
Noncircular Box $\frac{1}{\pi\sigma^2}\Pi\left(\frac{\lvert\boldsymbol{x}\rvert}{2\sigma}\right)$		Optics with the same properties as above, but with a noncircular aperture stop, as with adjustable iris diaphragms. The shape function reflects the shape of the aperture stop.
Gaussian $\frac{1}{2\pi\sigma}\exp\left(-\frac{\boldsymbol{x}^2}{2\sigma^2}\right)$		Widley used in order to describe the PSF. It can be shown that the Gaussian results from the superposition of Airy functions for a wavelength range in the case of polychromatic illumination.
Airy $\frac{2J_1(\lvert\boldsymbol{x}\rvert/\sigma)}{x/\sigma}$		Optical systems that are dominated by wave optics, with coherent and monochromatic illumination, mainly microscopic systems; σ depends on the wavelength.

distance from the plane of best focus. Then, the PSF can be described by a shape function S and a scaling factor σ that varies with the distance g':

$$PSF_Z(\boldsymbol{x}) = \frac{S\left(\frac{\boldsymbol{x}}{\sigma(Z)}\right)}{\int S\left(\frac{\boldsymbol{x}}{\sigma(Z)}\right) d^2x} \tag{4.45}$$

The denominator normalizes the PSF to $\int PSF_Z(\boldsymbol{x})d^2x = 1$, forcing gray-value preservation. In many cases it is sufficient to replace σ by the radius of the blur circle ϵ. The shape function can be completely different for different optical setups. Nevertheless, only a few shape functions are sufficient in order to describe the main properties of standard optics as summarized in Table 4.11.

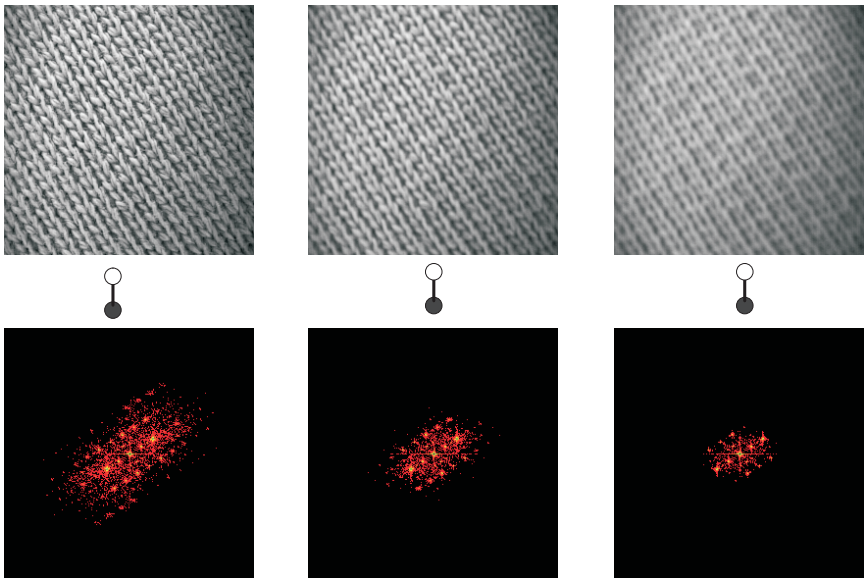


Figure 4.31: *Effect of defocus on images and their Fourier transforms. The cutoff of the higher wavelength is clearly observed with increasing defocus.*

Optical transfer function. In Fourier space, convolution turns into a multiplication of the Fourier transform of the object function with the Fourier transform of the PSF (Volume 2, Section 3.2.3). The latter is called the *optical transfer function* (OTF). Its values give the transfer coefficient for spatial structures of different wavelength through the optical system. A value of zero indicates that this particular wavelength cannot be seen by the optics

$$\begin{array}{lccccc} \text{spatial domain} & G(\boldsymbol{x}) & = & PSF(\boldsymbol{x}) & \otimes & O(\boldsymbol{x}) \\ & \circ\!\!| & & \circ\!\!| & & \circ\!\!| \\ \text{Fourier domain} & \hat{G}(\boldsymbol{k}) & = & \widehat{PSF}(\boldsymbol{k}) & \cdot & \hat{O}(\boldsymbol{k}) \end{array} \tag{4.46}$$

A typical OTF will act as a low-pass filter, eliminating higher spatial frequencies, that is, high resolution details. This is illustrated in Fig. 4.31 showing a series of images of fabric, taken with different focus setting, together with the corresponding Fourier transforms. A telecentric optics has been used in order to avoid scaling of the Fourier space due to change in image magnification. Clearly, the suppression of the higher spatial frequencies with defocus can be seen.

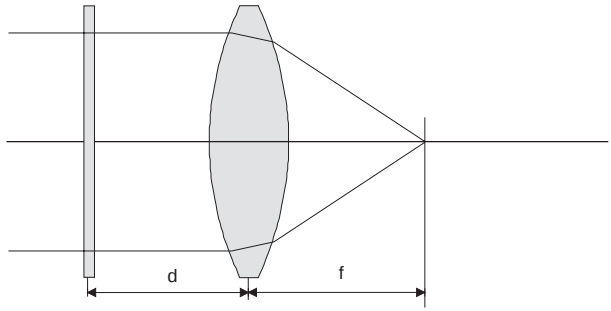


Figure 4.32: *Setup for optical Fourier transformation.*

4.7.2 Optical Fourier transform

One of the most useful properties of a convex lens is its ability to perform a 2-D Fourier transformation. The input image to be transformed has to modulate the amplitude of the incoming light. The simplest possible input would therefore be a monochromatic slide placed in front of the lens (Fig. 4.32). Of course, it is also possible to work with modulation by reflection instead of transmission.

For an infinite lens the intensity distribution in the rear focal plane is given by

$$I(\xi,\eta) = \frac{I_o}{\lambda^2 f^2}\left|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} T(x,y)e^{-i\frac{2\Pi}{\lambda f}(x\xi+y\eta)}dxdy\right|^2 \tag{4.47}$$

which is proportional to the power spectrum of the transmission function $T(x,y)$, that is, the input image. Changing the distance d between the input image and the lens only causes a phase shift and therefore has no influence on the intensity distribution.

To take into account the finite dimensions of the lens, a *pupil function* P is used that is 1 inside the lens and 0 outside the aperture. Thus arbitrarily shaped aperture stops can be described. Within the aperture, Eq. (4.47) changes to

$$I(\xi,\eta) = \frac{I_o}{\lambda^2 f^2}\left|\int\int T(x,y)P(x+\frac{d}{f}\xi, y+\frac{d}{f}\eta)e^{-i\frac{2\Pi}{\lambda f}(x\xi+y\eta)}dxdy\right|^2 \tag{4.48}$$

The amplitude and phase distribution in the rear focal plane correspond to the Fourier spectrum of the input image, and the intensity distribution to the power spectrum.

4.8 References

[1] Schröder, G., (1990). *Technische Optik*, 7th edition. Würzburg: Vogel Buchverlag.

[2] Hecht, E. and Zajac, A., (1977). *Optics*, 2nd edition. Addison Wesley World Student Series. Reading, MA: Addison Wesley.

[3] Schott. *Schott'96—Schott Optical Glass Catalog*. Schott Glass Technologies Inc., 400 York Avenue Duryea, PA 18642 USA, (1996). http://www.schottglasstech.com/SGTDnLoad.html.

[4] Shifrin, K. S., (1988). *Physical Optics of Ocean Water*. AIP Translation Series. New York: American Institute of Physics.

[5] Sullivan, S. A., (1963). Experimental study of the absorption in distilled water, artificial water and heavy water in the visible region of the spectrum. *Jour. Optical Soc. America*, **53**:962–967.

[6] Driscoll, W. E. and Vaughan, W. (eds.), (1978). *Handbook of Optics*. New York: McGraw-Hill Publishing Company.

[7] Tyler, J. E., (1978). Optical Properties of Water. In *Handbook of Optics*, W. E. Driscoll, ed. New York: McGraw-Hill Publishing Company.

[8] Goodman, J. W., (1996). *Introduction to Fourier Optics*, 2nd edition. New York: McGraw-Hill Publishing Company.

[9] Welford, W. T., (1991). *Useful Optics*. Chicago Lectures in Physics. Chicago and London: The University of Chicago Press.

[10] Welford, W. T., (1991). *Abberration of Optical Systems*. The Adam Hilger Series on Optics and Optoelectronics. Bristol: Adam Hilger.

[11] Smith, W. J., (1990). *Modern Optical Design—The Design of Optical Systems*. Optical and Electro-Optical Engineering Series. New York: McGraw Hill.

[12] Yariv, A., (1991). *Optical Electronics*, 4th edition. Fort Worth: Saunders College Publishing.

[13] Spinder & Hoyer, (1998). Gesamtkatalog G4. Göttingen: Spindler & Hoyer. http://spindlerhoyer.de.

[14] Klein, V., Miles and Furtak, E., Thomas, (1996). *Optics*, 2nd edition. New York: John Wiley & Sons (Sd). ISBN 0471872970.

5 Radiometry of Imaging

Horst Haußecker

Interdisziplinäres Zentrum für Wissenschaftliches Rechnen (IWR)
Universität Heidelberg, Germany

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

5.1 Introduction

Radiometry is the measurement of some radiometric quantity, such as radiance L, irradiance E, or intensity I. In terms of computer vision, it relates quantitatively the image brightness to radiometric properties of the observed objects. Thus, a radiometric analysis of images can be used to obtain important information about the underlying physical processes and object properties.

In Chapter 2 we defined the relevant radiometric and photometric quantities and detailed the basics of radiation. Chapter 3 showed how the radiation emitted from objects interacts with all materials that are encountered before it finally reaches the imaging system. In Chapter 4 the fundamentals of optical imaging were introduced.

This chapter concludes the radiometric considerations by combining the fundamental radiometric properties with the process of image formation and shows how quantitative radiometric measurements can be carried out with the imaging detector systems used in computer vision.

Starting at the object surface, we follow the radiation on its way through the camera system and analyze how it is changed by the optical imaging, converted into irradiance at the detector plane, and finally detected, contributing to a digital image.

5.2 Observing surfaces

Most applications of computer vision have to deal with images of opaque objects, which corresponds to images of object surfaces moving within the 3-D scenes. The "brightness" of these surfaces is usually taken for granted with the inherent assumption that they are Lambertian.

This assumption is frequently confused with constant brightness, although even Lambertian surfaces are subject to brightness changes under general conditions in terms of 3-D motion and illumination setups.

But what do surfaces look like, and which radiometric quantity can be remotely measured by an optical detector? In this section, we will address the following fundamental question: Which radiometric property of a surface is measured when it is observed by an optical detector system?

We will conclude that an imaging detector acts as a *radiance meter*, with an output proportional to the *radiance* of the imaged surface.

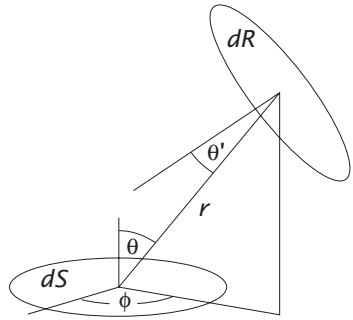


Figure 5.1: *Source-receiver geometry.*

5.2.1 Source-detector flux calculations

In order to measure radiation quantitatively, we need to know which portion of the radiation leaving the surface of an object finally reaches the detector. To derive the basic relations, we consider the geometric setup, illustrated in Fig. 5.1, where the radiative flux of a source is directly transferred (radiated) onto the detector without any imaging device.

Let $\mathrm{d}S$ and $\mathrm{d}R$ be infinitesimal surface elements of the source and the receiver (detector), respectively, separated by a distance r. The *radiance* L leaving the source element $\mathrm{d}S$ in the direction of the receiving surface $\mathrm{d}R$ can be computed from its initial definition Eq. (2.12) as

$$L = \frac{\mathrm{d}^2\Phi}{\mathrm{d}\omega\,\mathrm{d}S\cos\theta} \tag{5.1}$$

where θ is the angle between the surface normal on $\mathrm{d}S$, and the direction of the line connecting $\mathrm{d}S$ and $\mathrm{d}R$. With $\mathrm{d}\omega$ we denote the element of solid angle subtended by the area $\mathrm{d}R$ as observed from the source $\mathrm{d}S$. If $\mathrm{d}R$ is further inclined under an angle θ' with respect to the direction connecting the two surface elements, $\mathrm{d}\omega$ is given by

$$\mathrm{d}\omega = \frac{\mathrm{d}R\cos\theta'}{r^2} \tag{5.2}$$

Combining Eqs. (5.1) and (5.2), we get the infinitesimal element of radiative flux transferred between $\mathrm{d}S$ and $\mathrm{d}R$:

$$\mathrm{d}^2\Phi = L\frac{\mathrm{d}S\,\mathrm{d}R\cos\theta\cos\theta'}{r^2} \tag{5.3}$$

From this equation we can immediately infer the following basic properties of radiative transfer: The transfer of radiative flux is:

1. directly proportional to the *radiance* L of the emitting surface $\mathrm{d}S$;

2. directly proportional to the areas of the emitting and receiving surfaces $\mathrm{d}S$, and $\mathrm{d}R$, respectively;
3. inversely proportional to the square of the distance r between emitting and receiving surface (inverse square law); and
4. finally, it depends upon the orientation of the surface normals of $\mathrm{d}S$ and $\mathrm{d}R$ with respect to the direction connecting the two surfaces.

The most important fact is that the received flux is directly proportional to the radiance of the emitting surface. We will further show that this proportionality remains for all further considerations leading towards the final imaging detector. Thus, the basic property to be measured by radiometry is the *radiance* of the objects!

For finite size sources and detectors, we need to integrate Eq. (5.3) over the surface areas S and R of source and detector, respectively,

$$\Phi = \int_S \int_R L \frac{\cos\theta \cos\theta'}{r^2} \,\mathrm{d}S \,\mathrm{d}R \tag{5.4}$$

The average *irradiance* E of the receiving detector element is given by:

$$E = \frac{\mathrm{d}\Phi}{\mathrm{d}R} = \int_S L \frac{\cos\theta \cos\theta'}{r^2} \,\mathrm{d}S \tag{5.5}$$

The integrals Eq. (5.4) and Eq. (5.5) are the fundamental equations describing the transfer of radiation from a source surface to a detector surface [1]. These integrals, however, can only be solved analytically for simple geometrical setups.

For practical applications it is common to separate the geometrical aspects of the radiative transfer from the magnitude and spectral distribution of the radiance by defining a *configuration factor*. It is defined as ratio of the flux Φ_r on the receiver by the total emitted flux of the source, Φ_s [1]:

$$F_{s-r} = \frac{\int_S \int_R L \cos\theta \cos\theta' r^{-2} \,\mathrm{d}S \,\mathrm{d}R}{\int_S \int_{2\pi} L \cos\theta \,\mathrm{d}S \,\mathrm{d}\Omega} \tag{5.6}$$

where the integration of the denominator of Eq. (5.6) is carried out over the entire hemispheric enclosure. The indices F_{s-r} indicate the flux transfer from source to receiver. In case of a *Lambertian* source, the radiance can be drawn out of the integrals and Eq. (5.6) reduces to

$$F_{s-r} = \frac{1}{\pi S} \int_S \int_R \cos\theta \cos\theta' r^{-2} \,\mathrm{d}S \,\mathrm{d}R \tag{5.7}$$

which contains only geometrical quantities. For homogeneous Lambertian sources with radiance L_s, the exchanged flux is given by

$$\Phi_{s-r} = \pi L_s S F_{s-r} \tag{5.8}$$

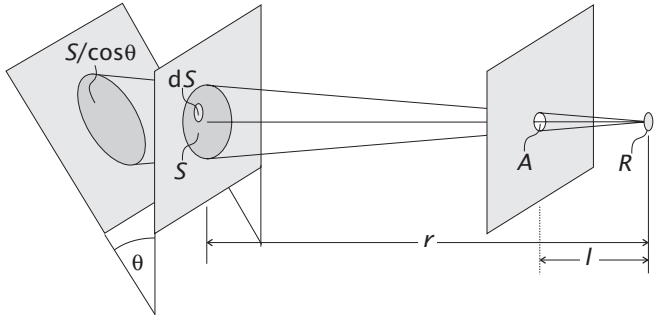


Figure 5.2: *Schematic setup of radiance meter (Gershun tube).*

Thus, if the geometry of the setup remains unchanged, the configuration factor stays constant and we can focus on the variable portions of the detector flux or irradiance.

More details on configuration factors are given by [2] and [1]. Tabulated values and approximated equations for a variety of geometrical setups can be found in [3].

If we just place a detector into a scene, all surface areas within the 3-D enclosure contribute to detector irradiance. Thus, we have to integrate Eq. (5.5) over the entire surface of all surrounding—arbitrarily shaped—objects. Apart from the mathematical difficulties, this integration yields the average irradiance of the detector surface element, rather than an "image" of the individual object surfaces. In order to resolve spatial variations of emitting surfaces, we need to restrict the allowed angles of incidence.

Section 5.2.2 outlines the principal setup of a point measuring radiometer and the basic radiometric properties, which constitute the basis for imaging systems.

5.2.2 Radiance meter

A simple *radiance meter* can be set up by a radiation detector R placed at the bottom of a tube of length l, with a welldefined aperture of diameter A on top. Such a device is commonly referred to as *Gershun tube* in the literature. Figure 5.2 illustrates the principal geometric relationships.

The entrance aperture of the radiometer limits incident radiation to a conical solid angle

$$\Omega_s = \frac{S}{r^2} = \frac{A}{l^2} \tag{5.9}$$

as observed from the center of the detector R. If the source S is inclined under an angle θ with respect to the axis of symmetry defined by the center of the detector R and the center of the aperture stop A (Fig. 5.2), the size of the observed source area S' is increased by:

$$S' = \frac{S}{\cos\theta} \tag{5.10}$$

Using the relation Eq. (5.4), we can derive the flux, which is emitted from the surface S' and received by the detector R. As the detector is arranged perpendicular to the axis of symmetry, $\theta' = 0$ (Fig. 5.1). For small detector elements $R \ll l^2$, we can assume the flux to be constant over R and replace the integration over the detector element $\mathrm{d}R$ by the area R. If we further assume the distance r to be much larger than the length l of the radiometer $l \ll r$, the distance r stays constant for all points on S' and can be removed from the integral, as well. Hence,

$$\Phi = \frac{R}{r^2} \int_{S'} L(\theta) \cos\theta \, \mathrm{d}S' \tag{5.11}$$

which simplifies to

$$\Phi = \frac{R}{r^2} \int_{S} L(\theta) \, \mathrm{d}S \tag{5.12}$$

using Eq. (5.10).

If the radiance $L(\theta)$ is constant over S, we can draw it out of the integral and Eq. (5.12) reduces to

$$\Phi = L(\theta) \frac{R}{r^2} \int_{S} \mathrm{d}S = L(\theta) \frac{RS}{r^2} = L(\theta) \frac{RA}{l^2} = L(\theta) c_g \tag{5.13}$$

where we have used Eq. (5.9) to replace the source-related properties S and r by the detector properties A and l. This yields a proportionality constant c_g given by the geometric proportions of the radiometer.

Thus, the flux received by the detector is proportional to the radiance of the source under the given direction θ, that is, the Gershun tube behaves like a *radiance meter*.

For Lambertian surfaces, that is, $L(\theta) = L$, Eq. (5.13) becomes independent from the inclination angle of the surface,

$$\Phi_L = L \, c_g \tag{5.14}$$

which means that a Lambertian surface shows equal brightness independent of the viewing direction!

However, if L is not constant over S the flux Φ is averaged over the entire area S of the source (Eq. (5.13)). As the total flux is proportional to the aperture size A, we need to increase the aperture in order to

collect a sufficient amount of radiation from faint sources. This does in turn increase the area S, therefore reducing the resolution of the radiometer. The only way to avoid this problem is to use an imaging optics, which allows the collected radiation to be increased without reducing the resolution of the system. This will be the topic of Section 5.4.

5.2.3 Revisiting Lambert: case studies

An important result derived in Section 5.2.2 is the fact that *Lambertian surfaces* appear to have the same brightness under all observation angles. This seems to be inconsistent with *Lambert's cosine law*, that is, a cosine dependence of emitted intensity (Eq. (2.27)).

To resolve this apparent contradiction, we need to distinguish carefully between solid angles and area sizes that are related to the detector and those that are related to the source. It is important to note that a detector observes the source under a fixed viewing solid angle Ω_s (Eq. (5.9)), which is given by the detector geometry and does not change with orientation of the source.

For finite source elements with an area smaller than that observed by the detector, the effective size of the source *decreases* with the projection of the surface on the direction perpendicular to the line between detector and source. Thus, the measured flux indeed shows cosine dependence.

In contrast to small source elements, extended sources show a different behavior. If the *total* area of extended sources subtends solid angles larger than Ω_s, the *effective* area observed by the detector *increases* with $\cos^{-1}\theta$, (Eq. (5.10), Fig. 5.2), which compensates the decreasing intensity in this direction. Thus, an infinitesimal source element shows the $\cos\theta$-dependence, but the number of such elements observed by the imaging system on an extended surface increases with $\cos^{-1}\theta$.

This fact is another manifestation of the definition of radiance, as opposed to radiant intensity, and shows that radiance is analogous to the visual sensation of perceived (imaged) brightness of the surface. As radiance is constant under all angles for Lambertian surfaces, the brightness of Lambertian surfaces remains constant under varying angles.

Case studies of surface properties. In the following examples we will illustrate the difference between Lambertian and non-Lambertian surfaces using a number of geometrical setups (camera and illumination).

Example 5.1: Surface under different observation angles

Consider a surface S to be illuminated by parallel light, using a directional illumination source under a fixed angle θ_i. The illumination

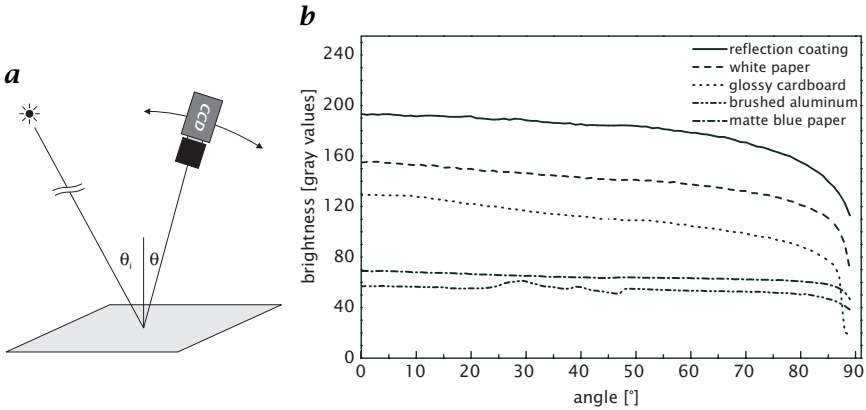


Figure 5.3: *Lambertian surface observed under different angles with fixed illumination:* ***a*** *experimental setup;* ***b*** *apparent surface brightness vs observation angle for different surfaces.*

source is considered to be far away from the surface, producing a constant irradiance $E(\boldsymbol{x}) = E$ over S. A camera is observing the surface under changing observation angles, $0 \leq \theta < 90°$ (Fig. 5.3a).

If the surface is Lambertian and the illumination source is fixed, the measured image brightness should remain constant with respect to θ (Eq. (5.14)).

For non-Lambertian surfaces, the brightness should exhibit a faster decrease with respect to the angle θ. Figure 5.3b shows the angular dependence of the apparent brightness for several different surfaces. The surface brightness is averaged over a fixed area of interest within the image for angular steps of 1°. The solid line corresponds to a commercial matte reflection coating with a high reflectivity of $\tilde{\rho} = 0.99$. It has the highest apparent brightness in the images. However, it does not show an angle-independent brightness. White paper shows a similar angular dependence but has a lower reflectivity. The fastest decrease in brightness with respect to angle can be observed for the glossy cardboard, which exhibits a mixture of specular and matte reflection.

A remarkable example for Lambertian surface characteristics can be observed with matte blue paper. Although having the lowest brightness, the angular dependence remains almost constant for angles up to 85°.

Brushed aluminum shows almost the same quasi-Lambertian behavior except that the specular component of the surface reflectivity leads to intermittent brightness changes with respect to the observation angle. It is important to note that all surfaces show a fast brightness decrease towards zero for angles close to 90°. The measurement, however, could not be carried out for angles above 88°.

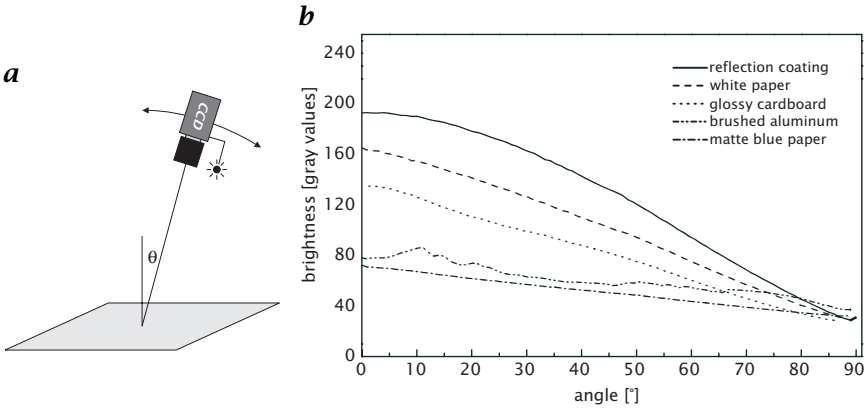


Figure 5.4: *Surface under different combined illumination/observation angles:* ***a*** *experimental setup;* ***b*** *apparent surface brightness vs observation angle for different surfaces.*

Example 5.2: Illumination under different angles

Instead of a fixed illumination, we attach the light source to the camera, which is again observing a surface under changing observation angles (Fig. 5.4).

Now, the irradiance of the surface S is changing with the angle of incidence, according to

$$E(\theta) = E(0)\cos\theta \tag{5.15}$$

as the same amount of radiative flux is spread over a larger area with increasing angle. Hence, even a Lambertian surface shows a cosine dependence of the image brightness with increasing angle. This is the case because the reflected radiance remains constant for changing viewing angles of Lambertian surfaces. The reflection, however, is proportional to the irradiance of the surface.

A non-Lambertian surface shows a much faster decrease with angle, as the decreasing irradiance and the angular decrease of the reflectivity add up. Figure 5.4b shows the angular dependence of the apparent brightness for the same surfaces as already shown in Example 5.1. Although they differ by the absolute brightness due to the different surface reflectivities, they all are dominated by the cosine relationship.

Non-Lambertian surfaces, however, show a faster decrease at large angles than that of Lambertian surfaces (compare Fig. 5.3b and Fig. 5.4b). Again, the brushed aluminum exhibits a strong variation of the surface brightness due to specular reflection.

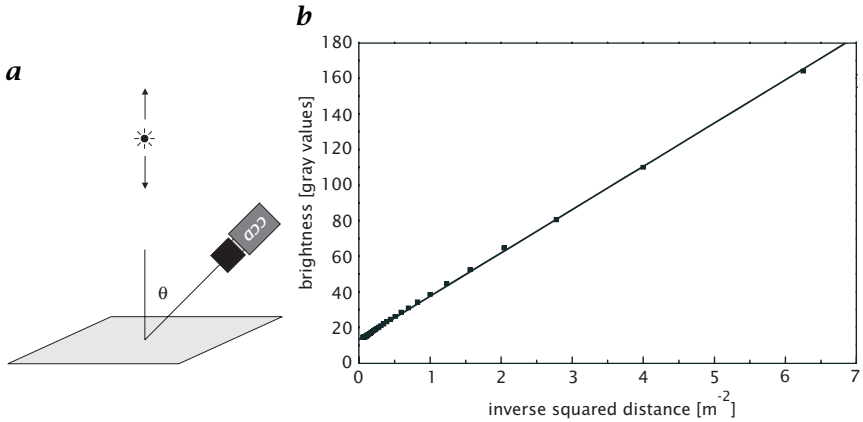


Figure 5.5: *Inverse-square law of irradiation on a surface:* ***a*** *experimental setup;* ***b*** *measured brightness vs distance, together with fitted inverse-square relationship.*

Example 5.3: Inverse-square law

In order to verify the inverse-square law, which has been derived in Chapter 2 (Eq. (2.26)), we need to move a point light source along a straight line perpendicular to the surface under observation. The experimental setup consists of a long (4 m) optical bench. The camera is fixed with respect to the surface, under an observation angle of $\theta = 45°$, such that it does not block light from the light source (Fig. 5.5a).

Figure 5.5b shows the resulting image brightness with respect to the distance d between light source and surface. The image of the surface was averaged over a fixed area of interest and plotted vs the inverse squared distance, d^{-2}. In this way, the inverse-square law reduces to a linear relationship. A linear fit of the measured data shows that the inverse-square law can be experimentally verified.

5.3 Propagating radiance

In Section 5.2 we learned that a radiometer serves as a *radiance meter*, which produces an output proportional to the radiance of the observed surfaces. Before we turn towards the question of how the radiance *distribution* of an object surface is converted into *irradiance* of the sensor plane by the optical image formation process, we need to consider exactly what happens to radiance when propagating through space and passing the camera lens system. We will derive a fundamental law of radiometry—referred to as *radiance invariance*—which constitutes the basis for all radiometric measurements. The derivation of this law follows McCluney [1] and Nicodemus [4].

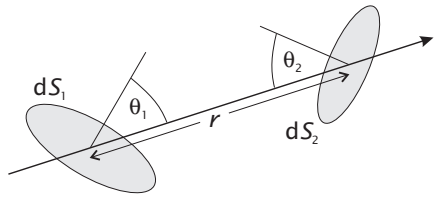


Figure 5.6: *Illustration of the radiance invariance.*

5.3.1 Radiance invariance

The concept of *radiance* is sometimes hard to grasp, as we intuitively think about radiation to be either absolutely parallel—in that case, we do not have a geometrical spreading and, hence, no radiance—or diverging in space. As radiance is defined as flux emitted into a unit solid angle, we always tend to think that it is diverging and, hence, becoming smaller, the farther it travels.

An important question in the context of imaging systems is whether the measured brightness is decreasing with increasing object distance or, in general, how the radiance is distributed over the lens system at all.

In order to derive the law of radiance invariance, we consider two "virtual" infinitesimal surface elements $\mathrm{d}S_1$ and $\mathrm{d}S_2$ placed along the propagation direction of the measured radiation (Fig. 5.6) at distance r. The surface normals of the two elements with respect to the direction of the connecting line are inclined under the angles θ_1 and θ_2, respectively. The incident flux on either of the two elements is considered to leave the element in exactly the same direction at the opposite side, without attenuation.

The flux *leaving* surface element $\mathrm{d}S_1$ is given by Eq. (5.3)

$$\mathrm{d}^2\Phi_1 = L_1 \frac{\mathrm{d}S_1 \cos\theta_1 \, \mathrm{d}S_2 \cos\theta_2}{r^2} \tag{5.16}$$

where L_1 denotes the incident radiance on the surface element $\mathrm{d}S_1$. Similarly, the *incident* flux on surface element $\mathrm{d}S_2$ is given by

$$\mathrm{d}^2\Phi_2 = L_2 \frac{\mathrm{d}S_2 \cos\theta_2 \, \mathrm{d}S_1 \cos\theta_1}{r^2} \tag{5.17}$$

Conservation of energy requires that both fluxes must be the same if no losses occur within the medium between $\mathrm{d}S_1$ and $\mathrm{d}S_2$, that is, $\Phi_1 = \Phi_2$. Using Eq. (5.16) and Eq. (5.17) we get

$$L_1 = L_2 \tag{5.18}$$

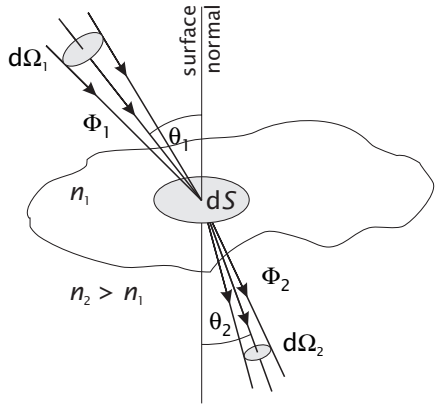


Figure 5.7: *Geometry for definition of radiance invariance at interfaces.*

As we have made no restrictions on the locations, orientations, or sizes of the surface elements, nor on the origin of the radiance, Eq. (5.18) constitutes a fundamental law, called *radiance invariance*.

Although this solution seems to be trivial, it is of major importance, as it proves, that the quantity of radiance is not changed along the ray of propagation in space. Thus, it makes absolutely no difference where we measure the emitted radiance of objects.

5.3.2 Radiance invariance at interfaces

In this section, we consider the question as to how radiance is changed at the interface between objects with different refractive indices. This extension of the radiance invariance constitutes the basis for radiometric measurements with optical systems.

At the interface between two media with different indices of refraction, not only the direction of propagation changes but also the radiance because the geometric spreading of the beam is altered. Figure 5.7 illustrates the geometric quantities at the transition from n_1 to n_2, for $n_2 > n_1$. As refraction is not linear in angle, the two bounding rays are refracted under different angles due to the slightly different angles of incidence.

The element of incident flux $\mathrm{d}\Phi_1$ is given by

$$\mathrm{d}\Phi_1 = L_1 \,\mathrm{d}S \cos\theta_1 \,\mathrm{d}\Omega_1 = L_1 \,\mathrm{d}S \cos\theta_1 \sin\theta_1 \,\mathrm{d}\theta_1 \,\mathrm{d}\phi \tag{5.19}$$

where $\mathrm{d}S$ denotes an infinitesimal surface area, and the element of solid angle $\mathrm{d}\Omega_1$ is replaced by spherical coordinates. Correspondingly, the element of refracted flux $\mathrm{d}\Phi_2$ is given by

$$\mathrm{d}\Phi_2 = L_2 \,\mathrm{d}S \cos\theta_2 \,\mathrm{d}\Omega_2 = L_2 \,\mathrm{d}S \cos\theta_2 \sin\theta_2 \,\mathrm{d}\theta_2 \,\mathrm{d}\phi \tag{5.20}$$

Conservation of energy requires

$$\mathrm{d}\Phi_2 = (1 - \tilde{\rho})\ \mathrm{d}\Phi_1 \tag{5.21}$$

accounting for reflection at the interface. Thus

$$1 = \frac{(1 - \tilde{\rho})\ \mathrm{d}\Phi_1}{\mathrm{d}\Phi_2} = \frac{(1 - \tilde{\rho})\, L_1 \cos\theta_1 \sin\theta_1\, \mathrm{d}\theta_1}{L_2 \cos\theta_2 \sin\theta_2\, \mathrm{d}\theta_2} \tag{5.22}$$

The relation between the angles of incidence and refraction is given by Snell's law (Eq. (3.15), see Chapter 3)

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \tag{5.23}$$

Differentiating both sides of this expression with respect to the angle yields

$$\frac{n_1}{n_2} = \frac{\cos\theta_1\, \mathrm{d}\theta_1}{\cos\theta_2\, \mathrm{d}\theta_2} = \frac{\sin\theta_1}{\sin\theta_2} \tag{5.24}$$

Combining Eq. (5.24) with Eq. (5.22) yields

$$\frac{(1 - \tilde{\rho})\, L_1}{n_1^2} = \frac{L_2}{n_2^2} \tag{5.25}$$

Ignoring reflection losses, the radiance is changed at the transition between two interfaces, but the quantity L/n^2 stays constant in any medium[1].

This leads to the conclusion that the radiance is not altered by optical components such as lenses and windows. Although the radiance within a lens is changed, the initial radiance is restored after exiting the lens at the second face. However, if the lens system is not loss-less due to reflections at all faces and internal absorption, only the fraction $\tilde{\tau}$ of the incident radiance is transmitted:

$$L_2 = \tilde{\tau} L_1 \tag{5.26}$$

5.4 Radiance of imaging

Now that we know that the radiance is conserved by passing through the optical system of a camera (with the exception of absorption and reflection losses), we need to know how the optical system changes the direction of propagation and the geometric spreading and how it turns the radiance distribution into an image. This section is meant to complement the basic considerations regarding the point measurement radiometer (Gershun tube) as described in Section 5.2.2 for an imaging radiometer.

[1] This fundamental law of radiometry can be compared to the invariance of the optical path nd in geometrical optics (see Chapter 4).

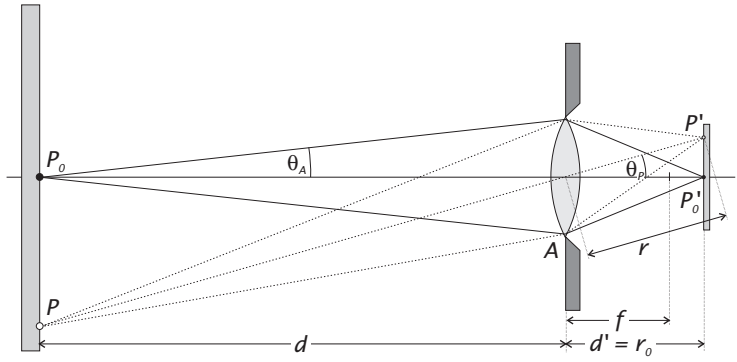


Figure 5.8: *Illustration of image formation by a perfect lens.*

5.4.1 Radiance and irradiance of images

Consider the imaging system to consist of a single circular lens, as illustrated in Fig. 5.8. We assume the lens to be perfect in terms of accurately focusing all radiation emerging from a point P at the object surface and collected by the lens aperture A, onto a single point P' on the sensor plane.

Let P'_o be the center point on the optical axis of the lens, that is, in the center of the image, and P_o the corresponding point at the object surface. The solid angles subtended by the lens aperture A, as observed from the point P_o, and from its image P'_o, are denoted by Ω and Ω', respectively.

The irradiance E' of the image point P'_o is simply given by integrating the radiance impinging onto this point from all angles within the solid angle Ω':

$$E'(P'_o) = \int_{\Omega'} L'(\theta', \phi') \cos\theta' \, \mathrm{d}\Omega' \tag{5.27}$$

where the primed letters refer to the quantities at the sensor side of the lens, that is, after passing the lens (Fig. 5.8).

Using the radiance invariance Eq. (5.26), we can replace L' by $L' = \tilde{\tau} L$, if we assume the lens to have a transmittance $\tilde{\tau}$, and L denotes the object radiance before reaching the lens. As the lens focuses all radiation, which is emitted by the point P_o into the solid angle Ω, we can replace the integration over the primed quantities in the image domain by an integration over the solid angle Ω in the object domain:

$$E'(P'_o) = \tilde{\tau} \int_{\Omega} L(\theta, \phi) \cos\theta \, \mathrm{d}\Omega \tag{5.28}$$

where $L(\theta, \phi)$ denotes the excitant radiance at the object point P_o.

For Lambertian surfaces, L is independent of the direction and can be removed from the integral. Thus,

$$E'(P'_o) = \tilde{\tau} L \int_\Omega \cos\theta \, \mathrm{d}\Omega = \pi \tilde{\tau} L \sin^2 \theta_A \tag{5.29}$$

with θ_A denoting the half angle of the lens aperture, as viewed from point P_o (Fig. 5.8). The larger the lens aperture, the more radiance is collected by the lens and the more irradiance is produced at the sensor. Hence, an optical imaging system allows the amount of collected radiative flux to be increased without reducing the spatial resolution, as opposed to the Gershun tube (Section 5.2.2). The maximum possible irradiance is collected for $\sin\theta_A = 1$, that is, for an infinite sized lens:

$$\max_{\theta_A} E'(P'_o) = \pi \tilde{\tau} L \tag{5.30}$$

which equals the *radiant exitance* of the surface at the point P_o (see Chapter 2, Eq. (2.14)), reduced by the transmittance of the lens.

Using the *f-number* n_f of the lens (Chapter 4), Eq. (5.29) can be rewritten as

$$E'(P'_o) = \pi \tilde{\tau} L \left(\frac{1}{1 + n_f^2} \right) \tag{5.31}$$

5.4.2 Field darkening

So far, we have considered only the central point P_o in the image, located on the optical axis of the lens. This section shows how the sensitivity of an extended detector decreases towards the edges of the sensor.

Off-axis irradiance. Let P' be an arbitrary image point located off-axis in the sensor plane. The corresponding point in object domain is denoted by P. Further, let P have the same radiance as the center point P_o, that is, we assume the object to have a constant radiance over the imaged area.

Now, the distance r from the center of the lens to the point P' will depend on the angle θ_P,

$$r = \frac{r_o}{\cos\theta_P} \tag{5.32}$$

where θ_P denotes the angle between the line connecting P and P' (passing through the center of the lens) and the optical axis, and r_o is the distance between the center of the lens and P_o (Fig. 5.8).

According to the inverse square law Eq. (5.2), the irradiance is proportional to $1/r^2$, which reduces the off-axis irradiance $E'(P')$ by the factor $\cos^2\theta_P$, compared to $E'(P'_o)$.

Another factor further reducing the irradiance $E'(P')$ is given by the fact that the solid angle Ω, subtended by the lens, decreases proportional to $\cos\theta_P$ (Eq. (2.5), see Chapter 2). Thus, the effective lens aperture is reduced by the projection onto the viewing direction.

Finally, the irradiance $E'(P')$ at the detector plane is proportional to the angle of incidence, which is also given by $\cos\theta_P$.

Combining all influences decreasing the irradiance E', we get the following result for off-axis points:

$$E'(P') = E'(P'_o)\cos^4\theta_P \tag{5.33}$$

This $\cos^4$-dependence is known as *field darkening*, reducing the irradiance towards the edge of the sensor plane.

Typical values of the relative decrease of irradiance at the edge of the image compared to the center point are in the order of 10 % and 0.5 % for $f = 25\,\text{mm}$ and 100 mm, respectively. With increasing focal length, the field darkening is expressed less. For wide-angle lenses, however, this effect can not be neglected. Volume 3, Fig. 32.3b shows an example of an image taken with a wide-angle endoscope optic. The field darkening is clearly visible.

Vignetting. In addition to the $\cos^4$-dependence of the irradiance across the sensor plane, other optical effects contribute to the resulting field darkening of an image. The term *vignetting* is used for effects blocking off-axis rays by internal aperture stops of the lens system or other beam-delimiting components [1]. Such effects produce an additional decline of the image irradiance towards the edge of the image.

5.5 Detecting radiance

The final step in the chain of radiometric imaging is the detection of radiation at the imaging sensor. Here, the irradiance of the sensor plane is converted into an electronic signal. Without going into details of solid state physics, this section outlines the basic properties of imaging detectors relevant for a quantitative radiometric interpretation of images. More detailed overviews of detectors for electromagnetic radiation can be found in the following excellent textbooks [1, 5, 6], as well as in standard handbooks on radiometry, such as [3].

5.5.1 Detector performance: figures of merit

Before we turn towards a classification of optical detectors in terms of their operational principle, we will summarize commonly used figures of merit, which allow us to compare the relative performance between

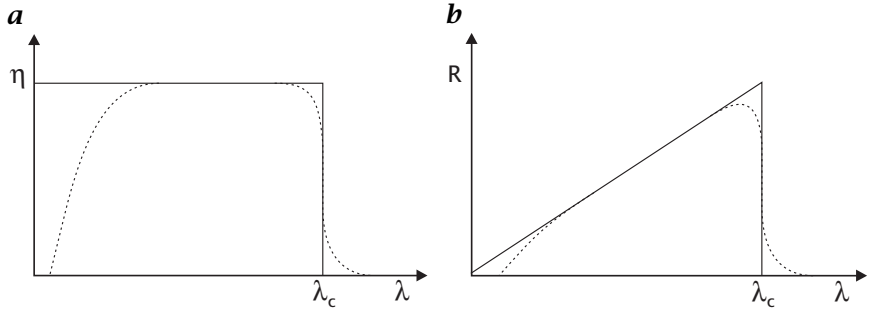


Figure 5.9: *Response of an ideal photodetector.* ***a*** *Quantum efficiency; and* ***b*** *responsivity. Solid lines correspond to ideal detectors and dashed lines to typical departures from ideal curves (After [5]).*

detectors. These quantities also constitute the link between the radiometric quantities of radiation impinging on the detector material and the final electrical detector output.

Quantum efficiency. *Quantum efficiency* $\eta(\lambda)$ relates the number of photons incident on the detector to the number of independent electrons generated. It counts only primary charge carriers directly related to the initial absorption process and does not count electrical amplification. Quantum efficiency takes into account all processes related to photon losses, such as absorptance of the detector material, scattering, reflectance and electron recombination.

In a more general sense, the CIE vocabulary defines quantum efficiency as the ratio of elementary events contributing to the detector output to the number of incident photons. This also accounts for detectors in which no charge carriers are directly released by photon absorption. The quantum efficiency can be expressed as

$$\eta(\lambda) = \frac{n_o}{n_p} \tag{5.34}$$

where n_p is the number of incident photons; n_o defines the number of output events, such as photoelectrons in photodiodes, and electron-hole pairs in semiconductors (Section 5.5.2).

The quantum efficiency is always smaller than one and is commonly expressed in per cent. Figure 5.9a shows the spectral quantum efficiency for an ideal photodetector. The ideal quantum efficiency is a binary function of wavelength. Above a certain *cutoff wavelength* λ_c, photons have insufficient energy to produce photogenerated charge carriers (Section 5.5.2). All photons with higher energy (smaller wavelengths) should produce the same output. Real photodetectors show a slightly different behavior. Near λ_c the thermal excitation of the detector material can affect the production of charge carriers by photon

absorption. Thus, the sharp transition is rounded, as illustrated by the dashed line. Another typical behavior of photodetectors is the decreasing quantum efficiency at short wavelengths.

Responsivity. An important quantity relating the final detector output to the irradiance is the *responsivity*, R, of the detector. It is defined as the electrical output signal divided by the input radiative flux θ:

$$R(\lambda, f) = \frac{V(\lambda, f)}{\phi_\lambda(f)} \tag{5.35}$$

where V denotes the output voltage and f is the temporal frequency at which the input signal is chopped. The frequency dependency accounts for the finite response time of detectors and shows the detector's response to fast changing signals. If the detector output is current, rather than voltage, V has to be replaced by current I. Depending on the type of detector output, the units are given as $\mathrm{V\,W^{-1}}$ (volts per watt) or $\mathrm{A\,W^{-1}}$ (amperes per watt).

For a photon detector (Section 5.5.2), the responsivity can be expressed by the quantum efficiency η and the photon energy $e_p = hc/\lambda$ as

$$R(\lambda) = \frac{\eta \lambda q G}{hc} \tag{5.36}$$

where q denotes the electron charge, $q = 1.602 \times 10^{-19}\,\mathrm{C}$. The *photoconductive gain* G depends on the geometrical setup of the detector element and material properties. The frequency dependent responsivity is given by

$$R(\lambda, f) = \frac{\eta \lambda q G}{hc\sqrt{2\pi f \tau}} \tag{5.37}$$

where τ denotes the time constant of the detector.

The ideal spectral responsivity of a photodetector is illustrated in Fig. 5.9b. As R is proportional to the product of the quantum efficiency η and the wavelength λ, an ideal photodetector shows a linear increase in the responsivity with wavelength up to the cutoff wavelength λ_c, where it drops to zero. Real detectors show typical deviations from the ideal relationship as illustrated by the dashed line (compare to Fig. 5.9a).

Noise equivalent power. Another important figure of merit quantifies the detector noise output in the absence of incident flux. The signal output produced by the detector must be above the noise level of the detector output to be detected. Solving Eq. (5.35) for the incident radiative flux yields

$$\phi_\lambda = \frac{V}{R} \tag{5.38}$$

where R is the responsivity of the detector. The *noise equivalent power* NEP is defined as the signal power, that is, radiative flux, which corresponds to an output voltage V given by the root-mean-square (rms) noise output, σ_n:

$$NEP = \frac{\sigma_n}{R} \tag{5.39}$$

In other words, NEP defines the incident radiant power that yields a signal-to-noise ratio (SNR) of unity. It indicates the lower limit on the flux level that can be measured. It depends on the wavelength of the radiation, the modulation frequency, the optically active detector area, the noise-equivalent electrical bandwidth Δf, and the detector operating temperature. Thus, it depends on a large number of situation-dependent quantities.

Detectivity. The *detectivity* D of a detector is the reciprocal of the NEP:

$$D = \frac{1}{NEP} \tag{5.40}$$

A more useful property can be obtained by incorporating the detector area and the noise-equivalent bandwidth Δf. The corresponding quantity, called *normalized detectivity* D^* or D-star is defined as:

$$D^* = \frac{\sqrt{A_d \Delta f}}{NEP} \tag{5.41}$$

where A_d denotes the optically active detector area. It normalizes the detectivity to a 1-Hz bandwidth and a unit detector area. The units of D^* are $\mathrm{cm\,Hz^{1/2}\,W^{-1}}$, which is defined as the unit "Jones". The normalized detectivity can be interpreted as the SNR of a detector when 1 W of radiative power is incident on a detector with an area of 1 cm.

Again, the normalized detectivity depends on the remaining quantities, the wavelength of the radiation, the modulation frequency, and the detector operating temperature.

5.5.2 Classification of optical detectors

Over the last one hundred years a variety of detectors for electromagnetic radiation have been developed. Recent developments in semiconductor technology have led to an increasing integration of large sensor arrays to produce high-quality focal-plane arrays suitable for computer vision applications. Other types of detectors are used as single-point measuring sensors, which scan the image area to produce higher-dimensional image data sets. Independent from the geometrical

setup, they all rely on inherent changes of a physical property of the detector material by absorption of radiation, which can be quantitatively measured.

According to the underlying physical process of converting radiative energy into an electrical signal, all detectors can be classified into three major types:

1. **Photon detectors.** These types of detectors respond directly to individual photons. Any absorbed photon releases charge carriers in the detector that produce an electric signal. Photon detectors are among the most important sensor types for computer vision applications. They cover the entire range of electromagnetic radiation from x-rays, to ultraviolet and visible light, up to the infrared region. The most prominent examples are photographic films and CCD arrays. Other important applications include light-amplifying cameras, such as micro-channel plate detectors and modern infrared focal plane array cameras.
2. **Thermal detectors.** Optical radiation incident on a thermal detector causes the detector temperature to increase due to the absorbed energy. The increased temperature changes some electrical property of the detector material. The output signal of thermal detectors is proportional to the total energy stored in the detector as opposed to the number of absorbed photons in photon detectors. The wavelength of the radiation is irrelevant, as the same output signal can be produced by photons at different wavelengths if the photon number compensates for the different photon energies. Thus, the responsivity of thermal detectors exhibits a broad wavelength dependency, dominated by the spectral absorptance of the photon-absorbing material.
3. **Coherent detectors.** The third class of detectors directly respond to the electric field strength of the electromagnetic radiation by interference of the electric field of the incident photon with the electric field of a reference oscillator. Coherent detectors can be used only for "low-frequency" radiation, primarily for detection of radio and submillimeter radiation down to the infrared region. Prominent examples of detector systems are radar satellites operating at microwave frequencies and radio telescopes used in astronomy.

In the remainder of this section we will give an overview of the most common detector types, relevant for computer vision, with regard to the principal physical mechanisms and radiometric properties.

5.5.3 Photon detectors

The class of photon detectors contains the most important detector types for computer vision. Apart from a few exceptions, such as pho-

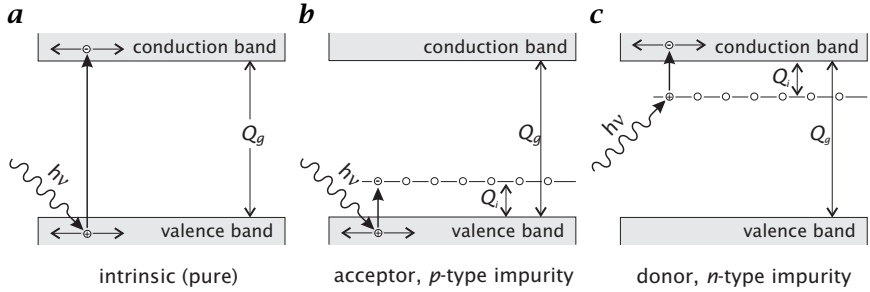


Figure 5.10: *Energy-band diagrams for* ***a*** *intrinsic photoconductors;* ***b*** *extrinsic p-type photoconductors; and* ***c*** *extrinsic n-type photoconductors.*

Table 5.1: *Intrinsic photoconductor materials.* [1] *Values taken from [6].* [2] *Values computed by the author.*

Material	η (%)	λ_c (μm)	T (K)
GaAs[2]	-	0.9	300
Si[2]	-	1.1	300
Ge[2]	-	1.9	300
PbS[1]	50	3	300
PbSe[1]	50	5	300
InSb[2]	-	6.9	77
HgCdTe[1]	60	25	77

tographic films, most photon detectors are solid state detectors, which make use of the fact that electrical properties of semiconductors are dramatically altered by the absorption of ultraviolet, visible and infrared photons.

Intrinsic photoconductors. *Photoconductors* respond to light by either changing resistance or conductance of the detector material. *Intrinsic photoconductors* are the most straightforward way to design a solid state electronic detector. They make use of the inherent electrical property of pure semiconductor materials without additional manipulations. At normal temperatures, relatively few electrons will be in the conduction band of a semiconductor, which results in a low electric conductivity of the material. Figure 5.10a illustrates the energy-band diagram for an intrinsic photoconductor.

In order to move from the valence band into the conduction band, an electron must have sufficient energy. By absorbing a photon whose energy is greater than that of the bandgap energy Q_g, an electronic bond can be broken and the electron can be lifted into the conduction

band, creating an electron/hole pair (Fig. 5.10a). Both the electron and the corresponding hole can migrate through the detector material and contribute to the conductivity. If an electric field is maintained across the detector, any absorbed photon results in a small electric current, which can be measured by a high-impedance amplifier.

As thermal excitation contributes to the conductivity in the same way as absorbed radiation, thermal noise will corrupt the signal, especially at high temperatures and low illumination levels. The number of thermally exited electrons follows the *Boltzmann distribution*:

$$n_t \propto \exp\left(-\frac{Q_g}{k_B T}\right) \tag{5.42}$$

where Q_g, k_B, and T are the bandgap energy, the Boltzmann constant, and the absolute temperature, respectively. As Q_g becomes smaller, the number of thermally exited charge carriers increases. One way to overcome this problem is to cool the detector down to cryogenic temperatures below 77 K (liquid nitrogen temperature), where thermal excitation is negligible.

The minimum photon energy that can be detected is given be the bandgap energy Q_g of the detector material. With the photon energy (Eq. (2.2))

$$e_p = h\nu = \frac{hc}{\lambda} \tag{5.43}$$

the maximum detectable wavelength λ_c, commonly referred to as *cutoff wavelength*, is given by

$$\lambda_c = \frac{hc}{Q_g} \tag{5.44}$$

Substituting for the constants, and correcting for units such that wavelengths are in microns and energy gap in electron volts yields the following rule of thumb:

$$\lambda_c[\mu\text{m}] = \frac{1.238}{Q_g[\text{eV}]} \tag{5.45}$$

Table 5.1 shows some examples of common materials used to manufacture intrinsic photoconductive detectors, together with the quantum efficiency, the cutoff wavelength, and the operating temperature.

Intrinsic photoconductor detectors can be made in large arrays and they have good uniformity and high quantum efficiency, typically in the order of 60 %. They are the basic components of CCD-arrays (charge coupled devices), which are the most widely used 2-D detectors in the visible, the near infrared, and—to some extent—in the x-ray and ultraviolet region using special semiconductor compounds. In the infrared region, semiconductors with a small bandgap have to be used. For highly

Table 5.2: *Extrinsic photoconductor materials.* [1] *Values taken from [6].* [2] *Values taken from [5].*

Material	η (%)	λ_c (μm)	T (K)	Material	η (%)	λ_c (μm)	T (K)
Ge:Hg[1]	30	14	4	Si:Cu[2]	-	5.2	-
Ge:Cu[1]	30	27	4	Si:In[1]	40	8	45
Ge:Be[2]	-	52	-	Si:Be[2]	-	8.3	-
Ge:As[2]	-	98	-	Si:Al[2]	-	18.5	-
Ge:P[2]	-	103	-	Si:Ga[1]	40	19	18
Ge:Ga[2]	-	115	-	Si:As[1]	40	24	4
Ge:B[2]	-	119	-	Si:B[2]	-	28	-
Ge:In[1]	30	120	4	Si:P[1]	40	29	12
Ge:Sb[2]	-	129	-	Si:Sb[2]	-	29	-

energetic radiation, such as x-rays, the energy exceeds the bandgap of any semiconductor. However, the absorption coefficient of most materials is extremely low at these wavelengths, which makes most sensors almost transparent to short-wave radiation. In order to deposit the energy in the detector, the semiconductor material must contain heavy atoms, which have a higher absorptivity in the x-ray region.

Extrinsic photoconductors. For longer wavelengths toward the infrared region, it is hard to find suitable intrinsic semiconductor materials with sufficiently small bandgaps. For wavelengths beyond 15 μm, materials tend to become unstable and difficulties occur in achieving high uniformity and making good electrical contacts. A solution to this problem is to use *extrinsic photoconductors*, that is, semiconductors doped with either p-type or n-type impurities.

The addition of impurities places available electron states in the previously forbidden gap and allows conductivity to be induced by freeing impurity-based charge carriers. Thus, smaller energy increments are required. As illustrated in Fig. 5.10b and c, only the gap between the valence band and the impurity level (p-type semiconductors) or the gap between the impurity level and the conduction band (n-type semiconductors) has to be overcome by absorption of a photon. In the former case, the conductivity is carried by holes and in the latter case free electrons in the conduction band contribute to the conductivity. The basic operation of extrinsic photoconductors is similar to that of intrinsic photoconductors, except that the bandgap energy Q_g has to be replaced by the excitation energy Q_i (Fig. 5.10b and c).

Table 5.2 shows some examples of common materials used to manufacture extrinsic photoconductive detectors, together with the quan-

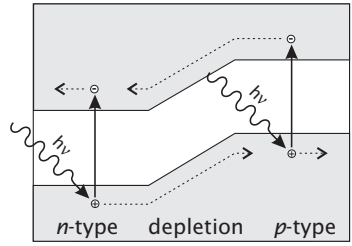


Figure 5.11: *Band diagram of the p-n junction in a photovoltaic detector (photodiode). In the p-type material, photogenerated electrons diffuse into the depletion region and are swept into the n-type region by the electric field. The same process occurs in the n-type material, except the roles of the holes and electrons are reversed.*

tum efficiency, the cutoff wavelength, and the operating temperature. The notation *semiconductor:dopant* is used to indicate the host semiconductor material and the majority dopant (impurity).

Although extrinsic photoconductors are an elegant way to get long wavelength response, they have some less desirable characteristics:

- Due to the smaller bandgap, extrinsic semiconductors are much more sensitive to thermal noise, which can be inferred from Eq. (5.42), and, therefore, require a much lower operating temperature than do intrinsic photoconductors (compare Table 5.1 with Table 5.2).
- Extrinsic photoconductors have a quantum efficiency that is substantially smaller than that of intrinsic materials (30 % compared to 60 % in average). This results from the fact that the impurities are necessarily more sparse than the host material, which leads to a smaller optical absorption cross section.
- The electrical conductivity of extrinsic materials differs fundamentally from that of intrinsic materials. In intrinsic photoconductors, electron/hole pairs are generated by the excitation process, both contributing to the charge transport (Fig. 5.10a). In extrinsic photoconductors, individual charge carriers are generated whose complementary charge resides in an ionized atom, which remains immobile in the crystal structure and cannot carry current (Fig. 5.10a and b).

As the number of semiconductor atoms always outnumbers the impurity atoms, the intrinsic effect dominates in both types of extrinsic material at high temperatures (where all impurity charge carriers are thermally excited) and for wavelengths smaller than the cutoff wavelength of the intrinsic material. To reduce the response from intrinsic conduction, all wavelengths below the anticipated long-wave radiation have to be blocked by spectral filters.

Table 5.3: *Photovoltaic (photodiode) detector materials. Values taken from [6].*

Material	η (%)	λ_c (μm)	T (K)
GaAsP	60	0.7	300
Si	65	1.1	300
InGaAs	86	1.7	300
Ge	64	1.8	300
InAs	40	3.3	77
HgCdTe	65	5	77
InSb	45	5.5	77

Photodiodes (photovoltaic detectors). A *photovoltaic detector* actively generates a voltage or current from incident electromagnetic radiation. The most common realization is based on a junction between two oppositely doped zones (p-n junction) in a semiconductor material. As this setup acts as a diode, this type of detector is also called *photodiode*.

Photodiodes allow large resistance and simultaneously high photoconductive gain within a small volume to be obtained. The n-type material has a surplus (and the p-type material has a deficiency) of electrons compared to the crystal bond of the semiconductor material. In the adjacent region of both oppositely doped zones, electrons migrate from the n- to the p-region acceptor atoms and holes migrate from the p- to the n-region donors, if thermal excitation frees them. Within the contact region all bonds are complete and the material is depleted of potential charge carriers. This results in a high resistance of this region, as opposed to the relatively high conductivity of the p- and n-type material. As the charge carriers diffuse, a voltage is established across the depletion region, called the *contact potential*, which opposes the diffusion of additional electrons. The net result is a permanent equilibrium voltage across the p-n junction. The resulting bandstructure across the contact zone is shown in Fig. 5.11.

Table 5.3 shows some examples of common materials used to manufacture photodiode detectors, together with the quantum efficiency, the cutoff wavelength, and the operating temperature.

When photons of energies greater than the forbidden gap energy are absorbed in, or close to a p-n junction of a photodiode, the resulting electron/hole pairs are pulled by the electric field of the contact potential across the p-n junction. Electrons are swept from the p-region into the n-region, and holes in the opposite direction (Fig. 5.11). As the charge carriers are spatially separated across the detector, a resulting

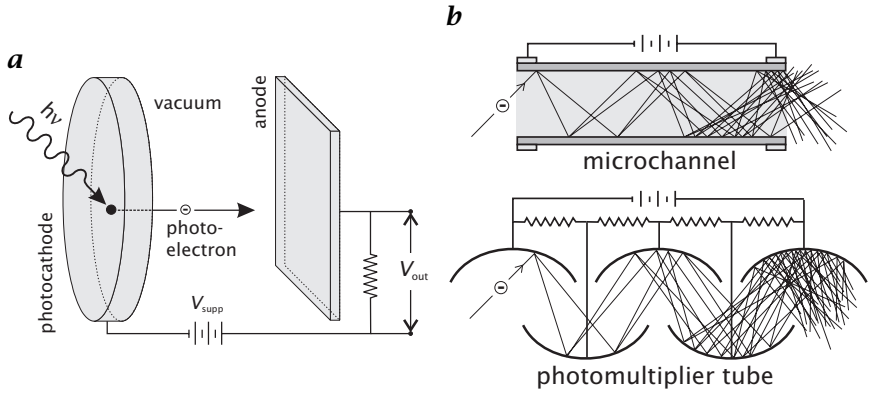


Figure 5.12: *Photoemissive detectors.* ***a*** *Detection process for a vacuum photodiode;* ***b*** *light amplification by a microchannel (top) and a photomultiplier tube (bottom).*

voltage can be measured. If the n- and the p-type region are connected, a small current will flow between both regions. This phenomenon is called the *photovoltaic effect*.

Because photodiodes operate through intrinsic rather than extrinsic absorption, they can achieve a high quantum efficiency in small volumes (Table 5.3). Photodiodes can be constructed in large arrays of many thousands of pixels. They are the most commonly used detectors in 1-6-μm region [5] (e. g., InSb infrared focal plane arrays) and are also used in the visible and near ultraviolet.

Photoemissive detectors. *Photoemissive detectors* operate with external photoelectric emission. The excited electron physically leaves the detector material and moves to the detecting anode. Figure 5.12a illustrates the principal setup. A conduction electron is produced in the photocathode by absorption of a photon with an energy greater than the intrinsic bandgap of the detector material. This electron diffuses through the detector material until it reaches the surface. At the surface of the photocathode it might escape into the vacuum. Using an electric field between the photocathode and the anode helps to accelerate the electron into the vacuum, where it is driven towards the anode and counted as current.

Suitable photocathode materials must have the following properties:

- high-absorption coefficient for photons
- long mean-free path for the electron in the cathode material (low transport losses of electrons migrating to the surface of the cathode)

Table 5.4: *Photocathode materials. Values taken from [6].*

Material	η (%)	λ_c (μm)
GaAsP (NEA)	30	0.9
Cs-Nag-K-Sb (S20)	20	0.9
Ag-O-Cs (S1)	1	1.1

- low electron affinity, that is, low barrier inhibiting the electron emission

Table 5.4 summarizes common materials used for the fabrication of photocathodes in photoemissive detectors.

The simple vacuum photodiode, illustrated in Fig. 5.12a, can be improved by electron multipliers, increasing the number of electrons contributing to the output current for each detected photon. A commonly used photoemissive detector is the *photomultiplier*, illustrated in Fig. 5.12b. It consists of a vacuum tube including several intermediate anodes. Each anode, called a *dynode*, is given a voltage higher than the previous one. The geometrical arrangement is such that emitted electrons are accelerated towards the next adjacent dynode. If the voltage difference is high enough, each photoelectron leaving a dynode gets fast enough to eject multiple electrons from the next dynode upon impact. This process is repeated until the avalanche of electrons finally reaches the anode. The voltages required for operation are provided by a single supply, divided by a chain of resistors. The photocathode is held at a large negative voltage in the order of several thousand volts relative to the anode.

Photomultipliers are large devices, restricted mainly to single detectors. A different form of electron multipliers, which is of practical relevance for computer vision, are made from thin tubes of lead-oxide glass. These microchannels have diameters of 8-45 μm and a length-to-diameter ratio of about 40 [5], and are suitable for integration into small-scale detector arrays. *Microchannel plates* are arrays of approximately one million channel electron multipliers, fused into solid wavers [7]. Figure 5.12b illustrates the principal mechanism of a single microchannel. The microchannel wall consists of three layers: an emitting layer; a conducting layer; and bulk glass. The conductive layer has a high resistance and allows a large voltage to be maintained across the ends of the tube. Electrons that enter the tube are accelerated along the tube until they collide with the wall. The inner surface layer, called the emitting layer, is made from PbO, which acts as an electron multiplier. Upon impact, the accelerated electrons create multiple secondary electrons that are accelerated by the voltage along the tube until they

strike the walls again and produce more free electrons. This operation is comparable to a continuous dynode chain and the gains are nearly as large as those of photomultipliers.

Microchannel plates are used in modern light intensifying cameras, suitable for low-illumination applications, such as fluorescence imaging and night vision devices.

5.5.4 Thermal detectors

The first detectors discovered were thermal detectors, which showed a response to the heating effect of radiation. Unlike photon detectors, they do not respond to charge carriers, directly excited by absorbed photons. Instead, the thermal energy of absorbed photons is detected by temperature-dependent physical processes. A thermal detector can be thought of as two essential parts: the absorber and the temperature sensor.

It is important to note that the net energy stored by absorption is given by the photon energy times the number of absorbed photons. Thus, low-energy photons can create the same detector output as high-energy photons, if the photon flux is higher and compensates for the lower energy. For this reason, the spectral response of thermal detectors is flat and determined by the spectral dependence of the surface absorptance.

Thermal detectors are either bulk devices or metal junction devices. The junction devices, such as the thermocouple and thermopile, rely upon the *Seebeck effect* or *thermoelectric effect*. Two separate junctions of two dissimilar metals generate a voltage proportional to the difference in temperature between them [1]. If one junction is kept at reference temperature, the series output will be proportional to the temperature of the other junction. In practical realizations of thermocouples, one junction is embedded into an absorbing material, while the other junction is thermally connected to the radiometer housing with a high thermal mass. *Thermopiles* are series of individual thermocouples, which substantially increases the sensitivity.

While thermopiles are mostly used as single detectors, another type of thermal detector, called a *bolometer*, is a bulk-type detector and can be easily integrated into large detector arrays. Bolometers take advantage of the high-temperature coefficient of resistance in semiconductors, which is similar to the principle of photoconductors. A detailed treatment of recent developments in the fabrication of microbolometer arrays is given in Chapter 10.

Recent developments in high-temperature (about 77 K) superconductivity made another type of thermal detectors available, which relies on the sharp resistance change with temperature in the superconducting transition region. These superconducting bolometers can also be

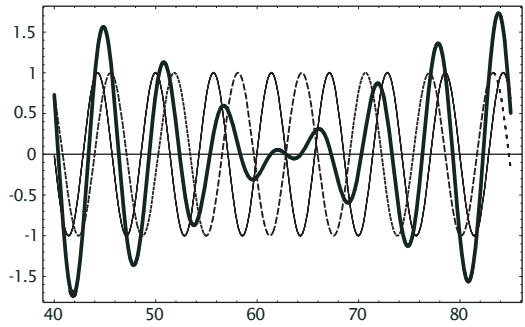


Figure 5.13: *Mixing of two periodic signals S_i, and S_m with slightly different wavelengths, $\lambda_i = 1.1\,\lambda_m$. The bold line shows the resulting signal $S = S_i + S_m$. The amplitude of the mixed signal is modulated by the difference, or* beat, *frequency.*

operated in two other modes that involve the breaking of *Cooper pairs* by the incident photons, thus destroying superconductivity [6].

Coherent detectors. Coherent receivers directly measure the electromagnetic field of the incident radiation. They mix the electromagnetic field of the incoming photons with an internal reference field of similar frequency, produced by a high-frequency oscillator. The resulting signal shows a strong modulation of the amplitude, which is given by the difference frequency of both signals—a physical effect commonly referred to as *beating*.

Let S_i and S_m be the incident, and the mixing signal (electric field), respectively, given in complex notation by

$$S_m = A_m \exp[i\omega t], \quad \text{and} \quad S_i = A_i \exp[i(\omega + \epsilon)t] \tag{5.46}$$

where ϵ is a small frequency shift compared to the main frequency ω. Linear superposition yields the following mixed signal:

$$\begin{aligned} S &= S_m + S_i &&= A_m \exp[i\omega t] + A_i \exp[i(\omega + \epsilon)t] \\ &&&= \exp[i\omega t]\,(A_m + A_i \exp[i\epsilon t]) \end{aligned} \tag{5.47}$$

which can be interpreted as an oscillation at the frequency ω, with an amplitude modulation at the difference (beat) frequency ϵ. This effect is illustrated in Fig. 5.13.

From the mixed field, the exact frequency can be extracted, as well as the amplitude and phase of the incident signal. In order to measure the electric field, the mixed field has to be passed through a nonlinear electrical element, called *mixer*, that converts power from the original frequency to the beat frequency.

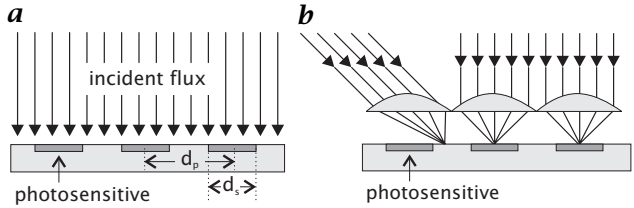


Figure 5.14: *Schematic illustration of the fill factor and microlens arrays on detector arrays.* ***a*** *Detector without a microlens array;* ***b*** *Detector with a microlens array.*

Unlike all other types of (incoherent) receivers, these *coherent* receivers obtain additional information about the wave number and phase of the signal. As the phase information is given, they can correlate measurements of different receivers to reconstruct the incoming wavefront by interferometry. Intercontinental baseline radio telescopes use this ability to combine several telescopes spread over the entire globe to enhance the resolution up to milliarc-seconds for astronomical applications.

A more detailed treatment of the theory of coherent receivers can be found in [8] and [5].

5.5.5 Characteristics of detector arrays

Fill factor. Most detector arrays used in computer vision are not photosensitive over the entire detector area. As all electrical contacts and microelectronic components have to be integrated into the chip surface, only a small portion is retained for the actual photosensitive detector area. Exceptions are 1-D detector arrays, where all electronic components and bonds can be arranged alongside the detector, or back-illuminated detector arrays.

The basic quantities defining the *fill factor* of the sensor are the pixel pitch d_p, which describes the center distance of two neighboring pixels, and the pixel size d_s, which is the extension of the photosensitive area. For nonsquare pixels, the dimensions on both directions have to be known.

Given a local irradiance $E_i(\boldsymbol{x})$ on the sensor, only the portion

$$E(\boldsymbol{x}) = E_i(\boldsymbol{x}) \frac{d_s^2}{d_p^2} \tag{5.48}$$

actually contributes to the signal at the point $\boldsymbol{x}$ (Fig. 5.14a). For nonsquare pixels/arrays, the squared quantities have to be replaced by the products of the corresponding quantities in the x- and y-direction, respectively.

Microlens arrays. A common technique to overcome the problem of reduced fill factor is to place microlens arrays over the detector area. An optimal microlens array covers the entire sensor surface, such that incident radiation is focused onto the individual photosensitive areas, as illustrated in Fig. 5.14b. In that way, the maximum possible radiative flux can be collected with low fill factors.

There are, however, two basic problems that have to be traded in, even for perfectly transparent lens-arrays:

- The incident radiation is focused onto a spot smaller than the photosensitive area, with the exact position depending on the angle of incidence (Fig. 5.14b). If the photosensitive area exhibits local inhomogeneities in the sensitivity, the detector output shows an angular dependence, given by the sensitivity distribution of the photosensitive area.
- For large angles of incidence, it might happen that the incident radiation is focused onto a point in between two photosensitive areas (Fig. 5.14b). Thus, the angular response suddenly drops to zero for a certain cutoff angle. This effect can be avoided if the geometric setup is such that no radiation beyond the critical angle can enter the optical system. The larger the focal lens of the optical system is, the smaller the maximum inclination angle.

Static noise pattern. It is impossible to manufacture large detector arrays in such a way that all individual sensor elements will be absolutely identical. Each pixel usually exhibits slightly different sensitivities, offsets, and gains. Thus, even absolutely uniform surfaces are imaged according to the intrinsic structure of the sensor array inhomogeneities. These patterns overlay all images and constitute some kind of "noise". Unlike other types of noise, this *fixed-pattern noise* is static and remains stable over a certain time span.

In principle, the fixed-pattern noise can be corrected for by radiometric calibration of the sensor. This procedure is commonly referred to as *flat fielding*, as a surface with uniform radiance is used to compute the local inhomogeneities.

If the fixed-pattern noise remains stable over the expected lifetime of the camera, it can be calibrated once by the manufacturer, and all pixel readouts can be automatically corrected for local offsets and gains. If the static noise pattern changes over longer periods, it might be necessary to repeat the calibration procedure more frequently.

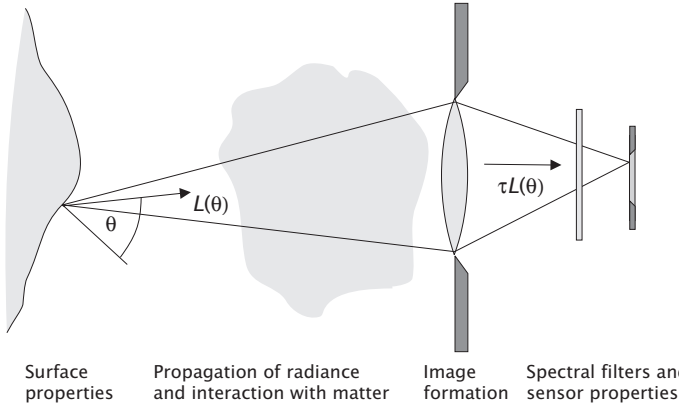


Figure 5.15: *The chain of radiometric imaging.*

5.6 Concluding summary

This chapter concludes with a summary of the basic results of the previous considerations about quantitative radiometry of imaging. Figure 5.15 summarizes the chain of events leading from emission of radiation to the final image formation.

The basic steps and results can be summarized as follows:

1. The detected flux is proportional to the radiance of the emitting surface with a proportionality constant given by the geometry of the optical setup.
2. The radiance stays invariant as it propagates through space. Thus, the radiometric measurement can be carried out at any position along the direction of propagation. This result, however, assumes that no losses occur along the propagation path. For effects such as scattering, absorption, refraction, etc., the radiance is decreased according to the interaction of radiation with matter (this was presented in Chapter 3).
3. The radiance is changed at the transition of interfaces separating two media with different refractive indices. In case the radiation penetrates a second interface (into a medium with the same refractive index as the initial one), this process is reversed. Thus, the initial radiance is restored after passing a lens system, but attenuated by the transmittance of the optical system.
4. By optical imaging, the radiance entering a camera lens is converted into irradiance of the detector. The irradiance distribution on the

detector plane shows a natural field darkening with decreasing irradiance towards the edges of the detector. This field darkening can be further amplified by vignetting and other optical effects blocking parts of the radiation.

5. The final output of the imaging detector depends on a variety of detector properties. If the conversion from incident flux to an electrical signal is linear, the output remains proportional to the object irradiance.

5.7 References

[1] McCluney, W. R., (1994). *Introduction to Radiometry and Photometry*. Boston: Artech House.

[2] Siegel, R. and Howell, J. R. (eds.), (1981). *Thermal Radiation Heat Transfer*, 2nd edition. New York: McGraw-Hill Book, Co.

[3] Wolfe, W. L. and Zissis, G. J. (eds.), (1989). *The Infrared Handbook*, 3rd edition. Michigan: The Infrared Information Analysis (IRIA) Center, Environmental Research Institute of Michigan.

[4] Nicodemus, F. E., (1963). Radiance. *Am. J. Phys.*, **31**:368–377.

[5] Rieke, G. H., (1994). *Detection of Light: From the Ultraviolet to the Submillimeter*. Cambridge: Cambridge University Press.

[6] Dereniak, E. L. and Boreman, G. D., (1996). *Infrared Detectors and Systems*. New York: John Wiley & Sons, Inc.

[7] Laurin Publishing, (1998). *The Photonics Design and Applications Handbook*, 44th edition. Pittsfield, MA: Laurin Publishing CO.

[8] Torrey, H. C. and Whitmer, C. A., (1948). *Crystal Rectifiers*, Vol. 15. New York: Massachusetts Institute of Technology Radiation Laboratory Series, McGraw-Hill.

6 Illumination Sources and Techniques

Horst Haußecker

Interdisciplinary Center for Scientific Computing
University of Heidelberg, Heidelberg, Germany

6.1 Introduction

In Chapters 2 and 3 the basics of radiation and the interaction of radiation with matter were introduced. How radiation is emitted from active sources and how incident radiation interacts with passive surfaces of objects in the scene were both demonstrated. However, we did not specify the characteristics of real radiation sources.

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

In this chapter we turn towards the question: How is the irradiance of surfaces generated in practical applications? We will introduce the most important radiation/illumination sources used in computer vision. After a short treatment of natural sources (such as solar and sky irradiance in Section 6.2), we will emphasize artificial sources for scientific applications and machine vision in Section 6.3.

The most important properties of illumination sources that have to be considered for practical applications are:

- spectral characteristics
- intensity distribution
- radiant efficiency (Section 2.4.3)
- luminous efficacy (Section 2.4.3)
- electrical properties
- temporal characteristics
- package dimensions

We will summarize these characteristics for each illumination source, depending upon applicability.

Single illumination sources alone are not the only way to illuminate a scene. There is a wealth of possibilities to arrange various sources geometrically, and eventually combine them with optical components to form an illumination setup that is suitable for different computer vision applications. In Section 6.4 we will show how this can be accomplished for some sample setups. The importance of appropriate illumination setups cannot be overemphasized. In many cases, features of interest can be made visible by a certain geometrical arrangement or spectral characteristics of the illumination, rather than by trying to use expensive computer vision algorithms to solve the same task, sometimes in vain. Good image quality increases the performance and reliability of any computer vision algorithm.

6.2 Natural illumination

For outdoor scenes, natural illumination sources, such as solar irradiance and diffuse sky irradiance, play an important role. In some applications, they might be the only illumination sources available. In other cases, they are unwanted sources of errors, as other illumination sources have to compete with them. Solar irradiance, however, is hard to overcome, as it covers the entire spectrum from the ultraviolet to the far infrared and has an enormous power in the order 10^3 Wm^{-2}, which is hard to achieve with artificial sources.

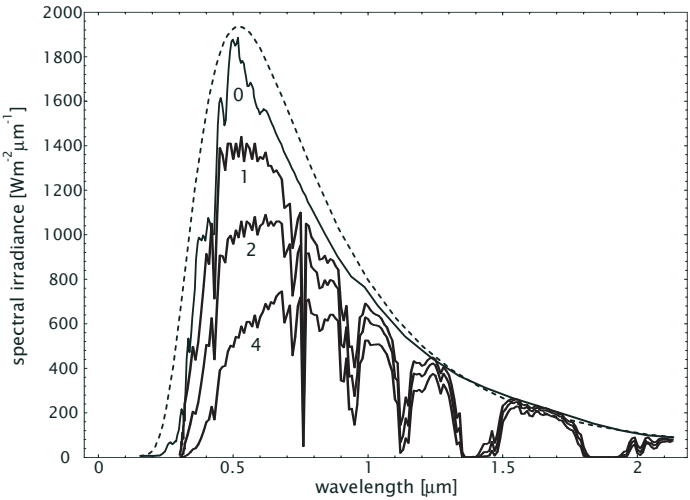


Figure 6.1: *Solar irradiance: Comparison of the solar spectrum (solid lines) at the top of the earth's atmosphere to a blackbody at a temperature of 6000 K (dashed line). Solar irradiance at sea level measured in multiples of the vertical path through standard atmosphere, denoted as ma. The figure shows the irradiance for ma = 0, 1, 2, and 4. With ma = 0, we denote the solar irradiance right above the earth's atmosphere, that is, without atmospheric absorption.*

6.2.1 Solar radiation

Although *solar radiation* has the principal shape of blackbody radiation (Fig. 6.1), the real origin is nuclear fusion rather than incandescence. Powered from internal nuclear power, the outer regions of the sun, heated up to a temperature of approximately 6000 K, emit thermal radiation. On its way through the colder parts of the solar atmosphere the radiation is subject to absorption (Section 3.4) from gases, which shows up as narrow absorption lines, known as *Fraunhofer lines*. These characteristic line spectra allow remote measurements of the presence and concentration of extraterrestrial gases along the optical path.

Within the earth's atmosphere additional absorption occurs. At sea level parts of the solar emission spectrum are extinguished while others remain almost unchanged (Fig. 6.1b). The latter parts are called *atmospheric windows* and are of major importance for long distance remote sensing. One example is the *visible window*, which is of major importance for terrestrial life. Strong absorption regions visible in the solar spectrum at sea level at about 0.9 μm, 1.1 μm, 1.4 μm, and 1.9 μm (Fig. 6.1b), are caused by water vapor (H_2O) and carbon dioxide (CO_2).

Another major attenuation line of CO_2 is located in the IR part of the spectrum at about 4.3 μm. This absorption line is of major impor-

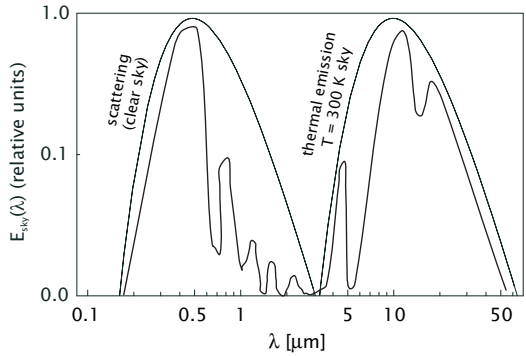


Figure 6.2: *Schematic illustration of the contributions from scattering and atmospheric emission to the diffuse background radiation.*

tance for the *greenhouse effect*, responsible for *global warming*. The increasing concentration of CO_2 in the atmosphere causes an increasing reabsorption of longwave IR radiation, which is emitted from the earth's surface, and thus increased heating up of the atmosphere.

The *radiation luminous efficacy* of solar irradiation can be determined to be approximately 90-120 $\mathrm{lm\,W^{-1}}$ for the lowest angle of incidence (midday).

6.2.2 Diffuse sky irradiation

In addition to direct solar irradiation, natural illumination consists of *diffuse sky irradiation*, commonly referred to as sky-background radiation. It is caused by two major contributions: *scattering* of the sun's radiation for wavelengths shorter than 3 μm; and *thermal emission* from the atmosphere for wavelengths beyond 4 μm (Fig. 6.2).

Depending on the cloud coverage of the sky, different scattering mechanisms dominate. As already outlined in Section 3.4.1, the two basic mechanisms are *Rayleigh scatter*, for particles smaller than the wavelength, such as atmospheric molecules, and *Mie scatter*, for particles with sizes about the wavelength of the radiation, such as microscopic water droplets. The solar scattering region dominates for wavelengths shorter than 3 μm because it is restricted to the region of solar irradiance. The spectral distribution changes depending on the scattering mechanism. For clear sky, Rayleigh scattering dominates, which has a λ^{-4} wavelength dependence. Thus short wavelengths are more efficiently scattered, which is the reason for the blue appearance of the clear sky. For cloud-covered parts of the sky, Mie scatter dominates the solar region. As this type of scattering shows a weaker wavelength dependency (which is responsible for the greyish appearance of clouds),

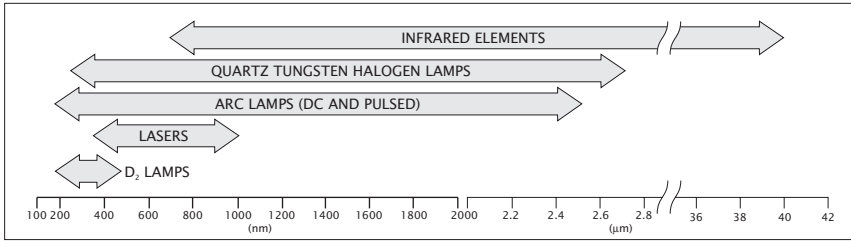


Figure 6.3: *Usable wavelength regions for commercially available illumination sources (Courtesy Oriel Corporation, 1994).*

the scatter spectrum is more closely approximating the solar spectrum, attenuated by the transmittance of the clouds. Additionally, the solar region of the scatter spectrum is modified by a number of *atmospheric absorption bands*. These are mainly the bands of water vapor at 0.94, 1.1, 1.4, 1.9, and 2.7 μm, and of carbon dioxide at 2.7 μm. The effect of these bands is schematically shown in Fig. 6.2.

The thermal region of the sky-background beyond 4 μm is represented by a 300 K blackbody irradiance. Figure 6.2 shows the corresponding blackbody curve. In this region, the absorption bands of the atmosphere have an inverted effect. Bands with strong absorption have a strong emission and will approach the blackbody curve appropriate to the temperature of the atmosphere. Conversely, bands with high transmissivity have correspondingly low emissivity and thus contribute only a small fraction of the blackbody irradiance. This effect is schematically shown in Fig. 6.2.

It is important to note, that the exact shape of the sky-background irradiance strongly depends on the elevation angle of the sun, as well as on meteorological parameters, such air humidity, air temperature, and cloud distribution.

6.3 Artificial illumination sources

Although being the basic physical process used in a large variety of illumination and radiation sources, thermal emission of radiation (Section 2.5) is only one among other possible mechanisms generating radiation. In this section, the most important commercial radiation and illumination sources are introduced, together with the underlying physical processes of radiation emission, practical implementation, and specifications.

Commercially available illumination sources cover the entire spectral range from the ultraviolet to the mid-infrared region. They are manufactured in a variety of package sizes and geometrical arrangements,

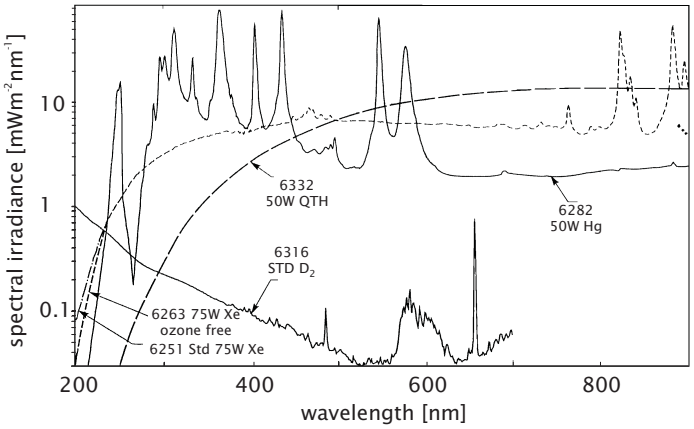


Figure 6.4: *Overview of spectral irradiance curves for arc, quartz tungsten halogen, and deuterium (D_2) lamps at a distance of 0.5 m (Courtesy Oriel Corporation, 1994).*

optimized for specified applications. Figure 6.3 shows an overview of available illumination sources for different spectral regions. In the following sections we will focus on the following illumination sources:

- incandescent lamps
- (arc) discharge lamps
- fluorescent lamps
- infrared emitters
- light emitting diodes (LED)
- laser

A more detailed overview can be found in [1], [2], and in catalogs of manufacturers, such as the one from the Oriel Corporation [3].

6.3.1 Incandescent lamps

Incandescent lamps are among the most popular all-purpose illumination sources. The most prominent examples are standard light bulbs used in almost every household. The classic light bulb uses a carbon filament, which is placed in an evacuated glass enclosure in order to avoid oxidation (burning) of the carbon filament.

More modern versions of incandescent lamps use tungsten filaments instead of carbon fibers. The practical setup of tungsten incandescent lamps are tungsten filaments of various shapes (rectangular dense and coiled filaments) in quartz glass envelopes (Fig. 6.5). The coiled filaments have an intensity distribution of circular symmetry about the

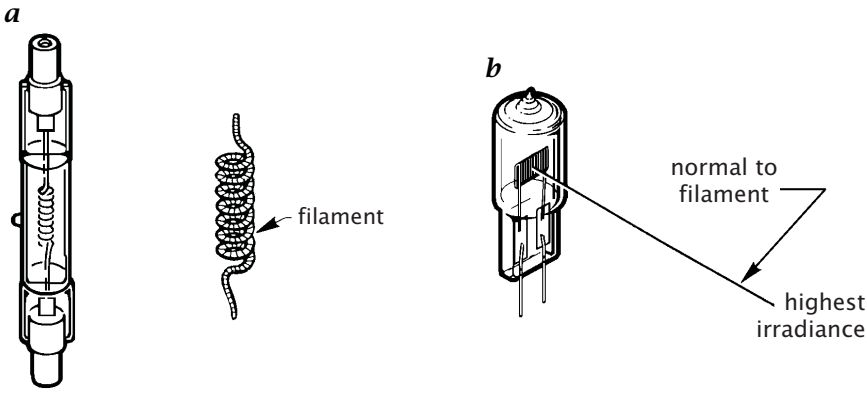


Figure 6.5: *Quartz tungsten halogen incandescent lamps:* ***a*** *setup of a coiled filament lamp;* ***b*** *setup of a rectangular filament lamp (Courtesy Oriel Corporation, 1994).*

long axis of symmetry of the housing. For the rectangular filaments, the light output strongly depends on the direction (Fig. 6.5b). The quartz glass housing is transparent only for wavelengths up to 3 μm. It does, however, heat up by absorption of long-wave radiation and thermally emits infrared radiation corresponding to the glass temperature extending the spectrum into the mid-infrared region.

Incandescent lamps have a high visible and near infrared output. With an emissivity of tungsten of about $\tilde{\varepsilon} = 0.4$ (in the visible), the spectral exitance of tungsten incandescent lamps is close to the exitance of a graybody. It does, however, deviate for wavelengths of about the peak wavelength and above. Figure 6.6 shows the spectral exitance of an incandescent tungsten surface, compared to a graybody with an emissivity of $\tilde{\varepsilon} = 0.425$ at a color temperature of 3100 K.

The *radiant efficiency* of incandescent lamps is in the order of 80 %, as incandescence very efficiently converts electrical input power into radiant output. The output within the visible region, however, is much lower. Operated at a color temperature of approximately 3000 K, tungsten incandescent lamps have a relatively low *radiation luminous efficacy* of $K_r = 21.2\,\mathrm{lm\,W^{-1}}$, as the main part of the spectrum lies in the infrared (Section 2.5.4). The *lighting system luminous efficacy* is only $K_s = 17.4\,\mathrm{lm\,W^{-1}}$. The values are taken for an individual tungsten incandescent light bulb [4] and are subject to fluctuations for individual realizations.

Two important modifications allow both radiant efficiency and the lamp life to be increased:

1. In all tungsten filament lamps, the tungsten evaporates from the filament and is deposited on the inside of the envelope. This blackens

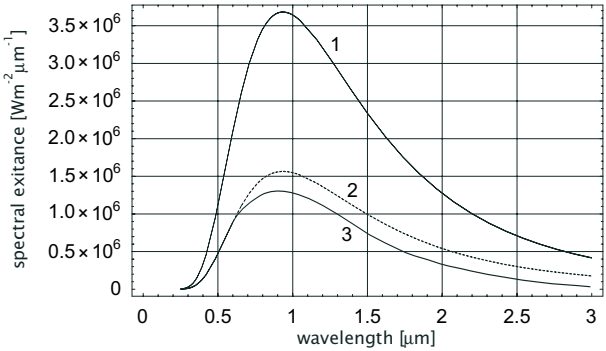


Figure 6.6: *Spectral exitance of (1) a blackbody; (2) a graybody with emissivity of $\epsilon = 0.425$; and (3) a tungsten surface, all at a temperature of 3100 K (Courtesy Oriel Corporation, 1994).*

the bulb wall and thins the tungsten filament, gradually reducing the light output. With tungsten halogen lamps, a halogen gas is filled into the envelope. The halogen gas efficiently removes the deposited tungsten and returns it to the filament, leaving the inside of the envelope clean, and providing long-term stability. This thermochemical process is called the *halogen cycle* [3].

2. Some manufacturers produce new-generation halogen lamps with infrared coatings on the envelope. These coatings are made such that infrared radiation is reflected back onto the tungsten filament. Thus, the temperature of the envelope and the infrared output of the lamp are reduced, which increases luminous efficacy. At the same time, the filament is heated by the emitted infrared radiation, which yields a higher radiant efficiency, as less current is needed to maintain the operating temperature. Both effects increase the lighting system luminous efficacy.

As the exitance of an incandescent lamp is given by the temperature, which does not immediately follow changes in the voltage, the light output does not follow rapid (kHz) voltage changes. It does, however, follow slow voltage changes, such as the net frequency under ac operation, with an amplitude in the order of 10% of the absolute exitance [3]. This effect might cause beating effects, if the frame rate of the video camera is at a similar frequency. For demanding radiometric applications it is recommended to use regulated dc power supplies. The smaller the filament, the lower the thermal mass and the faster the response of the lamp.

6.3.2 Discharge lamps

Discharge lamps operate on the physical principle of gas discharge. At low temperatures, such as ambient temperature and below, gases are nonconducting. The gas molecules are neutral and can not carry electrical current. In a statistical average, a small number of molecules is ionized due to natural radioactivity. These ions, however, have very short lifetimes and immediately recombine. In gas discharge lamps, a strong electric field is generated in between two electrodes, separated by distance d. Within this field, randomly generated gas ions are accelerated towards the electrodes of opposite charge. Upon impact on the cathode, the positively charged gas ions release electrons, which in turn are accelerated towards the anode. These electrons eventually hit other atoms, which can be excited and recombine under emission of light, corresponding to the difference between two energy levels.

Spectral lamps. *Spectral lamps* are plain gas discharge lamps without additional fluorescence coatings, as opposed to fluorescence lamps. As the energy levels of the light emission in gas discharge are characteristic for the gas molecules, gas discharge lamps emit the characteristic line spectra of the corresponding fill gas. A prominent example is the low-pressure sodium vapor lamp used for street illuminations. The bright yellow light corresponds to the Na-D line at a wavelength of 590 nm. Because the spectral exitance consists of a single line in the visible spectrum, the sodium vapor lamp has an extremely high *radiant luminous efficacy* of $524.6\,\mathrm{lm\,W^{-1}}$ (Osram GmbH). Accounting for the electrical power consumption yields a net *lighting system luminous efficacy* of $197\,\mathrm{lm\,W^{-1}}$.

In order to increase the luminous efficacy, the gas pressure within the lamp can be increased by allowing the bulb to heat up. As a consequence, the spectral lines of the exitance are widened. In extreme cases the spectral distribution shows a continuum without spectral lines.

Other examples of fill gases of discharge lamps are xenon (Xe), mercury (Hg), and mixtures of Xe and Hg. The spectral exitance of these gas discharge lamps is similar to that of arc lamps with the same fill gases (shown in Fig. 6.4).

Fluorescent lamps. The spectral output of gas discharge lamps, such as Xe or Hg lamps, shows a high contribution from the ultraviolet region well below 400 nm. Radiation at these wavelengths is invisible, causes severe sunburn, and damages the tissue of the eye's retina.

Fluorescent lamps are discharge lamps (usually filled with Hg) that are additionally coated with special fluorescent materials. These layers absorb ultraviolet radiation and convert it into longer wavelength radiation in the visible region, which is finally emitted. The exact spec-

tral content of the emitted radiation can be varied depending upon the compounds of the fluorescence layer. Examples are lamps with a high content of red light at 670 nm, which is photosynthetically active and can be used as an illumination source for greenhouses.

As the wavelength of light is shifted from about 250 nm towards 500 nm, the energy of the re-emitted radiation is only half the energy of the incident radiation. The remaining energy is absorbed within the fluorescence material. This energy constitutes the main energy loss in fluorescence lamps. Thus, the *lighting system luminous efficacy* is relatively high, compared to incandescent lamps. Typical values of the luminous efficacies are in the order of $K_s = 71\,\mathrm{lm}\,W^{-1}$, and $K_r = 120\,\mathrm{lm}\,W^{-1}$. The *radiant efficiency* lies in the order of $\tilde{\eta} = 50\,\%$. These high values are due to the fact that almost no heat is generated and the major part of the spectrum is emitted in the visible region. Fluorescent lamps are the perfect choice for low-energy room illumination.

For many years tube-shaped fluorescent lamps have been used in both homes and public buildings. Modern developments in lamp manufacturing have led to a huge variety of shapes and color temperatures of fluorescent lamps. They have most recently been advertised as low-energy substitutes for incandescent light bulbs. In order to reduce the size of the lamp and to overcome the elongated shape, narrow tubes are coiled to light bulb-sized compact illumination sources.

All gas discharge lamps are subject to fast brightness fluctuations when operated with ac power supplies. If stable illumination over time is required, these lamps have to be operated with special high frequency power supplies.

6.3.3 Arc lamps

For high currents, the electrodes of discharge lamps get extremely hot. At a certain temperature, the emission of electrons from the cathode is due mainly to incandescence of the electrode material, and the gas discharge is turned into an arc discharge. This effect can be facilitated by a cone shaped cathode, which focuses the electric field.

Xenon and mercury arc lamps. Figure 6.7a shows a diagram and the technical setup of commercial *arc lamps*. The anode and cathode are made of tungsten and sealed in clear quartz glass. The tungsten is doped with materials, such as thoria, to enhance electron emission. When the lamps run, the internal pressure increases to 15-75 bar, depending on the lamp type.

Arc lamps constitute the brightest manufactured broadband sources. The major light output is restricted to the arc, which can be made small depending on electrode geometry. The small radiating area makes

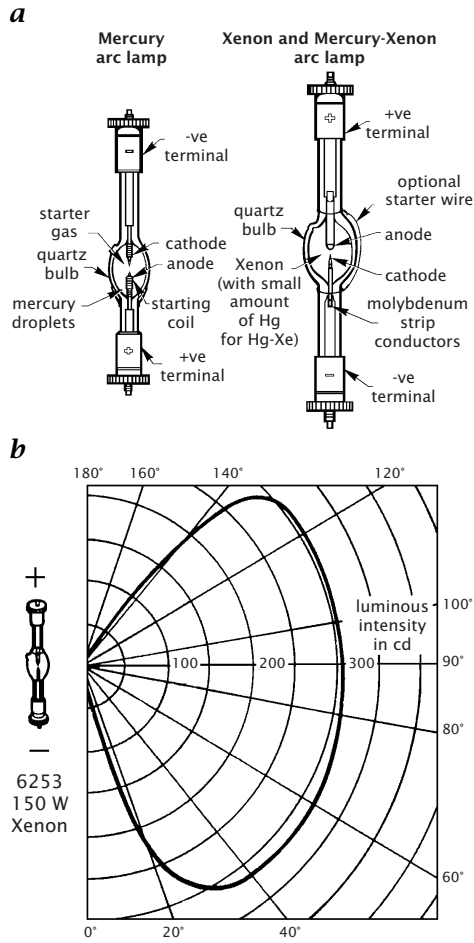


Figure 6.7: *Arc discharge lamps:* ***a*** *construction of arc lamps;* ***b*** *typical luminous intensity distribution of a xenon arc lamp (Courtesy Oriel Corporation, 1994).*

these sources suitable as point sources. The intensity distribution of arc lamps reflects the cylindrical shape of the electrodes and arc. The vertical brightness distribution is shown in Fig. 6.7b. The 3-D distribution is obtained by spinning this distribution about the vertical axis of symmetry of the lamp.

The two most common fill gases of arc lamps are mercury (Hg) and xenon (Xe). Figure 6.4 shows the spectral exitance of both gas types. Both gas types are broad continuum with discrete spectral emission lines. The spectrum of the *xenon arc lamp* closely matches the solar spectrum. The correlated color temperature lies at 5800 K. These lamps

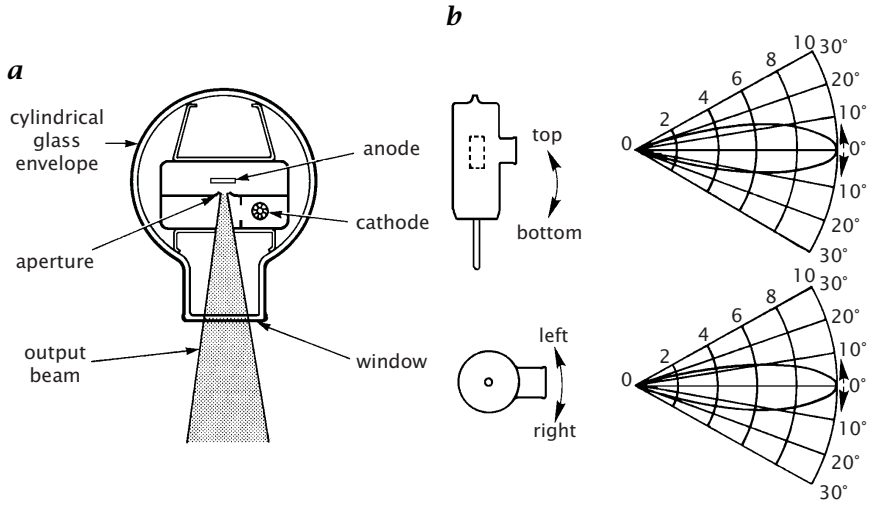


Figure 6.8: *Deuterium lamps:* ***a*** *construction of a deuterium lamp;* ***b*** *typical luminous intensity distribution of deuterium lamp (Courtesy Oriel Corporation, 1994).*

produce a bright white light, which makes xenon arc lamps suitable for solar simulators. They have a relatively smooth continuum from the ultraviolet to the visible region, with few strong emission lines in the near infrared. The luminous efficacy of xenon arc lamps ranges from 15 to 50 $\mathrm{lm\,W^{-1}}$ over a corresponding wattage range of 75 to 10,000 W.

Mercury arc lamps have a strong ultraviolet output with discrete lines and a smooth transition towards the near infrared. The luminous efficacy of mercury arc lamps ranges from 22 to 53 $\mathrm{lm\,W^{-1}}$ over a corresponding wattage range of 200 to 7000 W .

Deuterium arc lamps. If high ultraviolet light output with minimal infrared and visible output is required, *deuterium arc lamps* are the perfect choice. Figure 6.4 shows the spectral exitance of a deuterium lamp. It gradually decreases from a maximum output at 200 nm towards 500 nm. It shows very little output above a wavelength of 500 nm, except for a strong but narrow emission line at 660 nm.

Figure 6.8a shows a diagram and the technical setup of a deuterium lamp, respectively. The intensity distribution of a typical deuterium lamp is illustrated in Fig. 6.8b. It can be observed that these lamps emit directed radiation within a very narrow angular range.

Deuterium lamps emit high-intensity ultraviolet radiation. They have to be operated with extreme caution and protective eyewear and gloves are mandatory when working in the vicinity of these lamps.

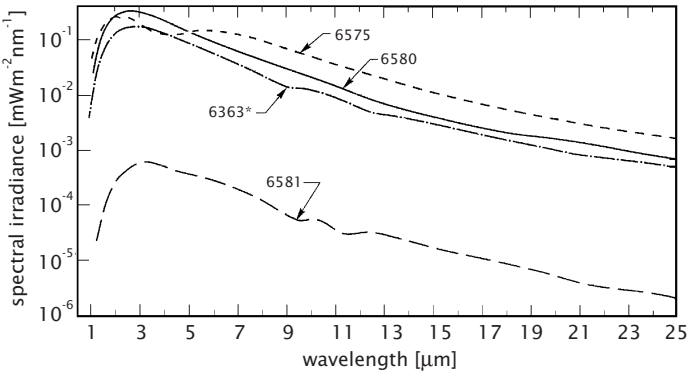


Figure 6.9: *Spectral irradiance of IR sources (Courtesy Oriel Corporation, 1994).*

6.3.4 Infrared emitters

Again, we return to thermal emission of radiation. As already discussed, quartz tungsten halogen lamps are excellent infrared sources, emitting into the far infrared. They do, however, also emit a large portion in the visible and near ultraviolet region. If this radiation is objectionable, other sources of infrared radiation have to be used.

Common infrared sources are basically electrical resistance heaters. The materials must have a high ohmic resistance to allow ohmic heating when an electric current is run through them. The surface must be covered with a material that has a high emissivity in the infrared region. Common materials are metals, ceramics, or carbon rods. Figure 6.9 shows several spectral outputs of commercially available infrared sources.

In the near infrared, a narrowband infrared emitter has become increasingly important: these are the infrared LEDs, which are discussed in Section 6.3.5.

6.3.5 Light-emitting diodes (LEDs)

This section is dedicated to *light-emitting diodes* (LEDs), small but nevertheless powerful light sources, which have gained in importance during the past few years. Originally intended as small signal lights for instrument panels, the performance of LEDs has dramatically increased, while simultaneously the package size has decreased. Light-emitting diodes are available in a huge variety of package sizes and spectral ranges from blue light up to the near infrared.

The most important advantages of LEDs in terms of illumination sources for computer vision can be summarized as follows:

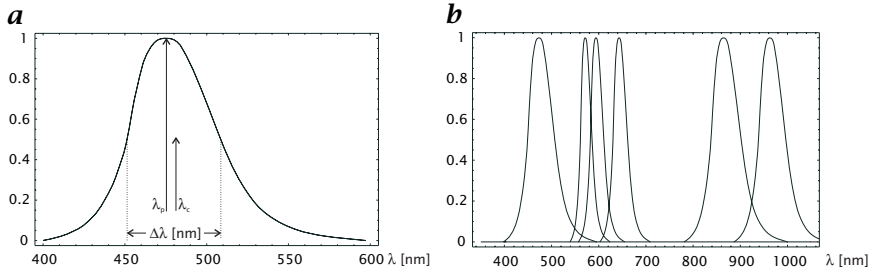


Figure 6.10: *Spectral emissivity of LEDs:* ***a*** *typical relative spectral distribution of an LED showing the location of the characteristic wavelengths and wavelength intervals;* ***b*** *relative spectral distribution of the radiation emitted by a series of typical LEDs and IREDs. Values have been normalized to unity at the peak wavelength.*

- high luminous efficacy
- small dimensions
- can be integrated into large arrays of arbitrarily any shape
- low power consumption, in the order of 20 mW
- fast response time, can be operated in pulsed mode

Basic operation. LEDs operate on the reversed physical process used in photodiodes (see Section 5.5.2). They are made from a semiconductor material with a p-n junction operated in forward bias direction, as opposed to the reverse bias operation of photodiode detectors (see Fig. 5.11). If an external voltage is supplied in forward bias direction, the sign of the applied potential is reversed, so it decreases the bias across the depletion zone. If the bias voltage exceeds the contact voltage, the junction becomes strongly conducting. Charge carriers that penetrate the p-n junction and enter the p- or n material can recombine under emission of radiation. The wavelength of the emitted radiation is given by the bandgap energy of the intrinsic semiconductor material. Due to thermal excitation and impurities, the potential energy transitions are not restricted to the bandgap energy but distributed about this energy. Thus, emitted radiation is not fully monochromatic. It is, however, limited to a narrow spectral range.

Spectral distributions. Figure 6.10a illustrates the typical shape of the spectral distribution of an LED exitance E_λ. The characteristic wavelengths are the *peak wavelength* λ_p and the *centroid wavelength* λ_c

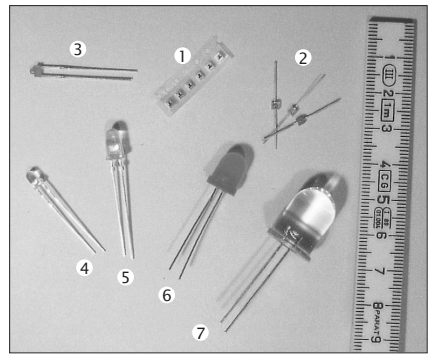


Figure 6.11: *Examples of various package shapes and LED types. (1) Superbright SMD miniature LED HSMC-H670 (Hewlett Packard); (2) Superbright SMD LED HLMP-Q105 (Hewlett Packard); (3) Miniature LED L10600ID (Kingbright); (4) 2-mm package LED (noname); (5) 3-mm package LED HLMA-CH00 (Hewlett Packard); (6) 5-mm package LED (noname); (7) 15-mm LED HLMP-8150 (Hewlett Packard).*

defined by

$$\lambda_c = \left(\int_{\lambda_1}^{\lambda_2} \lambda E_\lambda d\lambda \right) \left(\int_{\lambda_1}^{\lambda_2} E_\lambda d\lambda \right)^{-1} \tag{6.1}$$

with λ_1 and λ_2 denoting two wavelengths well below and above λ_c where E_λ has fallen to zero. It is important to note that the exact location of the centroid wavelength may be strongly affected by the very small values of the spectral distribution at the tails of the curve [5], when calculated from measured distributions. Another important quantity is the spectral bandwidth at half-intensity level, which is the difference between the two wavelengths on either side of λ_c, where the intensity has fallen to 50 % of the peak value.

Figure 6.10b shows spectral distributions for a selection of LEDs. They cover the entire range of the visible spectrum, extending into the near infrared region (IRLEDs). The spectral distribution depends on the material used in the semiconductor. Currently available light emitting diodes are made from III-V, II-VI, and IV semiconductors. The main materials used in the visible region are gallium arsenide phosphide $GaAs_{1-x}P_x$ (where the subscript x denotes the relative concentration of the constituents) and gallium phosphide. Gallium arsenide, another LED material, emits radiation around 900 nm, which lies in the near infrared and is not visible to the eye.

The efficiency of these materials is strongly dependent on the emitted wavelength and falls off drastically towards short wavelengths. For

Table 6.1: *Performance characteristics of different types of LEDs [1].*

Material	λ_p (nm)	color	K_s (lm W^{-1})
GaN	450	blue	-
GaP:N	565	green	0.6
$GaAs_{0.15}P_{0.85}$:N	590	yellow	0.36
$GaAs_{0.3}P_{0.7}$:N	630	orange	0.76
$GaAs_{0.6}P_{0.4}$	650	red	0.33
GaP:Zn,O	690	red	3.0
GaAs:Zn	900	infrared	-

photometric applications, this effect is less severe, as the luminous efficiency function of the human eye V_λ peaks at 555 nm and compensates for the decreasing efficiency above this wavelength. It is, however, extremely difficult to get LEDs at shorter wavelengths, such as blue light, because both the luminous efficiency of the human eye as well as the radiant efficiency decrease.

Table 6.1 summarizes the most important LED materials together with the peak wavelength, the apparent color, and the lighting system luminous efficacy K_s, if available.

Page dimensions and intensity distributions. Light-emitting diodes (LEDs) are available in a huge variety of package types and sizes. Figure 6.11 shows a selection of the most important packages. They range from bulky LEDs 15-mm in diameter, which resemble light bulbs more than LEDs, up to flat tiny surface mount (SMD) LEDs in the order 1 mm^2. It is important to note, when considering the use of LEDs, that the package size has nothing to do with the light output of the LED. This is due to the fact that the actual light emitting diode chip subtends only a small fraction of the surrounding plastic housing.

It can be shown that the LED chip is a very good approximation to a Lambertian surface. While GaAsP diode chips are nearly Lambertian, GaP are nearly isotropic. The actual intensity distribution depends strongly upon the shape and optical properties of the enclosing material. With a suitable design, the angular pattern can be changed from very broad to quite narrow. Some LED packages have cylindrical shape with a hemispheric top. These packages act as focusing lenses. If the LED chip is embedded in a depth corresponding to the focus of the lens, these devices produce a very narrow intensity beam. On the other hand, using diffusing materials yields a very broad distribution with good off-axis visibility but low luminance. In this context it is important to note

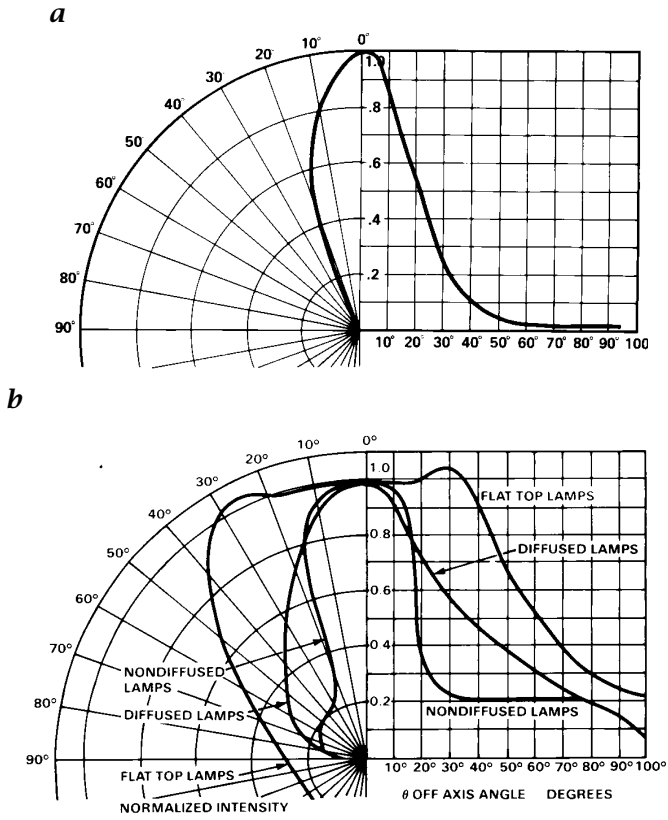


Figure 6.12: *Angular intensity distributions of two exemplary LEDs:* ***a*** *HLMP-K101, 'lens top';* ***b*** *HLMP-P105, 'flat top' (Courtesy Hewlett Packard Inc.).*

that most LEDs are optimized for photometric applications, such as instrument panel illuminations, rather than for illumination purpose.

Figure 6.12 shows two angular distributions for two different types of LEDs, one with a lens cover and one with a flat top. The intensity distributions differ significantly. There is a wide range of intensity distributions commercially available. Most manufacturers provide excellent data sheets, which show the averaged intensity distribution together with a variety of electrical and radiometric properties. Due to fluctuations in the manufacturing process, individual LEDs might show deviations from these averaged values, some in the order of 20 - 30 %. If the illumination setup requires narrow specifications, it is advisable to use optical precision LEDs, rather than bulk ware, designed for instrument panel signals.

An interesting example of a recently developed LED is shown in Fig. 6.13. This superbright miniature LED has the size of approximately

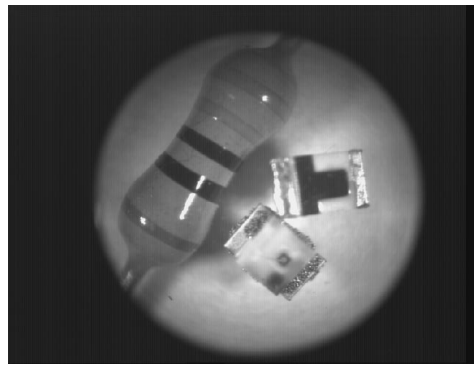


Figure 6.13: *Miniature LED HSMC-S690: size compared to an electronic resistor. Photo taken with an endoscopic optic.*

one square millimeter, but shows a luminous intensity of $I = 50\,\text{mcd}$ perpendicular to the surface with a total luminous flux of $\phi = 238\,\text{mlm}$, which is comparable to the brightest LEDs of larger package sizes. The intensity distribution is extremely flat and homogeneous. This LED is extremely useful for illumination purposes, as it can be integrated into arrays of high density with a corresponding high exitance and homogeneity.

There is no general agreement among LED manufacturers and users as to LED performance specifications, which leads to much confusion and misunderstanding. In manufacturer literature the most common quantity given to specify the directional output of an LED is luminous intensity. This term, however, is very often incorrectly used and the measured quantity is not the true intensity. In order to measure the intensity, the flux incident on a detector at a measured distance is used and the solid angle is computed by dividing the detector area by the squared distance. In real applications, the distance very often has to be chosen close to the emitting LED, which might not be large enough for the emitting area to behave like a point source. If the detector is too close, the LED acts as an extended source, which corrupts the angular distribution.

To avoid this problem and to pursue standardization in LED measurements, the international lighting commission, CIE, has defined a new term, called *averaged LED intensity* [5]. This term standardizes close range measurements by specifying the exact distance and size of the detector used for the measurement. The measurement geometries will be known as *CIE Standard Conditions* A and B. For averaged LED intensities measured under these conditions, the symbols $I_{LED\,A}$ and $I_{LED\,B}$ are recommended. Both conditions involve the use of a detector with a circular entrance aperture of $100\,\text{mm}^2$ (corresponding to a di-

ameter of 11.3 mm). The LED should be positioned facing the detector and aligned so that the mechanical axis of the LED passes through the center of the detector aperture. For conditions A and B, the distance d between the LED and the detector is 316 mm and 100 mm, respectively. This corresponds to solid angles of 0.001 sr for condition A and 0.01 sr for condition B. If the detector has been calibrated for illuminance E the averaged LED intensity can be calculated as

$$I_{LED} = \frac{E}{d^2} \tag{6.2}$$

Electrical properties. As the p-n junction of an LED becomes strongly conducting when operated in forward bias direction, LEDs always have to be operated with a protective resistance to avoid high currents, which will destroy the LED by thermal overheating. Currents are typically in the order of 20 to 50 mA, with a voltage drop across the LED of about 1 V. Thus, the power consumption of LEDs lies in the order of 20 to 50 mW.

As LEDs have very short response times, in the order of microseconds, they can be operated in pulsed mode with variable duty cycles. An important property of LEDs is the fact that they can be operated above the current limit for low duty cycle pulses. As the relationship between optical output and instantaneous forward current is linear over a wide region, very high intensity peak levels can be reached in pulsed mode. This technique, however, is not useful with GaP diodes, as they do not exhibit the linear relationship between current and luminous intensity, becoming saturated at moderate current levels. The maximum current depends on the duty cycle, as the average power consumption may not exceed the critical limit. For detailed information about the maximum current in dependence of the duty cycle, refer to data sheets provided by the manufacturer.

The pulsed-mode operation is especially useful for imaging applications. If LEDs are triggered on the frame sync of the camera signal, they can be pulsed with the frame rate of the camera. As the integration time of the camera only subtends a fraction of the time between two images, the LED output can be optimized by pulsed-mode operation. In order to operate the LED in pulsed mode, logical TTL-electronics can be used to generate an LED-pulse from the trigger signal of the camera. This signal can be used to switch the LED via transistors, as the TTL signal cannot be directly used for power switching of the LED. More detailed information about TTL electronics and interfaces driving optoelectronical components with TTL signals can be found in an excellent handbook on practical electronics by Horowitz and Hill [6].

LED arrays. The small package dimensions and the large variety of intensity distributions allow LEDs to be integrated into larger arrays

of arbitrary shape. Standard geometries include extended rectangular arrays, line arrays, and circular arrays, which can be used as ring illumination placed around the camera lens. In combination with additional optical components, virtually any intensity distribution can be achieved. For example, the use of diffusor plates creates very homogeneous extended illumination sources.

Combinations of different spectral LEDs can be used to produce color effects, depending upon the relative current distribution of the different LEDs. Most recently, tunable color LEDs have been commercially available. They combine three different LED chips with red, green, and blue output into one package. The light output consists of three spectral distributions, which are superimposed. The relative current input to the three diodes determines the color of the output.

6.3.6 Laser

Lasers are the most powerful monochromatic light source available. The word (acronym) LASER stands for light amplification by stimulated emission of radiation. The process of light generation is similar to that of other light emitting processes, where excited electron states recombine under emission of light. While the recombination of excited electrons usually happens randomly, the emission of light in lasers is stimulated by coherent radiation passing the laser material. Thus, the radiation is extremely coherent with a well-defined phase relation of all photons contributing to the light output. The final output is passed through an optical resonator, which allows only a very narrow spectral band of radiation to pass. Thus the radiation is essentially monochromatic with a very high spectral exitance.

While lasers usually have a very low radiant efficiency, in the order of 10 %, the radiation luminous efficacy might be quite high. A laser beam at a wavelength of 550 nm will have the maximum possible radiation luminous efficacy of $683\,\mathrm{lm\,W^{-1}}$. Lasers are available for a large variety of spectral ranges, from x-rays into the microwave region (MASER, microwave amplification by stimulated emission of radiation).

For illumination purposes in computer vision, the effect of coherence might cause problems. Due to the fixed phase relation, laser radiation is subject to interference, whenever it is scattered from objects with diameters in order of the wavelength. As almost any surface contains small-scale structures, or dust particles, surfaces illuminated with laser light show speckled structures, which move with the direction of observation. These speckles are randomly distributed points in space, where both constructive and destructive interference takes place. It is therefore hard to achieve a homogeneous illumination of a surface by laser light.

The output of a laser is usually confined to a very narrow, collimated beam of light. In order to get diffuse illumination, this beam has to be extremely diffused by optical components. On the other hand, the narrow laser beam makes it useful for applications where only a line pattern is needed. Using a scanning device or an optical component, such as a cylinder lens, light sheets can be created, which are commonly used for flow visualizations, such as particle imaging velocimetry or particle tracking. The thin light sheet illuminates only a 2-D subsurface of a 3-D volume, and allows optical slicing of the measurement volume of a partially transparent medium. Another application involves geometric measurements of object surfaces by the shape of a projected line, or other projected geometrical patterns, which can be conveniently created with laser beams.

6.4 Illumination setups

In Chapter 3 we showed how radiation can interact with surfaces and bulk properties of materials. The setup of illumination sources decides which radiometric/optical property of objects is encoded in the radiation received by the camera. It is a powerful tool to visualize object properties quantitatively and to optimize image quality.

Figure 6.14 shows several examples of different illumination setups, which will be further detailed in the remainder of this section. They are, however, only a small fraction of the almost unlimited possibilities to create problem-specific illumination setups that incorporate both radiometry and geometry of imaging.

6.4.1 Directional illumination

Directional illumination or *specular illumination* denotes a setup in which parallel light or light from a point light source is used to illuminate the object (Fig. 6.14a). This is the most simple type of illumination, as the setup basically consists of a single light source at a certain distance.

For *matte (Lambertian) surfaces*, directional illumination produces an irradiance, which depends on the angle of incidence of the light upon the surface. Thus, it can be used to determine the inclination of surfaces with respect to the illumination direction. At the edges of objects, directional illumination casts shadows, and does not illuminate occluded parts of objects. If the camera is observing the scene under a different angle, these shadows are visible in the image and might be confused with object borders.

For *specular reflecting surfaces*, directional illumination is not the appropriate illumination. These surfaces will appear black for all points

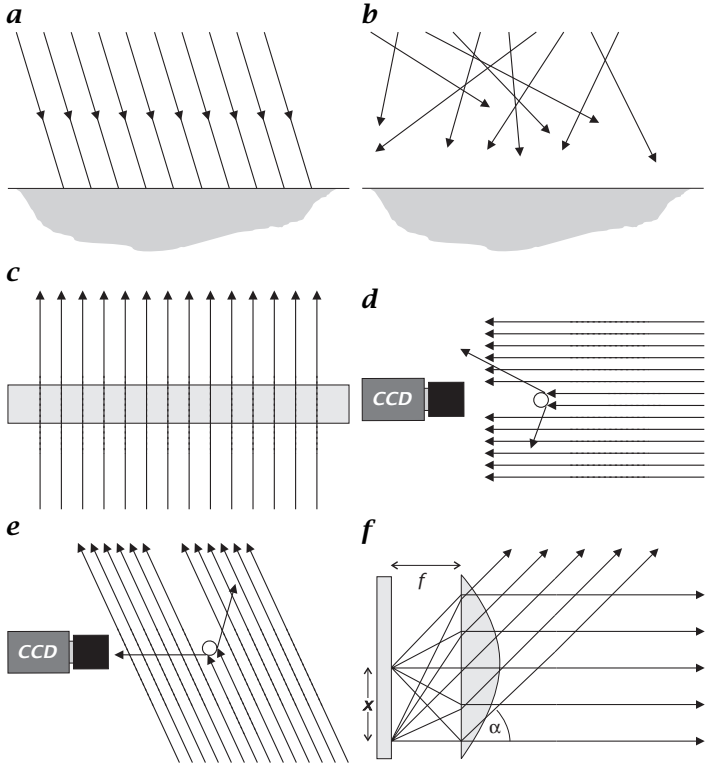


Figure 6.14: *Illustration of different illumination setups:* ***a*** *directed illumination;* ***b*** *diffuse illumination;* ***c*** *rear illumination;* ***d*** *light field illumination;* ***e*** *dark field illumination;* ***f*** *telecentric illumination.*

where the reflection condition is not met and show specular reflexes for the remaining points.

Most surfaces are mixtures of Lambertian surfaces with additional specular reflection. Thus, object surfaces show highlights that might be confused with surface structures or object edges. Furthermore, these regions might be overexposed and do not contain structural information. On the other hand, the position of specular highlights allows determination of the direction of the surface normal in these areas, as the exact reflection condition is fulfilled. This might be important information for 3-D reconstruction of the scene. Figure 6.15a shows an example of a plant leaf illuminated with directional illumination. The leaf shows highlights and a shadow is cast at the lower edge.

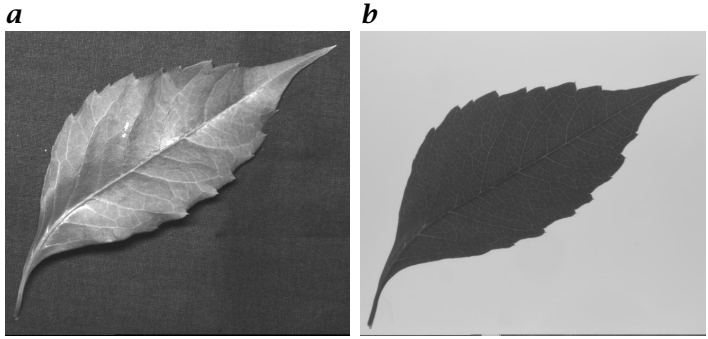


Figure 6.15: *Illustration of the difference between directed and rear illumination for a plant leaf.* ***a*** *Directed illumination.* ***b*** *Rear illumination.*

6.4.2 Diffuse illumination

A second type of front illumination is *diffuse illumination* (Fig. 6.14b). This illumination setup consists of an extended illumination source, which emits light under all directions. An optimal diffuse illumination creates an illuminance that is independent of the direction and impinges uniformly from the entire enclosing hemisphere. A good example of diffuse illumination is a completely overcast sky or heavy fog. Such an illumination is hard to realize in technical applications. Examples are extended diffusing plates or ring illuminations using LEDs or fiber optical illumination.

This type of illumination is well suited for both matte as well as specular surfaces. Although a diffuse illumination does not cast sharp shadows, thick edges of objects still partially block incident light. They appear as extended partially darkened regions, commonly referred to as *penumbra*.

6.4.3 Rear illumination

If only the geometrical outline of an opaque flat object is of interest, *rear illumination* is the common choice of illumination (Fig. 6.14c). Opaque objects appear as black objects without any structure. More interesting features can be obtained using rear illumination for semitransparent objects. For these types of objects, the transmitted radiation exhibits the entire spectrum of bulk-related interaction of radiation with matter, such as refraction, absorption, and scatter. Local inhomogeneities in the absorptivity show up as brightness patterns, integrated over the optical path of the radiation. Prominent examples of such images are x-ray images of medical applications. If the absorption is spectrally selective, the spectral content of the transmitted radiation carries additional information on the internal structure of objects.

Rear illumination can be set up with both directional as well as diffuse illumination. Figure 6.15b shows an example of a plant leaf illuminated by a diffuser screen behind the leaf. The background and the leaf show a well separated gray value distribution. The edge of the leaf is clearly visible. As the leaf is not totally opaque, it still shows fine structures, related to the more transparent water vessels.

6.4.4 Light and dark field illumination

Rear illumination can be considered to be a special case of *light field illumination*. Here a direct path exists from the light source to the camera, that is, the light source directly illuminates the sensor chip (Fig. 6.14d). As long as no object is present, the image appears bright. Any object in the light path diminishes the image irradiance by refraction, absorption, and scatter of light out of the illumination path. Thus, objects appear dark in front of a bright background. This type of illumination is commonly used to detect whether small objects (particles) are present in the volume between the illumination source and the camera (Volume 3, Section 29).

As opposed to light field illumination, *dark field illumination* inhibits a direct path between the light source and the camera (Fig. 6.14e). As long as no objects are present in the illumination path, the image appears dark. Objects in the illumination path become visible by scattering, reflecting, or refracting light into the camera. Thus, objects appear bright in front of a dark background. This type of illumination is as well used to detect small particles in the illumination path.

6.4.5 Telecentric illumination

Figure 6.14f illustrates the principal setup of a *telecentric illumination* system. It is used to convert the spatial radiance distribution of a light source into bundles of parallel rays that reflect the radiance (and spectral distribution) of a single point of the light source.

It principally consists of a large lens (often Fresnel lenses are used) which is placed at a distance of one focal length in front of an illumination source. A single point on the illumination source creates a bundle of parallel rays, leaving the lens into the direction of the line connecting the point and the center of the lens. The angle of the light bundle with the optical axis of the lens is given by the position on the focal plane using

$$\tan \alpha = \frac{x}{f} \tag{6.3}$$

where x is the distance between the intersection of the optical axis and the focal plane and f denotes the focal length of the lens. If the radiance of the light source is isotropic within the solid angle subtended by

the lens, the intensity emitted by the lens is constant over the lens aperture. For a nonisotropic radiance distribution (non-Lambertian source), the spatial distribution of the intensity of the emitted bundle of rays reflects the angular distribution of the radiance.

Thus, a telecentric illumination converts the spatial radiance distribution of an extended illumination source into an angular radiance distribution and the angular radiance distribution of a single point into a spatial distribution over the cross section of the bundle of rays. It is the basic part of various types of illumination systems.

6.4.6 Pulsed and modulated illumination

Pulsed illumination can be used for a variety of purposes, such as increasing the performance of the illumination system, reducing blurring effects, and measuring time constants and distances, to mention only a few of them.

Some illumination sources (e. g., special lasers) can only be fired for a short time with a certain repetition rate. Others, such as LEDs, have a much higher light output if operated in pulsed mode. As already outlined in Section 6.3.5, pulsed illumination has to be synchronized with the integration time of the video camera.

Instead of synchronizing the pulsed illumination with the camera integration both can be intentionally separated. Using a grated camera, with an adjustable delay after the illumination pulse, radiation is received only from a certain depth range, corresponding to the run time of the backscattered signal.

Pulsed illumination can also be used to image fast processes that are either blurred by the integration time of the camera or need to be imaged twice during the time between two consecutive frames. In the first case, a short pulse within the integration time restricts the accumulated irradiance to this time interval, independent from the integration time of the camera. The second case is commonly used in high-speed particle imaging velocimetry. Here the momentary distribution of the particle concentration in a liquid is imaged twice per frame by a fast double pulse. From the autocorrelation function of the image, the displacement of the particle pattern within the time between the two pulses can be computed.

Another important application of pulsed signals is time-of-flight measurements to estimate the distance of the scattering surface (see Section 18.5). Such measurements are demanding with electromagnetic waves, as the signal travels with the speed of light and time delays are in the order of nanoseconds. For acoustic waves, however, it is much easier to apply. These waves need about 3 ms to travel the distance of 1 m in air, as opposed to 3 ns for electromagnetic waves. Many liv-

ing species, such as bats and marine mammals, use acoustic signals to sense their 3-D environment in absolute darkness.

Instead of pulsing the illumination signal, it can also be *modulated* with a certain frequency. Examples can be found in scientific applications. Some processes that are visualized correspond with a certain time constant upon illumination with specific radiation. For example, active thermography uses infrared radiation to heat object surfaces and to observe temporal changes. Using a modulated thermal irradiance, the time constant of the processes related to the absorption and the internal transport of heat can be measured.

6.5 References

[1] Wolfe, W. L. and Zissis, G. J. (eds.), (1989). *The Infrared Handbook*, 3rd edition. Michigan: The Infrared Information Analysis (IRIA) Center, Environmental Research Institute of Michigan.

[2] Carlson, F. E. and Clarke, C. N., (1965). Light sources for optical devices. In *Applied Optics and Optical Engineering*, R. Kingslake, ed. New York: Academic Press.

[3] Oriel Corporation, (1994). *Light Sources, Monochromators & Spectrographs, Detectors & Detection Systems, Fiber Optics*, Vol. II. Stratford, CT: Oriel Corporation.

[4] McCluney, W. R., (1994). *Introduction to Radiometry and Photometry*. Boston: Artech House.

[5] CIE, (1997). Measurement of LEDs. CIE, Kegelgasse 27, A-1030 Vienna, Austria.

[6] Horowitz, P. and Hill, W., (1998). *The Art of Electronics*. New York: Cambridge University Press.

Part II

Imaging Sensors

7 Solid-State Image Sensing

Peter Seitz

Centre Suisse d'Électronique et de Microtechnique, Zürich, Switzerland

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

7.1 Introduction

As the name indicates, the field of *computer vision* has long been viewed as an essentially computational science, concerned only with the mathematical treatment of images whose origins are effectively ignored. This conventional view of computer vision (or machine vision), as perceived, for example, in the textbook by Gonzalez and Wintz [1], has slowly given way to a different, holistic comprehension of machine vision as the science of systems that extract information from wave fields (see also Chapter 1 or, for a practical implementation, Chapter 16). This systems approach, sometimes also called *electronic imaging* [2], has two immediate consequences: first, in a well-designed system, different components can compensate for the deficiencies in other components; practical examples of this capability include the digital correction of imaging lens distortions in photogrammetric applications (Chapter 17 or [3]), the significant increase of a system's dynamic range by nonlinear compression of the photosignal in the image sensor (Chapter 8 or [4]), and the digital compensation of offset and gain nonuniformities in the image sensor [5]. Second, the image acquisition process can become dynamic and adaptive, reacting to changes in the outside world by adapting the properties of the image capture and processing components in an optimal fashion. This powerful concept of *active vision* has already been proposed previously [6] but only now, with the recent development of custom solid-state image sensors, is it possible for active vision to reach its full potential, as described, for example, in Volume 3, Chapter 9. At the same time, new research opportunities are occurring in machine vision because new types of image processing algorithms are required that not only influence the image acquisition process but are also capable of exploiting novel imaging modalities [7].

This contribution should represent a comprehensive introduction to solid-state image sensing for machine vision and for optical microsystems, with an emphasis on custom image sensors that can be tailored to the requirements of individual imaging applications in research and industrial use.

The material presented here is organized in the form of a systematic exploration of the photosensing chain in Sections 7.2–7.5: Incident photons are followed on their paths into the interior of a semiconductor where most of the photons interact by producing electron-hole pairs. These photocharge pairs need to be separated in an electric field before they recombine again, leading to the flow of a photocurrent, which is proportional to the incident light intensity over many orders of magnitude (Section 7.2). The photocurrent can be manipulated and processed in many different ways before it is converted into a storable quantity at each pixel site. It is actually this large variety of processing capabilities that represents the true value of custom solid-state image sensing: by selecting and combining the required functionality for an imaging problem at hand, drawing from an extended "toolbox" of functional modules, the properties and the performance of an image sensor can be optimized for the given problem (Section 7.3). Finally, the preprocessed image information is stored at each pixel, often in the form of a voltage signal. During readout the individual pixels are interrogated either sequentially or several of them in parallel (Section 7.4). The stored pixel information is transmitted off-chip to the outside world, or additional processing steps (for example analog-to-digital conversion or even digital image processing) can be performed on the image sensor chip itself. An important part of the presented fundamentals of solid-state photosensing is the analysis of noise sources, noise reduction schemes, and the achievable signal-to-noise ratios (SNR) (Section 7.5). This leads us naturally to the basic reason for the development of modern *charge-coupled device* (CCD) technology and to the discussion of in which formats CCD image sensors might be replaced by CMOS-compatible image sensors in the near future.

Section 7.6 is devoted to an introduction of image sensor architectures. It covers the various types of CCDs employed today, the traditional photodiode array image sensor, and the active pixel sensor (APS) architecture. An external view of image sensors, as presented in Section 7.7, examines the different camera and video standards in use today. Although the conventional video standards as developed for TV applications such as CCIR, RS-170, PAL and NTSC still dominate today, new formats such as HDTV or nonstandard formats such as in some electronic still cameras are becoming more and more important.

The described image sensing developments, in terms of richness of functionality as well as the sharp decrease in price, have been possible only because of the amazing progress in semiconductor manufactur-

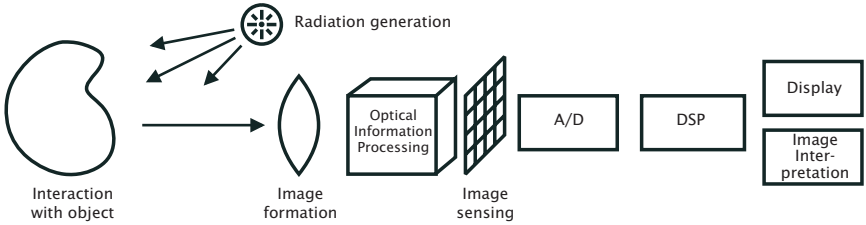


Figure 7.1: *Illustration of the photosensing ("electronic imaging") chain. It consists of a source of radiation, an interaction mechanism of the object under study with this radiation, shaping of the radiation field, conversion of radiation into electronic charge, the processing of this information, and the display for a human observer or the automatic extraction of pictorial information content.*

ing technology. A few aspects of this technology are presented in Section 7.8, wherein the aim is to gain insight into the consequences of the advances of semiconductor fabrication for solid-state image sensors. More predictions concerning the future of image sensing with regard to machine vision are offered in Section 7.10. Emphasis is placed on the custom functionality in hybrid systems, while in many practical applications the single-chip machine vision system does not make economical sense. As long as the fundamentals of the visual perception processes are not better understood, the realization of "seeing chips" will remain elusive.

Often ignored in the design of machine vision systems, the practical limitations of today's solid-state image sensors require special considerations for optimum system solutions. As described in Section 7.9, most of the shortcomings of the image sensors can be compensated by suitable calibration or correction procedures in an accompanying digital processor.

The concluding Section 7.11 reviews the most important aspects of custom image sensors, leading to the prediction that the large degree of freedom offered by the wide choice of image sensing functionality will result in many more applications where smart machine vision systems will be inexpensive, reliable, and yet provide high-performance solutions to optical measurement and visual inspection problems.

7.2 Fundamentals of solid-state photosensing

A generic machine vision or optical measurement system consists of the elements illustrated in Fig. 7.1. A suitable source of radiation, for example a light bulb, creates a wave field that can interact with the object under study. The part of the radiation that interacted with the object now carries information about it, which can be contained, for ex-

ample, in the spatial, temporal, spectral, or polarization modulation of the radiation. The returning information-carrying radiation is partially collected, often by making use of an imaging (lens) subsystem. A sensor converts the collected radiation into an electronic charge, which can be preprocessed using analog or digital electronics. The preprocessed information is converted into digital form for treatment in a specialized or general-purpose computer. The purpose of this image processing step is either to enhance certain aspects of the image information and display the modified image for inspection by a human observer, or to extract automatically certain types of pictorial content. This information can then be used to react to the perceived information content: for example, by interacting with the environment employing suitable actuators.

The work at present concentrates on the sensor and electronic preprocessing part of the whole electronic imaging chain using solid-state image sensors. The radiation that can be captured with these types of image sensors is restricted to electromagnetic waves extending from the x-ray region to the near infrared. This large spectral range covers most wavelength regions of practical importance, notably the visible spectrum.

Although any type of high-quality semiconductor can be employed for the conversion of electromagnetic radiation into photocharge and its electronic processing, the presentation in this work will be concerned mainly with *silicon*, due to its almost exclusive use in the semiconductor industry. As we will see, in most aspects this is not a real restriction, and the use of silicon for photoconversion and electronic processing is really an excellent choice.

In the following, a systematic exploration of the photosensing chain is presented ("from photons to bits"), as illustrated in Fig. 7.2. Incident photons are converted into charge pairs, leading finally to preprocessed image information at the output of the semiconductor chip.

7.2.1 Propagation of photons in the image sensor

Two types of interactions of photons with solid-state materials have to be considered for an understanding of an image sensor's properties: absorption and reflection (see also Sections 3.3 and 3.4). Before an incident photon can interact measurably in the bulk of a piece of semiconductor, it has to arrive there safely, crossing the interface between air and semiconductor surface. What can happen to an incident photon is illustrated schematically in Fig. 7.3, depicting the cross section through an image sensor. On top of the image sensor, we find scratch-resistant transparent covering and protective materials, often in the form of dielectric layers such as silicon dioxide or silicon nitride, with a typical thickness of a few μm. At the interface between cover and

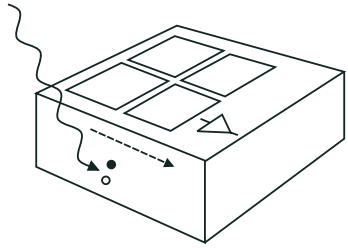


Figure 7.2: *Simplified sequence of events in semiconductor photodetection. Incoming radiation is converted into charge pairs in the bulk of the semiconductor, the charge pairs are separated in an electric field, and they are either stored in the pixel or the photocurrent is processed locally. The photosignal is subsequently transported to an electronic amplification circuit for detection.*

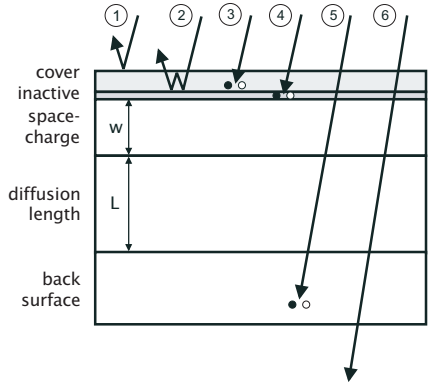


Figure 7.3: *Schematic representation of the optical losses encountered in semiconductor photosensors: (1) surface reflection, (2) thin-film interference, (3) absorption in the cover, (4) photocharge loss in inactive regions, (5) interaction deep in the semiconductor bulk, and (6) transmission through the semiconductor.*

actual semiconductor, there is a thin, essentially inactive zone. In the bulk of the semiconductor one encounters first a region that has been swept clean of mobile electronic charges. In this so-called space-charge region, usually a few microns deep, an electric field is present. Below this, the field-free bulk of the semiconductor follows, which can be as thin as a few μm or as thick as many 100 μm. The following identifies six different effects that prevent photons from being detected by the image sensor:

1. Due to the mismatch between the refractive index of top surface and ambient (often air), the incident photon is reflected and does not enter the image sensor. A typical value for this *reflection loss*

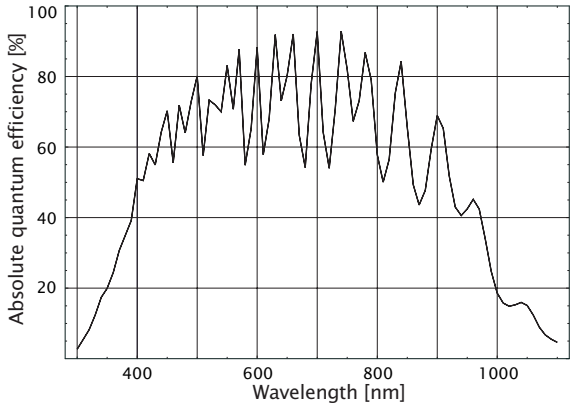


Figure 7.4: *Absolute quantum efficiency measurement of a silicon p-n junction realized with a standard CMOS process. This example illustrates the decay towards the blue (surface absorption) and red spectral region (interaction too deep in the semiconductor), as well as the oscillations due to thin-film interference.*

is obtained in the following way: using an index of refraction of n=1.5 for silicon dioxide, 4 % of the photons are reflected at normal incidence from air [8].

2. Multiple reflections in the covering thin layer lead to a strong spectral oscillation of the transmittance, as is apparent in the measurement shown in Fig. 7.4.
   Depending on the wavelength of the incident photon it is either transmitted well or it is preferentially reflected back. In good image sensors, this disturbing effect is virtually eliminated by the deposition of additional dielectric antireflection layers on top of the image sensor [8].
3. The covering layers are not perfectly transparent, leading to *absorption* of part of the incident photons already at this stage. The reduced blue response of CCD image sensors is a good example of this effect, caused by the low transmission of the covering polysilicon electrodes on the pixels.
4. Inactive regions near the surface of the semiconductor consist of semiconductor material with a very short lifetime of charge pairs. This is either caused by defects right at the interface (less than 1 nm), or by very high doping concentration near contacts [9]. Photogenerated charge pairs recombine so fast that their collection and electronic detection is improbable.
5. Photons that are absorbed very deeply in the bulk of the semiconductor result in photocharge that does not have a chance to reach the surface of the image sensor for collection in a pixel. As will be

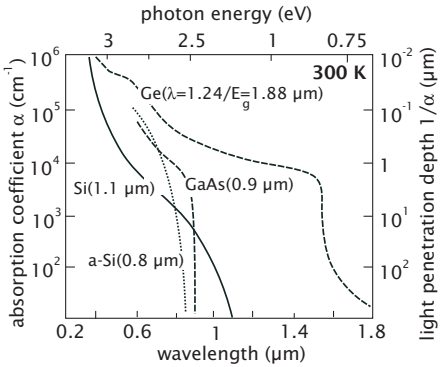


Figure 7.5: *Optical absorption coefficient and light penetration depth as a function of wavelength for various semiconductor materials. Data taken from Sze [10].*

described in what follows, the critical distance is the so-called diffusion length L, which can be many times 10 μm deep for low-doped semiconductors [9].

6. Finally, photons might travel through the image sensor without interaction, leaving it again at the back end.

7.2.2 Generation of photocharge pairs

Because of the sequential process of *photocharge generation*, virtually all photons that are absorbed in the semiconductor material are converted into an electronic charge [8]. There is a strong spectral dependence, however, of the mean absorption depth at which this photoconversion takes place, as illustrated in Fig. 7.5. Short-wavelength light is predominantly absorbed at the surface, while red light penetrates deeply into the bulk of the semiconductor. A major consequence of this effect is that the achievable spatial resolution degrades significantly with wavelength [11]: images taken in the red or infrared spectral region show much less contrast compared to images taken in green or blue light. For this reason, image sensors are often covered with an optical filter, cutting off the infrared portion of the incident light.

In the absorption process, a photon loses its energy by creating one or more charge pairs. In a photodetection event, no net charge is created and neutrality is always maintained. For this reason, charge pairs are created, consisting of an electron and a (positively charged) quasiparticle called hole [8]. The overall charge conversion efficiency of this process is usually measured with the *quantum efficiency* η, describing how many charge pairs are created and electronically detected per incident photon. Alternatively, this conversion efficiency can be described

with the *responsivity* R in units A/W, measuring how much current is flowing out of a photosensor per incident light power. The relationship between R and η is given by

$$R = \eta \frac{\lambda q}{hc} \tag{7.1}$$

Using Planck's constant h, the speed of light c, the unit charge q, and the photons' wavelength λ. As an example, consider a photodetector with an η of 0.9, illuminated with red light (λ =633 nm) from a HeNe laser. The corresponding responsivity is R =0.46 A/W.

In the visible and infrared portion of the spectrum, η is less than unity. This is illustrated in Fig. 7.4 with the actual measurement of an n^- p^- photodiode, manufactured with a standard CMOS process using silicon. The η decreases towards both the blue (incident light is already absorbed in the covering layers) and the infrared portion of the spectrum (light penetrates and interacts so deeply in the semiconductor that the created charge pairs recombine and disappear before they reach the surface where they could have been collected and measured). In the visible part of the spectrum, a rather high η of close to 100 % is observed. As no special antireflection coating is used in this photodiode, spectral oscillations can be seen in the η curve, caused by multiple reflections of the incident light within the covering layers [8], so-called thin-film interference. For improved performance, antireflection coatings are employed, reducing this effect significantly.

If a photon has a sufficiently high energy such as in x-rays, one photon can create many charge pairs. In silicon a mean energy of 3.8 eV is required for the creation of one electron-hole pair [12]. As an example, consider a soft x-ray photon with an energy of 1000 eV, corresponding to a wavelength of 1.24 nm. The absorption of this x-ray photon results in the creation of 263 charge pairs. Because silicon starts to become transparent for x-ray photons with an energy of more than a few 1000 eV, silicon is not an efficient solid state detector for such energies. Other semiconductors, consisting of high-density materials with atoms of high atomic numbers, are more appropriate for x-ray detection [13].

7.2.3 Separation of photogenerated charge pairs: photocurrents

Once a charge (electron-hole) pair has been created, it must be separated within a certain time before it recombines again and loses all information about the previous presence of the photon that generated the charge pair. This recombination lifetime τ depends critically on the quality and purity of the semiconductor [9]. In high-quality low-doped silicon used in CMOS processes, for example, the lifetime can be as large as several tens of microseconds. This is the time available for

separating the photocharge and moving the different charge types to suitable storage areas.

Two physical effects dominate the motion of electronic charge in semiconductors: drift in an electric field and diffusion caused by the random thermal motion of the charge carriers. The presence of an electric field E causes charge carriers to move with the velocity v

$$v = \mu E \tag{7.2}$$

with the *mobility* μ. As an example, the mobility of electrons in low-doped silicon at room temperature is about 1350 cm^2/Vs. Above a certain field strength, the velocity saturates, taking on a constant value v_{sat}. For silicon, this saturation velocity is about 10^5 m/s [10].

Even in the absence of an electric field, charge can move: the thermal random motion causes diffusion, a tendency of charge carriers to equilibrate their distribution. The thermally induced velocity v_{diff} of the charge carriers can be very high: an electron at room temperature has an average velocity of $v_{\text{diff}} = 10^5$ m/s. This random motion causes an average [root-mean-square (rms)] displacement L of a single electron, depending on the time t given for the diffusion process

$$L = \sqrt{Dt} \tag{7.3}$$

with the diffusion constant D. Silicon exhibits a typical electron diffusion constant of about $45\,cm^2/s$ at room temperature. For the recombination lifetime τ already mentioned, the corresponding average displacement L is called *diffusion length*. This is the average distance over which a charge carrier can move without the influence of an electric field and without recombining. As an example, consider $\tau = 10\,\mu s$ and D =45 cm^2/s, resulting in L =212 μm. This implies that the diffusion process can be extremely important for the collection of charge carriers over significant distances. This also means that charge carriers photogenerated deeply in the semiconductor have a high chance of reaching the surface, where they can be collected and where they contribute to a severe reduction of the contrast, especially for small pixel periods. As mentioned in the preceding, this can be counteracted only by filtering out the long-wavelength photons that would penetrate deeply into the semiconductor.

Photogenerated charge carriers moving under the influence of an electric field represent a current, the so-called *photocurrent*. This photocurrent is proportional to the incident light intensity over 10 orders of magnitude and more [14]. It is this strict linearity of photocurrent with incident light over a wide dynamic range that makes semiconductor photosensors so attractive for many applications in image sensors and optical measurement systems.

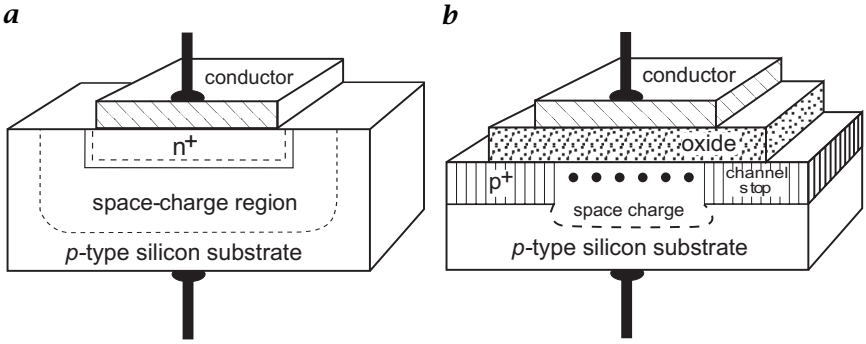


Figure 7.6: *Cross sections through the two major types of electrical field generating and charge storing devices in semiconductors:* ***a*** *photodiode, consisting of a reverse-biased p-n junction;* ***b*** *MOS capacitance, consisting of a (transparent) electrode on the semiconductor material, separated by a dielectric insulation.*

7.3 Photocurrent processing

All the information a photosensor can extract from the light distribution in a scene is contained in the spatial and temporal modulation of the photocurrent in the individual pixels. For this reason, it is of much interest to process the pixels' photocurrents accordingly, in order to obtain the relevant modulation parameters in the most efficient manner [7]. Traditionally, only the integrated photocurrent could be extracted; today a large variety of photocurrent preprocessing is available, making it possible to optimize the photosensor acquisition parameters to a given problem. In the following, a few examples of such photocurrent preprocessing are presented.

7.3.1 Photocharge integration in photodiodes and charge-coupled devices

The simplest type of photocurrent processing is the integration of the photocurrent during a certain time, the exposure time. In this way an integrated charge is obtained that is proportional to the number of photons incident on the pixel's sensitive area during the exposure time. This functionality is very easy to implement by employing the capacitance of the device used for generating the electric field for photocharge separation. Figure 7.6 illustrates this principle for the two most important photosensitive structures, the *photodiode* (PD) and the *metal-oxide-semiconductor* (MOS) capacitor as used in the *charge-coupled device* (CCD) image sensors. Both devices are easily fabricated with standard semiconductor processes.

A photodiode consists of a combination of two different conductivity types of semiconductor, as illustrated in Fig. 7.6a. In the junction between the two types of semiconductor, an electric field in the so-called space-charge region exists, as required for the separation of photogenerated charge carriers. At the same time, this space-charge region has a certain capacitance, varying with the inverse of the space-charge region width. Photodiodes are typically operated by biasing ("resetting") them to a certain potential and exposing them to light. Photocharge pairs entering the space-charge region are separated in the PD's electric field, a photocurrent is produced, and the photocharge is accumulated on the PD's capacitance, lowering the voltage across it. After the exposure time, the residual voltage is measured, and the voltage difference compared with the reset voltage level is a measure for the amount of light incident on the pixel during the exposure time.

The MOS-capacitance illustrated in Fig. 7.6b consists of a thin layer of oxide on top of a piece of semiconductor. The oxide is covered with a conductive material, often a metal or highly doped polycrystalline silicon (polysilicon). As in the case of the PD, the MOS structure is biased to a suitable voltage, leading to a space-charge region of a certain extent in the semiconductor. Again, photocharge is separated in the electric field and it is integrated on the MOS capacitance, collected at the interface between semiconductor and oxide.

A typical value for the PD and MOS area capacitance is 0.1 fF/μm^2. Assuming a maximum voltage swing of a few volts, this implies a storage capacity of a few thousand photoelectrons per μm^2. Once this storage capacity is exceeded, additional photocharge in the corresponding pixel starts to spill over to neighboring pixels. This effect is called *blooming*, and well-designed image sensors provide special collecting ("*antiblooming*") structures for a reduction of this effect [15].

7.3.2 Programmable offset subtraction

Several machine vision and optical metrology problems suffer from small spatial contrast [7]. In such cases in which the spatial signal modulation is small compared to the background light level, one would profit from an *offset subtraction* mechanism in each pixel. This can be realized, even programmable in each pixel, with the offset subtraction mechanism proposed by Vietze and Seitz [16]. Each pixel contains a photodiode in series with a programmable current source, as illustrated in Fig. 7.7. This current source is easily realized with a MOSFET, whose gate voltage can be preset to a certain voltage level with a second MOSFET, and by using a capacitance for the storage of this gate voltage. The MOSFET is operated in the so-called weak-inversion regime, where the drain current depends exponentially on the gate voltage; the current typically doubles with each increase of gate voltage by about 30 mV. In

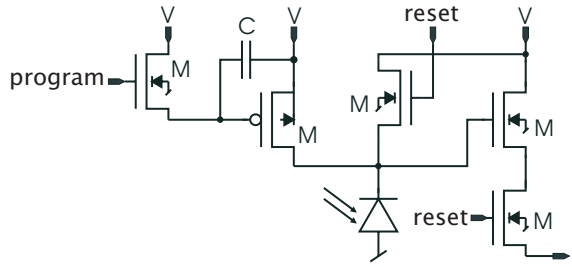


Figure 7.7: *Schematic diagram of the offset pixel with current source transistor M_{cur}, reset transistor M_r, row-select transistor M_{sel}, and sense transistor M_s. The value of the offset current is stored on the switched offset memory capacitor C_M with the programming switch M_p [16].*

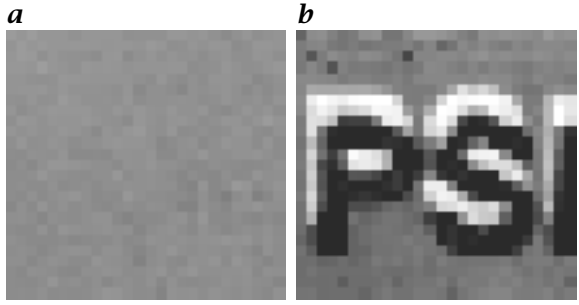


Figure 7.8: *Application example of the offset pixel—motion detector realized with a 26× 28 pixel CMOS image sensor [17]:* ***a*** *sensor image of a simple scene (black letters "PSI" on white paper) after adjusting the pixels' individual offset current to a medium gray level;* ***b*** *sensor image after moving the scene slightly downwards and to the right. Pixels with changed values appear either black or white.*

this way, the offset current can be varied easily between 1 fA up to several tens of μA [17]. The same integration mechanism as presented in Section 7.3.2 is employed for the collection of signal photocharge, representing the difference between total photocharge minus offset photocharge. Using this method, a dynamic range exceeding 150 dB can be reached, and several interesting applications can be realized very easily. An example of this is a simple *change detector*, implemented as a two-stage process. In a first stage, the offset current in each pixel is programmed such that the net result is zero; the offset currents effectively cancel the local photocurrents. In a second stage, the image is simply observed for non-zero pixels, indicating that there was a change in the present scene compared to the original "reference" scene: a change in the scene has occurred!

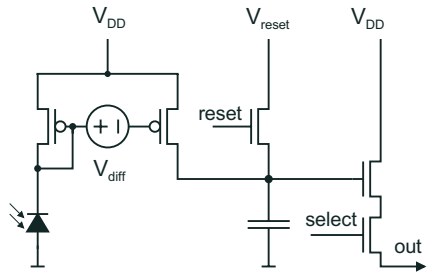


Figure 7.9: *Schematic diagram of the gain pixel, consisting of a modified current mirror [17], with which a photocurrent multiplication with a factor ranging between 10^{-4} up to more than 10^{4} can be realized.*

The realization of such a change detector is illustrated with an experimental offset pixel image sensor with 28×26 pictures, fabricated with standard CMOS technology [17]. In Fig. 7.8a the result of offset cancellation for a stationary scene containing the letters PSI is shown: a uniform gray picture. Once the object is moved (the letters are shifted downwards), the resulting pixels appear as bright where the dark object was, or as dark where the bright background was, see Fig. 7.8b.

7.3.3 Programmable gain pixels

Another local operation desirable in an image sensor is the individual multiplication of the photocurrent with a programmable factor. This can be achieved with a modification of a simple electronic circuit called *current mirror*, consisting of two transistors. In the standard configuration, the gate terminals of the two transistors are connected. In the modification proposed in Vietze [17], a voltage difference between the two gates is applied, as illustrated in Fig. 7.9. This voltage difference is either fixed (e.g., by semiconductor process parameters), or it can be implemented as individually programmable potential differences across a storage capacitor. The photocurrent produced by a photodiode in the first branch of the modified current mirror results in current in the second branch that is given by the photocurrent times a factor. By using a similar physical mechanism as in the offset pixel, the gain pixel shows a current doubling (or halving) for each increase (decrease) of the voltage difference by about 30 mV. In this way, current multiplication (division) by several orders of magnitude can easily be obtained. As before, the multiplied photocurrent is integrated on a storage capacitor and read out using conventional circuitry.

An application of this is a high-sensitivity image sensor as described in [17], in which each pixel has a fixed gain of about 8500. In this way, a sensitivity (see Section 7.5.1 for the definition) of 43 mV per photo-

electron has been obtained, and an input-referred rms charge noise of better than 0.1 electrons at room temperature. As will be discussed in Section 7.5, this impressive performance must come at a price. In this case it is the reduced bandwidth of the pixel, reflected in the low-pass filter characteristics at low photocurrents with response times of several milliseconds.

7.3.4 Avalanche photocurrent multiplication

The multiplication mechanism described in the foregoing is based strictly on the use of electronic circuitry to achieve gain. In semiconductors there is a physical mechanism that can be exploited to multiply charge carriers before they are detected. This effect is called avalanche multiplication, and it is used in so-called *avalanche photodiodes* (APDs) [18]. If the electric field is increased to a few times 10^5 V/cm, charge carriers are multiplied with a strongly field-dependent factor. Depending on the specific doping conditions in the semiconductor, the necessary electric fields correspond to breakdown voltages between a few volts and a few hundred volts. The strong dependency of the multiplication factor on voltage is illustrated with a model calculation for a breakdown voltage of 40 V, shown in Fig. 7.10 [19].

The APDs are commercially available and, because of the high achievable gains, they are even suitable for single-photon light detection [20]. Due to the unusual voltages, the complex voltage stabilization/homogenization circuits and the nontrivial readout electronics in each pixel, most APDs are only of the single-pixel type. The development of APD line and image sensor arrays has only just started. Nevertheless, the fabrication of reliable APD image sensors with CMOS processes is an active topic of research, and promising results have already been obtained (see, for example, Mathewson [21].

7.3.5 Nonlinear photocurrent to signal voltage conversion

Image processing algorithms are often motivated by solutions found in biological vision systems. The same is true for different types of photodetection strategies, especially for the realization of image sensors offering a similarly large dynamic range already inherent in animal vision. The fact that the human eye shows a nonlinear, close to *logarithmic sensitivity* has been exploited, for example, in the artificial retina described in Mahowald [22]. The realization of CMOS pixels offering a logarithmic sensitivity is particularly easy to achieve: one can use the logarithmic relationship between gate voltage and drain current in a MOSFET operated in weak inversion, already described in Section 7.3.2. The resulting pixel architecture, shown in Fig. 7.11 and exploited in Chapter 8, is particularly easy to implement in a CMOS

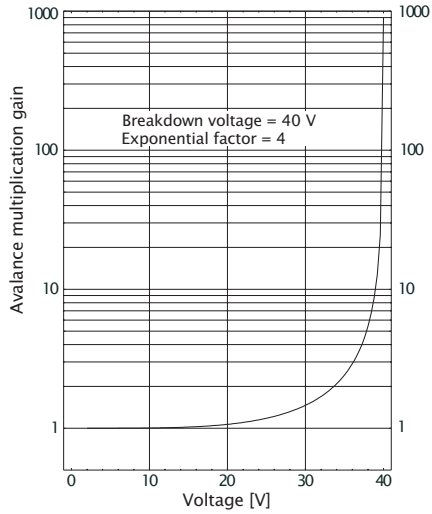


Figure 7.10: *Empirical relationship between applied voltage and obtained current gain in an avalanche photodiode, for which a breakdown voltage of* $V_B = 40$ *V and an exponent of* $n = 4$ *have been assumed.*

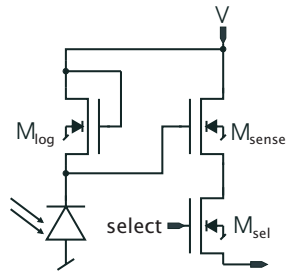


Figure 7.11: *Schematic diagram of a pixel with logarithmic response, consisting of just one photodiode and three MOSFETs. Implemented with a standard CMOS process, such a pixel shows an output voltage increase of about 40 mV per decade of incident light power.*

process because a pixel consists of just a photodiode and three MOS transistors [23]. A typical photoresponse of about 40 mV per decade of optical input power is obtained with such logarithmic pixels, and their useful dynamic range exceeds 120 dB. Practical examples of scenes requiring such a high dynamic range are illustrated in Fig. 7.12, with the actual measurements taken with a logarithmic image sensor exhibiting 64×64 pixels. In the image of a light bulb, the glowing filament as well as the background are clearly visible at the same time. Back-illuminated scenes, such as a portrait in front of a window, are dreaded by photog-

Figure 7.12: *Four examples of pictures taken with a small-area logarithmic image sensor with 64 × 64 pixels, illustrating the large dynamic range of ≥ 120 dB:* ***a*** *electric light bulb where the glowing filament and the background are visible simultaneously;* ***b*** *back-illuminated scene of a portrait in front of a window;* ***c*** *parking garage application with its notoriously high dynamic range (headlights compared to dark corners) and low average light levels;* ***d*** *welding application in which the object and the welding arc can be observed at the same time without blooming.*

raphers, but they are easily handled by logarithmic pixels. In a parking garage, it is difficult to image dark corners and the interior of cars without being blinded by car headlights. Welding applications profit from the simultaneous imaging of the welding arc and its environment.

In contrast to other pixel types in which photocharge is integrated as discussed in Section 7.3.1, the logarithmic pixel measures the *voltage* at the drain of the MOSFET in series with the photodiode. For this reason, the dynamic behavior of such a logarithmic pixel depends on the photocurrent: the darker a scene (the lower a diode's photocurrent), the longer it takes until this MOSFET is in equilibrium again. Therefore, logarithmic pixels react much more slowly at low than at high illumination levels.

Besides their high dynamic range, logarithmic pixels have a property that should make them extremely interesting for image processing applications: an object with a given local contrast, which is imaged with a logarithmic sensor, results in an image with local pixel differences that

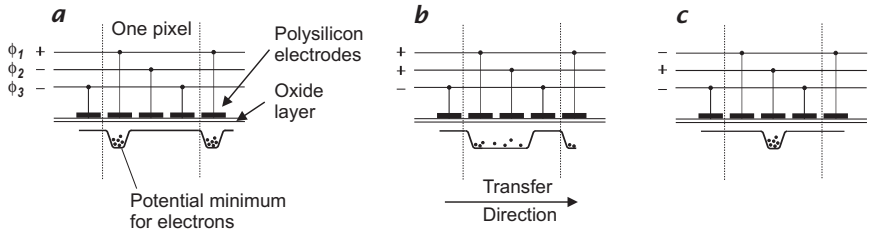


Figure 7.13: *Illustration of the charge transport principle in CCDs. Different stages of the electrode clocking and charge shifting sequence are shown in* ***a****,* ***b*** *and* ***c****.*

are independent of the scene illumination level. This property is easily explained with the observation that a (local) light intensity ratio I_1/I_2 results in a signal given by $\log(I_1) - \log(I_2)$, and a proportional intensity change of $c \times I$ results in a signal given by $\log(c) + \log(I)$. The same object under brighter illumination looks the same in the logarithmic image, except for an additive shift of the background level.

7.4 Transportation of photosignals

The different types of image sensors described in the preceding produce an electrical quantity as a measure for a certain property of the incident light. The electrical quantity can be an amount of charge (e. g., the integrated photocharge), a current (e. g., the photocurrent) or a voltage level (e. g., the voltage difference of a discharged photodiode). This signal has to be transported as efficiently as possible to an output amplifier, responsible for making this signal available to the off-chip electronics.

7.4.1 Charge-coupled device photocharge transportation

In the case of CCDs, the photocharge is stored under a precharged MOS capacitance. The basic CCD idea is to combine a linear array of such MOS capacitances, so that a stored photocharge can be moved laterally under the influence of appropriate MOS electrode voltage patterns. This principle is illustrated in Fig. 7.13, showing a *surface-channel CCD* (S-CCD). In the semiconductor, photocharge pairs are created under the influence of light. Moving by diffusion and by drift, the photoelectrons can find their way to positively biased MOS electrodes, also called gates, where they are stored at the interface between semiconductor and thin oxide. The photogenerated holes are repelled by the positive gate voltage, and they move around by diffusion until they finally combine in the silicon substrate.

It is important to note that a CCD pixel is not represented only by the positively biased gate because this electrode can receive diffusing and drifting photoelectrons from its environment. A pixel's geometry is therefore rather defined in terms of "effective photocharge collection area," extending about halfway to the next positively biased electrode. This also shows that a pixel does not have sharply defined edges; the extent of the charge collection area representing a pixel depends on the wavelength, the electric field distribution, and the diffusion properties of the semiconductor. Generally, longer wavelength light results in a lower contrast and offers reduced resolution, as discussed in Section 7.2.2.

In Fig. 7.13, the potential distribution under the electrodes right at the surface is indicated. Photocharge accumulates in the shown "potential wells." By changing the gate voltage patterns, the potential wells can be widened, leading to a broadened distribution of photoelectrons. Using a suitable gate voltage pattern, one can also reduce the extent of the potential wells, and photoelectrons move again to regions with the lowest potential. As illustrated in Fig. 7.13, it is physically possible to transport photocharge. This transport mechanisms works rather well up to frequencies of a few MHz. In good S-CCDs, only about 0.01 % of the photocharge is lost on average in transporting a photoelectron packet from one gate to another, neighboring gate. Instead of this charge transport loss, one often uses the *charge transfer efficiency* (CTE) concept, defined as the complement to 100 %. The CTE amounts to 99.99 % in the case of a good S-CCD.

In long CCD lines, a CTE of 99.99 % is still not good enough. Charge is trapped at the surface, making it hard to improve the CTE. For this reason, another type of CCD has been invented, the *buried-channel CCD* (B-CCD), in which the transport takes place in the bulk of the semiconductor, a few 100 nm away from the surface. Those CTEs of up to 99.99995 % can be obtained in B-CCDs, and all commercially available CCD line and image sensors are of this type.

Above a limiting clock frequency a CCD's CTE starts to degrade rapidly. Nevertheless, CCDs have been operated successfully at very high clock frequencies. For silicon, 1 GHz has been achieved [24], while GaAs CCDs have reached 18 GHz clocking frequency [25]. Such high clock rates require special precautions in the CCD fabrication process, usually not available for standard video sensors. Today's technology limits the analog bandwidth of CCDs to about 40 MHz. This is sufficient for standard video imagers according to the European CCIR or the American RS-170 black-and-white video standard. For HDTV sensors, however, the required pixel rate is around 75 MHz, making it necessary to operate two outputs in parallel in HDTV CCD imagers.

7.4.2 Photodiode photocharge signal transmission

The CCD technology provides a clean separation of the acquisition of photocharge and its electronic detection. This is achieved by transporting the photocharge with the almost perfect CCD transportation principle. Traditional photodiode arrays operate differently, by supplying each photodiode (PD) with its individual switch (see also Fig. 7.17 and Section 7.6.4), and by connecting many switches to a common signal ("video") line. This video line is most often realized using a well-conducting metal strip, leading to a common output amplifier structure. In a PD array, the image acquisition process proceeds in the following way: assume that all PDs are initially precharged to a certain reverse bias, typically a few volts and that all switches are closed. Incident light generates photocharge pairs in each pixel, leading to the flow of a photocurrent due to the separation of photocharge pairs in the electrical field region of the PDs. As a PD also represents a capacitance, this capacitance is discharged by the photocurrent. After a certain time (the exposure time), a pixel can be interrogated by connecting the PD via the appropriate switch to the video line. The output amplifier resets the photodiode to its initial voltage value through the conducting line, while measuring how much charge is necessary to do so. This charge is (apart from noise effects) the same as the accumulated photocharge in this pixel. This means that—in contrast to CCDs where the actual photocharge is transmitted and detected—a PD array works by charge equilibration in a usually long conducting line. As we will see in Section 7.5.2, this charge equilibration process introduces noise in the signal detection process, which is proportional to the video line's total capacitance: the larger the number of pixels, the larger the video line capacitance and the larger the image noise. It is this physical effect that made PD image sensors so unattractive compared to CCDs in the early 1980s and which led to their almost complete replacement by CCD image sensors.

7.4.3 Voltage signal transmission

Not all pixel types depend on the transmission of charge signals, as indicated by several examples of pixel functionality discussed in Section 7.3. Voltage signals are sometimes generated in the individual pixels and these voltage signals must be transmitted to an output amplifier structure. A similar architecture as described in the preceding is used for this, consisting of individual switches in each pixel that connect the local voltages to a common amplifier structure. In such an architecture the voltage signal transmission task is much easier to accomplish than the charge signal transmission just discussed here: Johnson noise in the conducting video line, filtered with the video line's RC low-pass fil-

ter characteristics results in voltage noise that is proportional to one over the square root of the video line's capacitance [26]. The larger this capacitance, the lower the voltage noise. For this reason, voltage signals can be transmitted with much less noise and higher measurement precision than (small) charge signals. This implies that image sensor types offering voltage transmission architectures, such as that provided by the logarithmic pixel described in Section 7.3.5, have an inherent noise advantage over conventional PD architectures. This will be discussed in more detail in Section 7.3.3.

7.5 Electronic signal detection

The basic task of electronic signal detection is the precise measurement of voltage signals offering low noise levels and a wide dynamic range. These input voltage signals have either been produced by the conversion of photocharge into a voltage, for example by employing a capacitance, or they are the result of more elaborate photocharge preprocessing as was already described here. The output of the signal detection electronics is usually a voltage that should be proportional to the input voltage over a large dynamic range. An important property of the signal detection electronics is that its output should have very low impedance, that is, the output voltage should be stable and must not depend on the amount of current drawn. As we will see in what follows, the electronic signal detection noise is today's limiting factor in increasing an image sensor's sensitivity and its dynamic range.

7.5.1 Signal-to-noise and dynamic range

For a numerical description of the voltage or charge-noise performance of an electronic circuit, two values are often used, the *signal-to-noise ratio* SNR and the *dynamic range* DR. The SNR is defined by comparing an actual signal level V with its rms noise ΔV, according to:

$$\mathrm{SNR} = 20\,^{10}\log \frac{V}{\Delta V} \tag{7.4}$$

The DR compares the maximum signal level $\Delta v_{\max}$ with the minimum rms noise level ($\Delta V_{\min}$, in an image sensor typically obtained in the dark

$$\mathrm{DR} = 20\,^{10}\log \frac{V_{\max}}{\Delta V_{\min}} \tag{7.5}$$

As an example, consider a CCD image sensor whose maximum charge ("*full well charge*") is 50,000 electrons, and for which a dark noise of

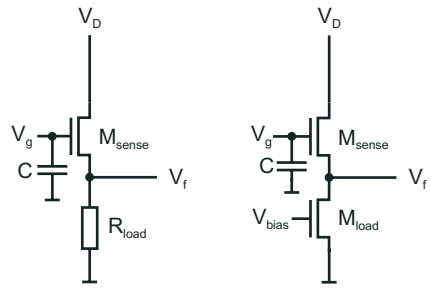


Figure 7.14: *Schematic diagram of the source follower circuit realized with a resistor (left) or with a so-called active load MOSFET (right). This is the most often used electronic circuit for photocharge detection in semiconductor image sensors. Photocharge deposited on the gate capacitance leads to a gate voltage V_g, which in turn produces a linear change in output voltage V_f.*

50 electrons rms is observed. This image sensor has a dynamic range of 60 dB.

It should be mentioned that the preceding definitions of SNR and DR in image sensors are not consistent with usage elsewhere in optical physics: as the measured voltage at the image sensor's output is usually proportional to the incident optical power, a factor of 10 in front of the logarithm should be used instead of the employed factor 20. However, because electrical engineers are used to associate power only with the square of voltage levels, the definitions given here are the ones employed almost exclusively for all image sensor specifications.

7.5.2 The basic MOSFET source follower

Although elaborate circuits exist for the desired conversion of voltage signals into other voltage signals, most image sensors employ the simplest type of voltage measurement circuits, the *MOSFET source follower*. As shown in Fig. 7.14, this circuit consists of just one transistor and one resistor, which is often implemented as another transistor called active load [27]. The output voltage of this source follower circuit is essentially given by

$$V_{out} = fV_{in} - V_0 \tag{7.6}$$

with a transistor-dependent multiplication factor f of 0.6-0.8 and an offset voltage V_0 of several hundred millivolts. In practice, one or a few such source follower stages are employed in series, to obtain low enough output impedance while maintaining the required read-out speed. At first sight it is surprising that such a simple circuit with a

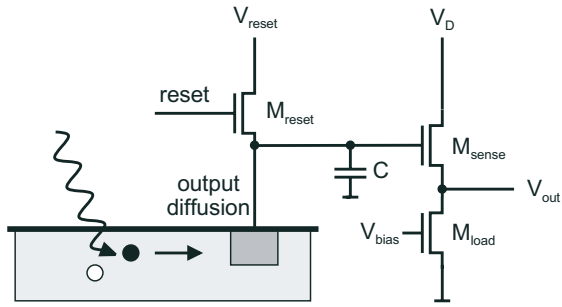


Figure 7.15: *Complete single-stage output circuit of a typical image sensor, consisting of a floating diffusion, a reset transistor, and a single-stage source follower as shown in Fig. 7.14.*

gain of less than unity is used in high-sensitivity image sensors. The reason for this is that the photocharge conversion gain is provided by the effective input capacitance, which is kept as small as possible. Today's best image sensors have an effective input capacitance of around 15 fF, corresponding to a voltage increase of around 10 μV per electron. Taking the circuits' overall gain of less than unity into account, one arrives at the so-called sensitivity of the image sensor, expressed in μV per electrons. Typical sensitivities of state-of-the-art CCD and CMOS image sensors are between 5 and 10 μV per electron.

7.5.3 Noise sources in MOSFETs

Based on a source follower circuit, a typical output stage of an image sensor consists of the components shown in Fig. 7.15. The photocharge is transported to a diffusion (either the output diffusion of a CCD or the photodiode itself) that is connected to the gate of the source-follower MOSFET. Before measurement of each individual photocharge packet, the diffusion and the connected gate are biased to a reference voltage using a so-called reset MOSFET. Three main noise sources can be identified in such a circuit [26], whose influences are referenced back to the input of the source-follower MOSFET, contributing to an effective rms charge measurement uncertainty ΔQ.

Reset or kTC noise. The channel of the reset transistor exhibits Johnson noise similar to an ordinary resistor. This causes statistical fluctuations in the observed reset voltage levels, which result in effective charge noise ΔQ_{reset} given by

$$\Delta Q_{reset} = \sqrt{kTC} \tag{7.7}$$

for the effective input capacitance C, at the temperature T and using Boltzmann's constant k.

Flicker or 1/f noise. Statistical fluctuations in the mobility and charge carrier concentration of the source follower transistor's channel cause an effective charge noise $\Delta Q_{flicker}$ described by

$$\Delta Q_{flicker} \propto C\sqrt{\frac{I^A B}{g_m^2 f C_{ox} W L}} \tag{7.8}$$

at frequency f, for current I, bandwidth B, transistor length L, and width W, oxide capacitance C_{ox}, process-dependent flicker noise constant A, which is typically between 0.5 and 2, and the transistor's transconductance g_m.

Thermal noise. Johnson noise in the source follower transistor's channel can also be referred back to the input, resulting in thermally generated charge noise $\Delta Q_{thermal}$ given by

$$\Delta Q_{thermal} = C\sqrt{\frac{4kTB\alpha}{g_m}} \tag{7.9}$$

using the same parameters as in the preceding.

In practice, the first two noise sources can be essentially eliminated by a signal-processing technique called *correlated double sampling* (CDS) [28]: reset noise is canceled by a two-stage process, in which the diffusion is preset to a reference voltage and a first measurement is made of this voltage level. In a second step, the photocharge is transferred to the diffusion, and a second measurement is made. The difference between these two measurements is free of reset noise and contains only information about the photocharge of interest. Because CDS is a temporal high-pass filter, flicker noise with its low-frequency dominance is effectively canceled at the same time.

The thermal noise contribution cannot be reduced using signal-processing techniques, and it is obvious from Eq. (7.9) what can be done to minimize thermal noise. Reduction of temperature (in astronomical applications down to -120 °C) not only lowers charge noise levels [29] but the *dark current* contribution can be reduced to values as low as one electron per day per pixel. As a rule of thumb, dark current in silicon doubles for each increase in temperature of around 8–9 °C.

Often the reduction in temperature is combined with a reduction of the readout bandwidth to 50–100 kHz, leading to a charge noise level of around one electron [30]. Another technique of bandwidth reduction is the repetitive, nondestructive measurement of photocharge with output signal averaging, as carried out in the Skipper CCD [31]. Charge

noise levels of 0.3 electrons rms have been obtained in this way. As can be seen in Eq. (7.9) the dominant factor in noise performance is the effective input capacitance. This has been lowered to values of less than 1 fF using the so-called double-gate MOSFET [32], corresponding to a sensitivity of more than 200 μV per electron and an effective charge noise level of less than one electron at room temperature and at video frequencies. As the maximum photocharge such an output stage can handle is about 10,000 electrons, the DR is limited to about 80 dB.

7.6 Architectures of image sensors

For the acquisition of 1-D and 2-D distributions of incident light, arrays of pixel are required. Such arrays can be realized as an arrangement of CCD columns or as suitably placed and interconnected individual photodiodes as described in Section 7.3.1. Depending on the choice of arrangement and interconnection, different types of image sensors result.

7.6.1 Frame-transfer charge-coupled devices

The simplest type of CCD image sensor is the *frame-transfer* (FT) CCD. As illustrated in Fig. 7.16, it consists of three CCD sections. One CCD area (A register) is used for the conversion of photons into photocharge during the exposure time and for the storage of this photocharge in the pixels. This 2-D photocharge distribution is subsequently shifted down into another CCD area (B register), which is covered with an opaque metal shield. From the B register, an image row at a time is shifted down into a CCD line (C register), with which the photocharges are transported laterally to the output amplifier, so that the content of this image row can be accessed sequentially.

The disadvantage of the FT-CCD principle is the after-exposure of bright areas that can occur when the photocharge pattern is transported from the A register into the light-shielded B register. This occurs because the A register remains light-sensitive during the vertical photocharge transportation time. The after-exposure effect in FT-CCDs can create saturated (“bright”) columns without any contrast information. For this reason, high-quality FT-CCD cameras employ a mechanical shutter, shielding the A register from incident light during the vertical photocharge transportation time.

The big advantage of the FT-CCD is that the whole A register area is photosensitive; one speaks of an *optical fill factor* of 100 %. Because the A register is covered with polysilicon CCD electrodes that tend to absorb in the blue and UV, an FT-CCD is not very sensitive in the blue spectral region. For special applications this can be remedied by thin-

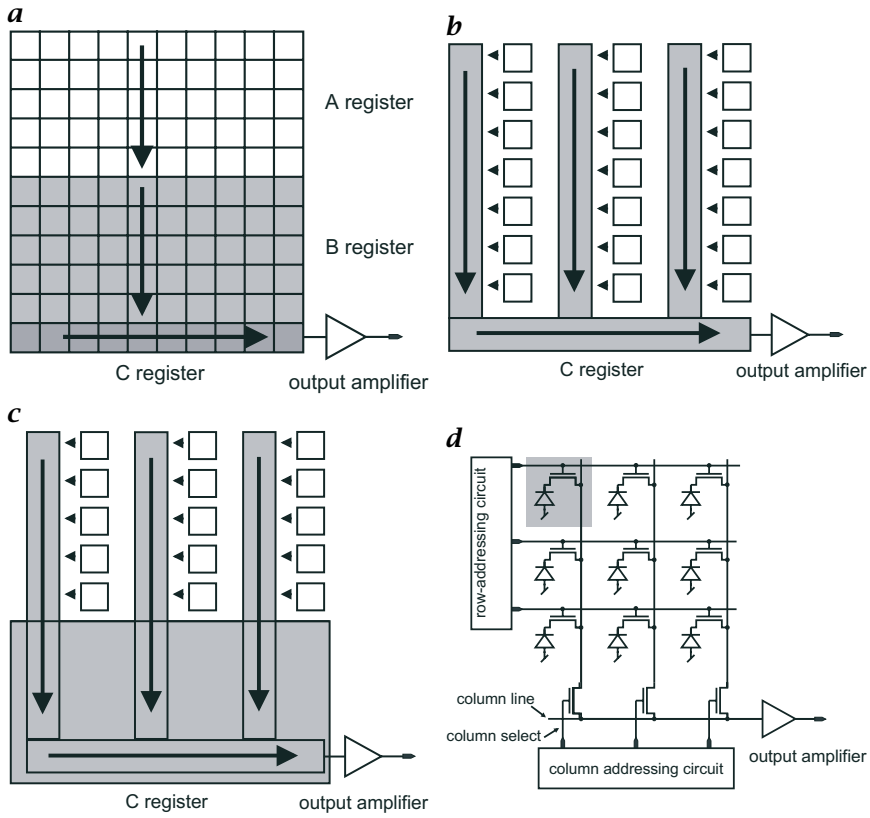


Figure 7.16: *The four most important architectures of solid-state image sensors:* ***a*** *frame-transfer (FT) CCD with its three registers;* ***b*** *interline-transfer (IT) CCD with column light shields for vertical charge transfer;* ***c*** *field-interline-transfer (FIT) CCD, combining FT-CCD and IT-CCD principles for studio and broadcast applications;* ***d*** *traditional photodiode array image sensor with one photodiode and one selection transistor per pixel.*

ning down an FT-CCD to about 10 μm thickness and by illuminating it from the back. Such back-side illuminated FT-CCDs offer 100% fill factor, an excellent response over the whole visible spectrum, and they are the image sensors of choice for scientific and astronomical applications.

7.6.2 Interline-transfer charge-coupled devices

In consumer applications, a mechanical shutter is impractical to use, and for this reason FT-CCDs are rarely used in video and surveillance cameras. Rather, the *interline-transfer* (IT) CCD principle is employed,

as illustrated in Fig. 7.16b. Photocharge is collected in the individual pixels, and after the exposure time the photocharge is transferred via the pixels' transfer register into a corresponding vertical CCD column. These CCD columns are shielded from light with an opaque metal layer. A 2-D photocharge distribution can therefore be shifted downwards, one row at a time, into the horizontal output register, from where the photocharge packets are read out sequentially. As the vertical CCD columns are shielded, the after-exposure problem is much less severe than in FT-CCDs. One pays for this with a reduced fill factor, because the column light shields reduce the available photosensitive area on the image sensor's surface. The typical fill factor of an IT-CCD is about 30 %, reducing the total sensitivity to about a third of that observed in FT-CCDs.

With the IT-CCD principle a very useful functionality becomes available: because there is essentially no time-constraint in exposing the pixels and transferring their accumulated photocharge under the shielded columns, one can implement an *electronic shutter*. The exposure time can be as short as a few 10 μs, extending up to several seconds in cameras not conforming to a video standard. The exposure time is essentially bounded by the dark current, which depends strongly on temperature, as described in Section 7.5.2. The desirable properties of the IT-CCD make it the image sensor of choice for most of today's video and surveillance cameras, especially for consumer applications. In order to increase the optical fill factor of IT-CCDs, some manufacturers supply each pixel with its own *microlens*, so that more light can be directed to the IT-CCD's photosensitive surface. An even more efficient, albeit more expensive improvement is the coverage of an IT-CCD with amorphous silicon, with which the optical fill factor can be increased further, close to 100 %.

7.6.3 Field-interline-transfer charge-coupled devices

Although the column light shield in the IT-CCD is an efficient light blocker, there is always some residual photocharge seeping into the columns from the sides. For this reason, an IT-CCD can still show some after-exposure effects. For professional applications such as video broadcasting, this is considered not acceptable, and a combination FT- and IT-CCD principle has been invented to overcome this problem, the *field-interline-transfer* (FIT) CCD, illustrated in Fig. 7.16c. The upper part of a FIT-CCD really consists of an IT-CCD. The lower part, however, is realized like the B and C registers of an FT-CCD. The FIT-CCD is operated by acquiring an image conventionally, making use of the IT-CCD's variable exposure time functionality. The resulting 2-D photocharge distribution is then shifted quickly under the shielded vertical columns, from where it is transported very fast under the completely shielded in-

termediate storage register. The sequential row-by-row readout is then effectuated from the B and C registers, exactly as in FT-CCDs.

7.6.4 Conventional photodiode (MOS) arrays

A photodiode or MOS array image sensor consists of a 1-D or 2-D arrangement of PDs, each provided with its own selection transistor, as illustrated in Fig. 7.16d. For a description of the PD image sensor's operation, assume that all PDs are precharged to a certain reverse bias voltage, typically 5 V. Under the influence of the incident light, each pixel is discharged to a certain level. A pixel is read out by addressing the corresponding row and column transistors, providing a conducting line from the pixel to the output amplifier. Using this line the pixel is charged up again to the same reverse bias voltage as before. The amplifier measures how much charge is required to do so, and this charge is identical to the photocharge (plus dark current charge) accumulated at the pixel site. In this way, each pixel can be read out individually, at random, and the exposure time is completely under the control of the external addressing electronics.

The random addressing freedom, however, comes at the price of a large capacitance of the conducting line between pixel and output amplifier of several pF. As is obvious from the inspection of Eq. (7.9), this leads to noise levels one or two orders of magnitude larger than in corresponding CCDs image sensors. For this reason, the usage of such traditional PD image sensors has been restricted to applications where the random pixel access is an absolute must. In video applications, CCD technology is used almost exclusively.

7.6.5 Active pixel sensor technology

As just discussed, the noise performance of PD array image sensors is much worse than that of a CCD because of the large effective capacitance the first MOSFET in the output amplifier sees. The logical conclusion is that it should be possible to realize CMOS-compatible PD array image sensors with a noise performance comparable to CCD imagers when this first MOSFET is placed in each pixel. It took surprisingly long until this seemingly trivial observation was made. As a consequence, it led directly to what is called today "*active pixel sensor*" (APS) imaging technology [33]. It is apparently not sufficient just to move the first MOSFET into the pixel, because its input requires a reset mechanism. For this reason, the simplest APS image sensor pixel consists of one photodiode and three MOSFETs as illustrated in Fig. 7.17.

With the reset MOSFET the photodiode and the gate of the source follower MOSFET are precharged to a voltage of typically 3-5 V. The

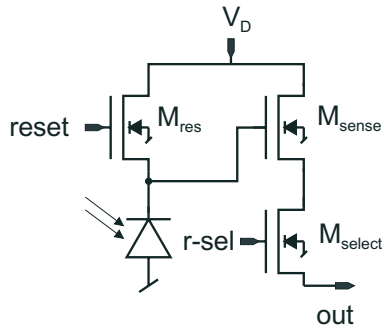


Figure 7.17: *Schematic diagram of an APS pixel, consisting of a photodiode, a reset transistor, a sense transistor, and a row-select transistor. The active load transistor that completes the source-follower circuit is shared by all pixels in a column, and it is therefore needed only once per column.*

photocurrent produced by the photodiode (plus the dark current) discharges the capacitance of the reverse-biased PD. The resulting voltage can then be sensed efficiently with the source-follower MOSFET with a sensitivity that is comparable to that of CCD image sensors. As in the PD array, the third MOSFET is employed as a selection switch with which a row is selected. The active load MOSFET of this APS pixel can be shared by all the pixels in a column, and it does not need to be included in the pixel itself.

The APS technology is very attractive for several reasons: (1) APS image sensors can be produced in standard CMOS technology, opening the way to image sensors with integrated electronic functionality and even complete digital processors; (2) The pixels offer random access similar to PD arrays; (3) The pixel readout is non-destructive, and it can be carried out repeatedly for different exposure times; (4) The exposure times can be programmed electronically; (5) APS image sensors dissipate one or two magnitudes less electrical power than CCDs; (6) APS imagers show less blooming (spilling of electronic charge to adjacent pixels). And (7) APS pixels are more robust under x-ray radiation.

Disadvantages of APS image sensors include the reduced optical fill factor (comparable to that of IT-CCDs), the increased offset noise due to MOSFET threshold variations (see Section 7.9) and the impossibility of performing correlated double sampling for noise reduction as discussed in Section 7.5.3. Fortunately, a combination of APS and CCD technology has been proposed, and the resulting photogate APS pixels offer this functionality [34].

Active pixel image sensors with up to $2\text{k} \times 2\text{k}$ pixels have been realized, with speeds of several thousand frames per second, with an input-referred charge noise of about 30 electrons at room temperature

and video speed, and with a DR of up to 84 dB. Many experts do not doubt, therefore, that CMOS imagers using APS techniques can replace CCD image sensors in many practical applications, and several consumer products in the electronic still and video camera market already contain CMOS imagers.

7.7 Camera and video standards

Although it is possible to realize custom image sensors according to application-specific requirements at lower and lower prices, off-the-shelf standard imagers are likely to be much less expensive. Therefore, one always has to inspect the growing list of image sensors conforming to one of the popular standards, whether or not it might be possible to use one of them for a given application. In the following, today's most important video standards are summarized, together with their salient properties.

7.7.1 RS-170, CCIR, NTSC and PAL

The electrical power systems of most countries in the world offer a mains frequency of either 50 or 60 Hz. As a large proportion of illumination sources operate on this basic frequency or a harmonic of it, the adopted video standard should work with a field or frame rate conforming to it. The obvious reason for this is that beat frequencies between the temporal illumination modulation and the periodic sampling of the camera should be avoided because the resulting aliasing would lead to annoying low-frequency intensity modulation of artificially lighted video sequences. Two major black-and-white *video standards* have therefore been defined, *RS-170* for 60 Hz (as used, e. g., in the U.S. and Japan) and *CCIR* for 50 Hz (as employed in Europe).

Both video standards use *interlacing*, a technique where each *frame* (a complete image in a video sequence) is split into two so-called *fields*. The first field consists of all odd lines in a frame, the second field consists of all even lines. Psychophysical reasons for doing so can be found, for example, in Pratt [35]. Because these standards were defined for a completely analog signal transmission chain, it was never necessary to specify an exact number of pixels per line. In the summarizing Table 7.1, such a number has been calculated, based on the assumption that the digitized pixels are square. This is not part of the standard, however, and relates solely to this mentioned (but by no means unique) choice of square pixel shape.

More information on TV and video standards can be found in Benson [36], especially on the techniques by which timing and synchronization information can be included in the same analog signal waveform.

Table 7.1: *Video image sensor properties*

Video image sensor property	CCIR	NTSC
Frame rate (image sampling rate)	25 Hz	30 Hz
Field rate	50 Hz	60 Hz
Line rate	15.625 kHz	15.735 kHz
Number of lines in a frame	625	525
Number of active lines with video information	576	486
Aspect ratio (width to height ratio of an image)	4:3	4:3
Calculated number of square pixels per line	768	648
Analog video bandwidth	5 MHz	4 MHz
Video information modulation amplitude	700 mV	700 mV
Synchronization information amplitude	-300 mV	-300 mV

Table 7.2: *Video imager formats*

Video imager format	1"	2/3"	1/2"	1/3"	1/4"	1/6"
Image sensor height (mm)	9.6	6.6	4.8	3.6	2.4	2.0
Image sensor width (mm)	12.8	8.8	6.4	4.8	3.2	2.7

The actual geometry of an image sensor depends on the semiconductor technology used for its manufacture. The more advanced the technology, the smaller the pixels and the sensor. Dating back to the times of vidicon vacuum tubes for image sensing, solid-state sensor geometries are specified in terms of equivalent vidicon tube diameters in inches, as listed in Table 7.2.

Today's state-of-the-art video image sensors are already fabricated in 1/4" format, offering an effective *pixel size* of around 4.8 μm. It is not difficult to predict that shrinking geometries of semiconductor processes will make it possible to reduce this small pixel size even further. For video image sensors a continuing significant pixel size reduction will not make much sense, however, because the imaging quality of the TV lenses and the diffraction limit represent a lower bound to the reasonable pixel pitch, which is estimated to be around 3-4 μm [29].

The RS-170 and the CCIR standard do not foresee any color information in the video signal. Because color can be such an important source of information, color extensions to the existing black-and-white video standards were defined. The two most important color video standards are NTSC (for 60 Hz systems) and PAL (for 50 Hz systems). Both rely on a separation of the luminance (black-and-white signal) and the chrominance (two basic color channels). Whereas luminance is transmitted

Table 7.3: HDTV sensor properties

HDTV sensor property	value
Frame rate	30 Hz
Field rate	60 Hz
Line rate	33.75 kHz
Number of lines	1050
Aspect ratio	16:9
Number of square pixels per line	1868
Analog video bandwidth	75 MHz

in exactly the same way as in the black-and-white standard and the chrominance is just a high-frequency modulation of the luminance signal, a low-pass filtered version of the color signal again becomes a valid black-and-white video signal according to the corresponding standard [36].

7.7.2 High-definition television

The forthcoming extension of today's video standards, the so-called *high-definition television* (HDTV), consists essentially of a doubling of the number of lines, while maintaining the basic interlacing and field/frame repetition rate. Although it would be desirable to migrate from interlaced to progressive transmission (instead of showing two fields per frame in succession, full frames are transmitted), the resulting doubling of video bandwidth is currently difficult to justify for video applications. Additionally, the aspect ratio is changed to 16:9 to reflect viewer preference of more elongated picture formats. An HDTV image sensor conforming to the 60-Hz standard in the U.S. will therefore have the properties summarized in Table 7.3.

Most of today's state-of-the-art HDTV image sensors are 2/3" devices with around $1\text{k} \times 2\text{k}$ pixels (so-called 2M imagers), exhibiting a pixel pitch of around 5 μm. The very high pixel rate of 75 MHz is technically so difficult to realize that HDTV sensors usually have two output amplifiers (two taps) that are operated in parallel, each of which offers a pixel rate of 38 MHz.

As soon as HDTV image sensors and cameras realized around them are available at prices of around $ 1000, they will be embraced not only by the electronic photography field but certainly also by machine vision and automatic fabrication markets. For the next few years, however, HDTV technology is still considered to be too expensive for such applications.

7.7.3 Random pixel access and format

The emerging CMOS (APS) imager field not only brings low-cost and low-power image sensors for video and electronic photography applications to the market, but new modes of utilizing an image sensor become possible. In contrast to a CCD imager where a full frame must be read out, CMOS imagers offer random access and random format generation by suitable pixel sequence readout. No standards have been defined yet, and it is also not expected that standards for APS image sensors will be defined. Nevertheless, such image sensors share some properties in their addressing, making it worthwhile to discuss their mode of operation.

The CMOS imagers of APS or PD type are operated by selecting a row of pixels with a suitable digital address, for example 9 bits for one out of 512 lines. With another digital address, a column is selected, so that one individual pixel can now be accessed, whose content is available either as an analog or a digital value at the output of the image sensor chip. In the case of a traditional PD array or an APS with photogates, this readout is destructive, and it implies a reset of the pixel. In a standard APS image sensor, however, such readouts can be repeated at different times without destroying the collected photocharge. Finally, a reset signal needs to be applied at a suitable time, which often causes a reset of a complete row of pixels whose row address is selected at that time.

The advantage of this addressing scheme is that individual pixels can be accessed at random. Programmable pixel patterns become possible, such as subsampled 2-D arrays, circular or linear 1-D arrays, or small 2-D regions of interest sampled at a very high rate. It is obvious that in this way rectangular arrays of arbitrary, programmable aspect ratio can be generated, so that image formats can be changed dynamically, for example, adapted to a problem at hand. With the advent of such programmable CMOS imagers, novel types of image processing strategies can be developed, as foreseen by various *active vision* concepts [6].

Since an increasing number of image sensing applications are no longer based on video standards that call for interlaced signals (two successive fields form a complete frame), many new types of image sensors and camera types can be read out one full frame at a time. This so-called *progressive scan* read-out implies that all pixels on an image sensor are transmitted sequentially, starting at the top and ending with the last line; de-interlacing (for display on a computer screen, or for image processing or printing) is no longer required. Another consequence of the departure from video standards is the implementation of image sensor formats and pixel numbers that are related in some form to existing computer display or storage standards: High-speed image

sensors are available commercially offering a resolution of 256×256 or 512×512 pixels. For videoconferencing and other low-resolution imaging applications, image sensors according to the CIF ("*common intermediate format*") standard with a resolution of $288\text{v} \times 352\text{h}$ have been developed, that is, 288 pixels in the vertical direction and 352 pixels in the horizontal direction. An increasingly popular type of image sensor for consumer applications is the VGA ("*Video Graphics Array*") imager with a resolution of $480\text{v} \times 640\text{h}$ pixels. For applications requiring higher quality picture acquisition, image sensors in SVGA ("*Super Video Graphics Array*") format with a resolution of $600\text{v} \times 800\text{h}$ pixels and in XGA ("*Extended Graphics Adapter*") format with $768\text{v} \times 1024\text{h}$ pixels are offered commercially. In many conventional video image sensors, the pixels were rectangular with an aspect ratio deviating from the 1:1 square format. As many image-processing applications would profit from square pixels for the extraction of correct geometric information without reformatting, most of today's image sensor pixels are square. This simplifies interpretation, display and printing, as well as the use in metrological applications of the acquired images.

In summary, the recent trend in image sensors is away from the traditional video standards towards close connections with computer standards and digital image-processing applications. Modern solid-state image sensors and digital cameras are no longer regarded as dumb external data gathering devices; instead they are increasingly being made part of an integrated, intelligent and dynamic information acquisition and extraction system.

7.7.4 Analog signal transmission of video information

For the past 70 years, the preferred way of transmitting video information has been in the form of an analog electrical signal over *coaxial cable*. Even today, this medium is used in demanding professional applications as well as in bringing a multitude of TV channels to the house ("cable TV"). Coaxial cable has a central conducting core that carries the signal. It is completely surrounded by a cylindrical shielding electrode, acting also as the ground terminal. The two conductors are separated by a stiff, insulating plastic material, making the coaxial cable robust against mechanical forces, so that the cable's electrical parameters are precisely known. For an understanding of the transportation properties of electrical signals, the coaxial cable must be modeled as a transmission line [37] with its characteristic impedance, transmission speed, and frequency-dependent attenuation. Typical values for the characteristic impedance are 50 or $75\,\Omega$ and the transmission speed is about half the value of the speed of light in vacuum, that is, about 150,000 km/s. Because signal attenuation occurs exponentially with the product of transmission distance and the square root of frequency

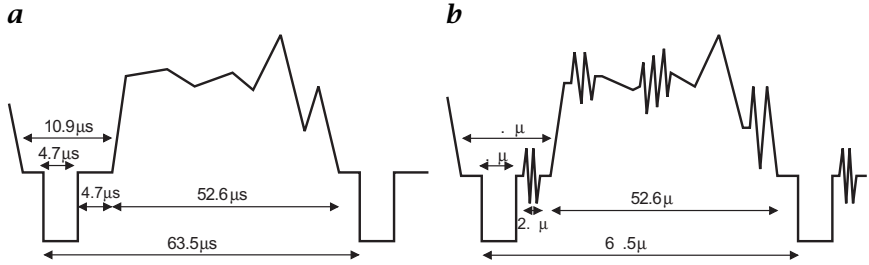


Figure 7.18: *Illustration of how image and synchronization information is combined in the video signal according to the RS-170 and NTSC standard:* ***a*** *one line of black-and-white video information according to the RS-170 standard;* ***b*** *one line of color video information according to the NTSC standard. Note the short color burst after the horizontal sync pulse and the high-frequency modulations in the image signal indicating saturated colors.*

[38], the longer the transmission distance the smaller the transmittable bandwidth. In practice, a maximum bandwidth of less than 1 GHz over a transmission distance of a few 1000 m is employed [39].

The bandwidth of video signals according to one of the TV standards summarized in Section 7.7.1 is restricted to 4–5 MHz. It must be noted, though, that high-quality image sensors used for video applications are capable of delivering much more detailed image information. The necessary bandwidth for such a high-quality video signal can be estimated as follows: according to the RS-170 black-and-white video standard (see Table 7.1), 30 frames/s have to be transmitted, each with a total number of $525 \times 648 = 340{,}200$ square pixels. This corresponds to more than 10 million pixels/s. According to the Nyquist criterion, the analog bandwidth required for the transmission of this sampled data has to be at least twice as large, so that an analog bandwidth of around 25 MHz is necessary for the full exploitation of the original video signal's contents.

Since the image sequence is sent line for line and picture for picture as a continuous electrical signal over a single coaxial cable, it is necessary to provide this video signal with all *synchronization* information, so that the receiver can reconstruct the original image sequence. In the following, typical times are given that were adopted for the RS-170 black-and-white video standard [36].

The synchronization information is provided by preceding each line with a negative voltage pulse of -300 mV that lasts for 4.7 µs, the so-called horizontal synchronization ("sync") signal (Fig. 7.18a). The horizontal sync pulse is followed by a voltage level that corresponds to no light ("black") in a picture. This black level is very useful because temperature-dependent effects (such as dark current) can cause the dark signal level to change with time. After this black level, which lasts

for another 4.7 µs, the video information of the corresponding image line is transmitted, taking 52.6 µs ("active video"). This time together with the horizontal blanking time of 10.9 µs is the time taken by a complete video line, that is, 63.5 µs in the RS-170 standard. The signal voltage swing of the active video is 700 mV so that the video signal shows a peak-to-peak voltage of 1 V. The information that a new field or a new frame starts is transmitted by a specified sequence of sync pulses, typically lasting several 100 µs. With this, the complete information is available for the receiver to reconstruct the black-and-white image sequence.

To make color video signals compatible with black-and-white video signals, the color information is encoded in the amplitude and the exact phase of high-frequency oscillations overlaid on the black-and-white video signal [40]. For this, high-frequency color synchronization signals (2.3 µs long color "bursts," consisting of about 10 cycles) are introduced as part of the black-level signal. They are illustrated schematically in Fig. 7.18b, together with the typical high-frequency oscillations representing saturated colors in the video signal.

As mentioned before, video information is transported as a travelling wave along the transmission line represented by the coaxial cable. At the end of an open transmission line, such a wave would be reflected, traveling back to the signal source and distorting the video signal. It is important, therefore, to dissipate the wave energy at the end of the transmission line by terminating it with a resistor of the same value as the characteristic impedance, that is, 50 or 75 Ω for standard coaxial cable used for video signals.

Although the coaxial cable can be tapped with high-impedance devices, for example by using T-connectors, the branches should be relatively short, so that the wave reflection effects remain insignificant. Again, the transmission line must be terminated once at the end of this side-branch coaxial cable. Professional video equipment such as a monitor often provides built-in termination resistors that can be switched on and off, and connectors are provided for coax cable coming in and going out, so that the video signal can be looped through or terminated in the case of the last piece of equipment in the line. If the rules of proper line termination and short side branches are not followed, different types of ringing and ghost artifacts result in the pictures.

7.7.5 Color chips and color cameras

The goal of any high-performance camera system is to capture accurately the perceptible contents of a scene for subsequent faithful reproduction. The black-and-white image sensors and cameras discussed so far can do this only for the brightness sensation; the very rich perception of *color* requires additional information, as described in Sec-

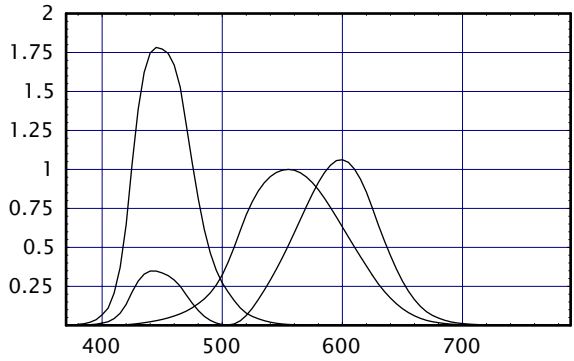


Figure 7.19: *CIE tristimulus curves $\overline{x}$, $\overline{y}$, and $\overline{z}$ as a function of wavelength. The $\overline{y}$ curve is the standard luminosity function describing the human perception of brightness.*

tion 11.3. It is surprising to realize that, according to Grassman's Laws [40], only two additional spectral representations of a scene are sufficient for the complete representation of a color scene as it can be perceived by a human observer: according to the trireceptor theory of vision, only three different types of color receptors are present in the human eye, each of which exhibits a different spectral sensitivity distribution. It is sufficient, therefore, to acquire a color scene through three different types of spectral filters for a complete representation of its perceptible content, describable as local "*brightness*," "*hue*" and "*saturation*."

To provide colorimetric standards and objective measurements of colorimetric performance, in 1931 the Commission Internationale de l'Eclairage (CIE) adopted the so-called CIE tristimulus curves for the "standard observer," $\overline{x}$, $\overline{y}$ and $\overline{z}$, illustrated in Fig. 7.19 [41]. These curves were chosen such that $\overline{y}$ represents the standard "luminosity" function, describing the spectral distribution of the human sensations of brightness. Loosely speaking, the three tristimulus curves correspond to the sensations of red (R), green (G) and blue (B). Any type of color camera must acquire three types of images through spectral filter functions, each of which is a different linear combination of the CIE tristimulus curves.

For the best performance, a color camera is built by providing special beam-splitting optics and by arranging three black-and-white image sensors so that they see an identical portion of a scene. Each image sensor is covered with its own color filter, as just described, and together the three image sensors acquire the complete colorimetric information about a scene. Such three-chip color cameras are employed in professional and studio cameras. They are quite expensive, unfortunately,

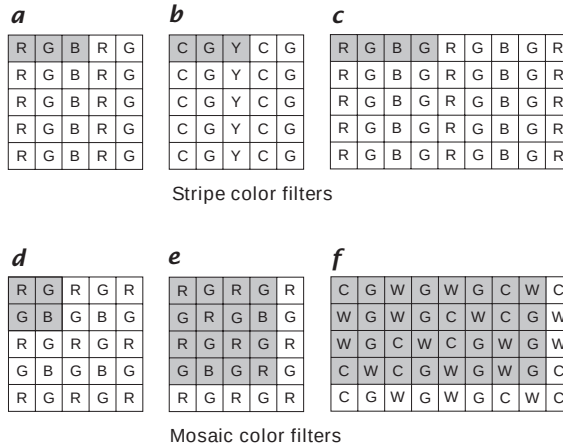


Figure 7.20: *Illustration of different color filter types for single-chip color sensors. The unit cell (basic arrangement of color filter patches that is periodically repeated on the image sensor) is shown as shaded rectangle:* ***a*** *primary color (RGB) stripe filter with* 3×1 *unit cell;* ***b*** *complementary color (CGY) stripe filter with* 3×1 *unit cell;* ***c*** *primary color (RGB) stripe filter with* 4×1 *unit cell;* ***d*** *Bayer color mosaic filter with* 2×2 *unit cell;* ***e*** *Bayer color mosaic filter with* 4×4 *unit cell;* ***f*** *shift-8 color mosaic filter using complementary colors in an* 8×4 *unit cell.*

because they have to employ costly beam-splitting objects, the three image sensors have to be aligned according to close tolerances (registration to sub-pixel accuracy), and three high-quality image sensors must be used, each requiring its proper driving electronics.

For these reasons, it is highly desirable to realize a color camera with just one single black-and-white image sensor and a suitable pattern of pixel-individual color filters on top. Several techniques have been used for the implementation of such a single-chip color camera. They are either based on 1-D color stripe filters (Fig. 7.20a-c) or on 2-D color mosaics (Fig. 7.20d-f).

The simplest arrangement is the RGB color stripe pattern shown in Fig. 7.20a. Its obvious drawback is its sensitivity to periodic objects, producing so-called moiré and color-aliasing effects [15]. Instead of the primary RGB filters, one can also use the complementary colors cyan (C=G+B), yellow (Y=R+G), and magenta (M=R+B), or even transparent white (W=R+G+B). An example of such a complementary stripe filter pattern is shown in Fig. 7.20b. Compared to the primary color stripe filter in Fig. 7.20a, this filter can be simpler to fabricate, and because it accepts more light, it might offer an improved signal-to-noise performance. Another example of a stripe filter is shown in Fig. 7.20c, illustrating the use of more green than red or blue information and the larger filter period of four pixels. This reflects the property of the hu-

man eye that spatial resolution is largest in the green, less pronounced in the red, and least developed in the blue spectral band. Much better performance is achieved with 2-D mosaic color filters. A popular color filter is the Bayer pattern with its 2×2 pixel unit cell shown in Fig. 7.20d [42]. An improved form makes even better use of the different spatial resolution for the three filter curves, resulting in the 4×4 pixel unit cell shown in Fig. 7.20e [42]. In this filter pattern, half of the color filters are green, 3/8 are red and only 1/8 are blue. The larger the unit cell period, the better a color filter's ability to prevent aliasing and moiré effect. A very effective color pattern making use of complementary colors is shown in Fig. 7.20f [43]. It uses a 4×8 pixel unit cell in such a way that the required signal processing is relatively simple to realize using conventional electronics [44]. The least amount of aliasing is produced by a color mosaic with an aperiodic color pattern. Although this is well known in theory, no commercial product has been offered yet with such a random color pattern, which would also require precise knowledge of the image sensor's complete color pattern for the accurate extraction of color information.

7.7.6 Digital camera technology

For the foreseeable future, solid-state cameras are based on the linear conversion of the local intensity of incident light into a proportional electronic charge or voltage. For this reason, they have to be considered analog devices, working over an amazingly large dynamic range of at least ten decades [14]. As an increasingly large number of applications call for a digital representation of the acquired images, it becomes more and more desirable to work with standardized formats for the transmission of sequences of digital images. Examples of such standards include the *digital studio standard CCIR-601*, the compressed *videoconference standard CCITT H.261*, and the compressed multimedia standards of ISO's *MPEG* working group [45]. The number of digital image standards proliferates because the use of computers for storing, processing, and displaying images makes it so easy to implement (or convert) any type of digital scene representation.

The traditional approach for the conversion of a camera's analog image sequence into a stream of digital image data is the use of a frame-store. Built around an analog-to-digital converter (ADC) a frame-store digitizes the incoming analog video signal into a pixel stream by making use of the synchronization information contained in the video signal (see Section 7.7.4) or provided separately. To simplify the acquisition of digital images, it is highly desirable to replace the conventional combination of analog camera plus frame-store with a "digital camera," which provides directly digital image data in a suitable format. As already discussed, the exact nature of this digital image format is of reduced

significance because it is no longer difficult to convert this digital image format into another, more appropriate one for a specific application.

The problem of digital cameras is not one of digital image format but rather one of fast and reliable transmission of this digital image information from the camera to receiving equipment such as a computer. For simplicity, this transmission should make use of existing digital communication lines. Since a typical uncompressed image contains several hundred thousand to a few million bytes, the transmission speed of the digital line is an important issue. The ubiquitous serial communications standard RS-232C is limited to some hundred kbits/s and does not lend itself very well to the fast transmission of image information. The parallel port according to the Centronics standard can be used for the transmission of digital image information at a data rate of about 100 kBytes/s. Improvements of the parallel port standard—the Extended Capability Port ECP and the Enhanced Parallel Port EPP—allow the byte-wise bidirectional exchange of information at data rates of up to 2 MBytes/s.

The increasingly popular USB (*Universal Serial Bus*) standard allows the transmission of 12 Mbits/s, and the first digital cameras with a USB interface are commercially available [46]. For the uncomplicated transmission of large amounts of image data in real-time applications, the *IEEE 1394* (*FireWire*) serial bus is the medium of choice. Present specifications allow a data rate of 400 Mbit/s and future extensions (e.g., the proposed IEEE 1394b standard) are foreseen to offer more than 1Gbit/s. Since FireWire interfaces are more complex and more expensive than USB interfaces, FireWire cameras are more expensive than USB cameras, and the two standards are likely to coexist for quite some time because they serve different application fields [46].

7.8 Semiconductor technology for image sensing

Driven by the apparently insatiable demand for faster digital processors and memories with ever-increasing storage capacity, silicon fabrication technology develops at an amazing pace. Minimum feature dimensions are shrinking and diameters of wafers are increasing continuously. More functionality on larger and less expensive chips is the results of this development. Image sensors profit directly from this development, and what is true for digital computing components is also true for image sensors.

7.8.1 Shrinking design rules for more and smaller pixels

One of the crucial parameters of semiconductor technology is the minimum feature size, also called design rules. As illustrated in Fig. 7.21,

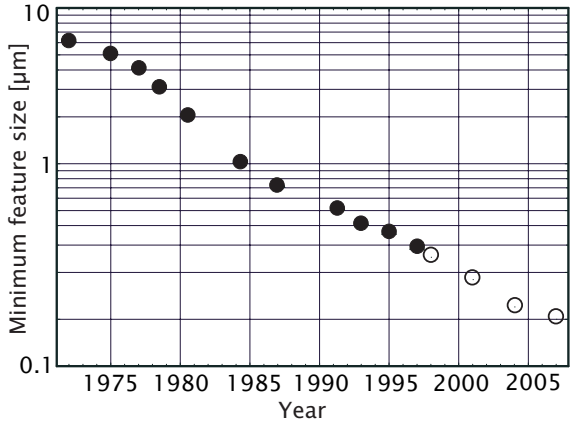


Figure 7.21: *Evolution of the minimum feature size in silicon-based semiconductor technology, showing a reliable exponential reduction of about 10 % per year. Full circles indicate values achieved by advanced manufacturers; open circles represent the semiconductor industry association's roadmap predictions.*

the design rules have been reduced by about 10% each year for the past several decades. This trend is expected to continue for at least another ten years. As a direct consequence of this, the pixel size in CCD image sensors has also been reduced continuously, and a similar pixel pitch reduction of about 10% per year can be observed, as shown in Fig. 7.22. It is doubtful whether CCD image sensors and semiconductor technology will be further developed in parallel in future. As mentioned in Section 7.7.1, the optical resolution of TV lenses and the diffraction limit will probably pose a lower limit of 3-4 μm to the pixel size of practical image sensors. As can be seen in Fig. 7.22, this limit has been reached today. It is suspected that smaller pixels might make sense only for special functionality such as high-resolution color pixels, or high-speed image acquisition with storage in each pixel, etc. [7]. Another development of silicon technology, the increase in wafer diameter, has led to wafer-size image sensors with multimillion pixel resolution. The evolution of CCD image sensors with a record number of pixels is plotted in Fig. 7.23. There was a marked lack of progress in the number of pixels in the years 1975-1983. Our interpretation of this phenomenon is that the number of pixels in image sensors was increased rapidly by different research groups, until enough pixels on an image sensor were available for the realization of solid-state video cameras. After this initial period of research activity, it took significant time and effort to develop the semiconductor technology that was necessary for the mass-fabrication of these devices with high enough yield. It was only then, after 1983, that the technology was pushed again, and im-

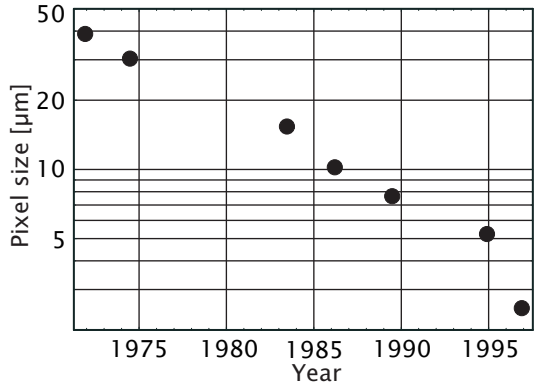


Figure 7.22: *Evolution of the minimum pixel size in CCD image sensors, following the exponential decrease of the minimum feature size shown in Fig. 7.21: an average reduction rate of about 10% per year is observed. The current record is a pixel pitch of 2.4 µm [48].*

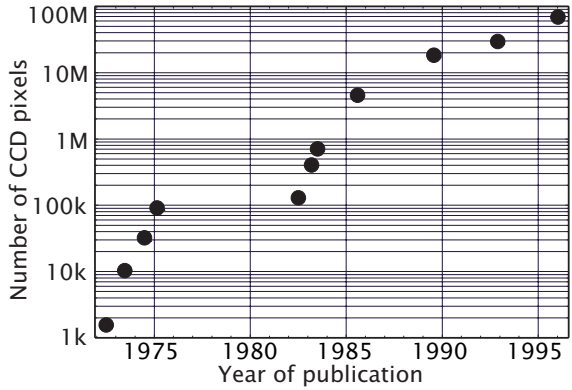


Figure 7.23: *Evolution of the maximum number of pixels on a CCD image sensor. Today's record is held by a wafer-scale CCD with 66 million pixels on an area of* $9 \times 12\,cm^2$.

age sensors with increasingly large numbers of pixels were fabricated. The current world record is held by a $9 \times 12\,mm^2$ large CCD image sensor offering $7168 \times 9216 = 66$ million pixels, fabricated on a 150-mm diameter silicon wafer [47]. Because of the large cost of such devices, these huge image sensors find applications only in special fields such as astronomy.

7.8.2 Multi-project fabrication processes for low-cost prototyping

The fabrication of silicon circuits is not restricted only to the production of large quantities of ICs. Today, many so-called silicon foundries offer their production services for varying numbers of fabricated integrated circuits, down to prototyping quantities of just 5-10 pieces. Such a service has become possible through the sharing of costs for photomask generation and silicon wafer processing: several users share the total costs, resulting in a reasonable cost for the individual customer, who obtains only a small number of fabricated ICs. Such multi-project wafer (MPW) services are available mainly for CMOS technology, but there are also silicon foundries offering CCD and CMOS/CCD processes. The individual customer just sends in his electronically generated circuit layouts, and 8–10 weeks later he receives the ordered number of finished ICs.

A typical MPW price for about 10 fabricated and packaged integrated circuits in 1 μm CMOS technology, each with an area of about 5 mm^2, is around $ 4000. For educational institutions, much lower rates are offered by government sponsoring agencies such as MOSIS (see reference on MOSIS fabrication service, MOSIS [49]).

Using such MPW services, it has become possible not only to predict the behavior of custom image sensors and analog and digital signal processing circuits by computer simulation, but one can also realize quite quickly and inexpensively prototypes with which the salient properties of the application-specific photosensors can be verified in practice.

7.9 Practical limitations of semiconductor photosensors

Due to the analog nature of the pixels in a semiconductor photosensor, it is not possible to fabricate all pixels with identical properties, and often some pixels on an imager will be defective. It is therefore important for a machine vision system architect to have an idea about typical limitations and shortcomings of practical image sensors.

7.9.1 Pixel nonuniformity and dead pixels

Because of slightly varying geometries of CCD and APS pixels, their effective area and therefore their gain are not identical. These gain variations are of the order of 1-5 %, and for precision measurements, a multiplicative correction of this effect is required.

In APS pixels, where the individual source-follower transistors in the pixels show offset voltage fluctuations, an offset uncertainty of the order of 10 mV is observed. This results in APS pixel offset variations of around 1-2 %. These offset variations have to be corrected additively for precision measurements. Because the CCD principle is based on

the virtually complete transfer of photogenerated charge packets from pixel site to pixel site, CCD pixels do not show this type of offset variation.

In applications where dark currents become significant, offset variations are obtained in APS as well as in CCD image sensors because dark current densities can vary from pixel to pixel in any type of semiconductor image sensor. It might even be possible that the dark current is so high in a few so-called "hot pixels" that these pixels are completely filled with thermally generated charge during the exposure time. This effect can only be reduced by lowering the temperature or by shortening the exposure time.

Digital memories do not suffer from most localized defects on the semiconductor surface because there are redundant memory cells on the integrated circuit that can replace defective storage cells. In an image sensor, this is of course not possible. For this reason, it is rather difficult to produce a perfect image sensor without any defects. It is not uncommon, therefore, that a few defective ("*dead*") pixels can be encountered on an image sensor. Usually, the position of these dead pixels is stored, and the image content at this place is computed as a function of neighboring values. Such pixel defect densities occur quite infrequently with a percentage of typically less than 0.001-0.01 %.

In CCDs, another type of defect is more consequential, when complete dead columns are encountered; the required correction computation is much more expensive than with single dead pixels. Fortunately, dead columns usually are only encountered in megapixel CCDs of lower grade, while smaller area CCDs for video applications are free of this type of defect.

7.9.2 Sensor nonlinearity

The conversion of light into photocharge is a highly linear process. In silicon, this has been verified for a large dynamic range of at least 10 orders of magnitude [14]. Unfortunately, much of this linearity is lost in the photocharge detection principle that is mainly used in image sensors. Photocharge is stored as the state of discharge of a precharged capacitance, either an MOS capacitance or a photodiode. As the width of the space-charge region depends on the discharge level, the spectral sensitivity and the photometric linearity are a function of the amount of photocharge already stored.

The same problem is encountered in the electronic charge detection circuits that are implemented as source followers after a floating diffusion (see Fig. 7.15). The capacitance of the floating diffusion depends on the voltage on it and therefore on the charge state. This causes nonlinearities in charge sensing.

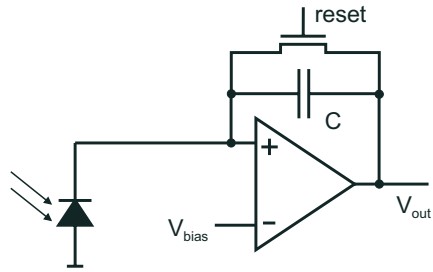


Figure 7.24: *Schematic diagram of a charge detection circuit, providing a high photodetection linearity by keeping the photodiode voltage constant. If the feedback capacitance is replaced by a resistor, a so-called transimpedance amplifier results, converting photocurrent in a proportional voltage with very high linearity.*

The degree of the nonlinearity depends very much on the charge detection (or voltage) range that is used. For differential measurements over a few hundred mV in the middle region of the analog sensor output, nonlinearities can be below 0.1 % [50]. Over the full sensing range, nonlinearities may be as large as a few percent. If the measurement should be highly linear, a proper electronic charge detector circuit must be used in which the voltage at the input is kept constant. Such a charge detector circuit, illustrated in Fig. 7.24, requires a certain amount of silicon floorspace. With state-of-the-art semiconductor technology, pixels become so large that only 1-D arrays have been realized with this technique [51]; in image sensors it is not yet realistic to implement such charge detectors in each pixel. For this reason, image sensing applications for optical metrology in which sub-percent linearity is demanded have to resort to accurate calibration and off-chip digital correction techniques [5].

7.10 The future of image sensing

We have seen that modern semiconductor technology makes it possible to tailor custom photosensors with application-specific functionality to many practical problems. To make this capability widely available, researchers in the field are exploring systematically the possibilities and limitations of silicon photosensing, creating the “photosensor toolbox.” This development is leading to integrated machine vision systems for dedicated applications, and one day perhaps even to “seeing chips” that can perceive in certain ways their environments visually.

Figure 7.25: *Simplified signal chain in traditional solid-state image sensors. Incident light generates a photocurrent in each pixel. The photocurrent is integrated and stored. During sequential scanning, photocharge is detected electronically and read out.*

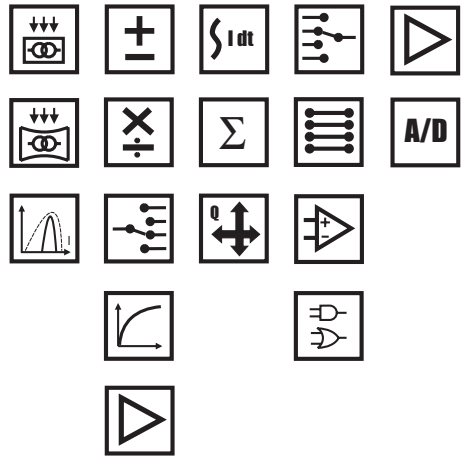


Figure 7.26: *Enhanced image acquisition and processing chain ("photosensor toolbox"), made possible by modern, silicon-based fabrication technology. This picture is a symbolic, incomplete representation of the possibilities offered by image sensors with smart pixels for application-specific photosensors.*

7.10.1 Custom functionality with the photosensor toolbox

In a traditional image sensor, the detection of light is restricted to the simplified signal chain illustrated in Fig. 7.25. A photodiode or a MOS capacitance is employed for the separation of photogenerated charge pairs. This photocurrent is integrated over a certain time, the so-called exposure time, and the photocharges are retained on a suitable storage device. The individual pixels are then sequentially scanned with a suitable switching mechanism. The pixels' charge signals are read out, and they are amplified, one by one, to complete the detection process.

Modern semiconductor processes and the reduced feature sizes for electronic circuits are the basis for functionality in the individual pixels that is much increased above what is illustrated in Fig. 7.25. Some of the possibilities and novel functionality offered at the different stages of the image acquisition chain are symbolized in Fig. 7.26. This forms the basis of the photosensor toolbox, an assortment of well-characterized

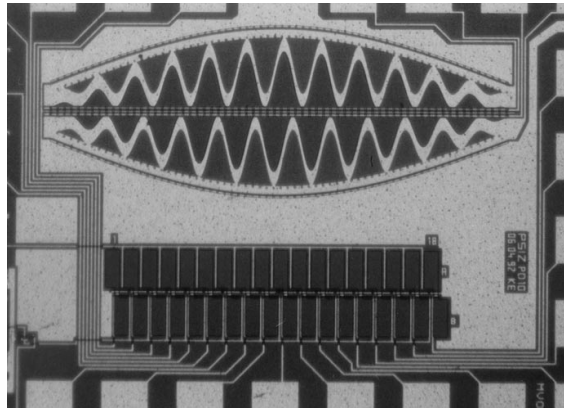


Figure 7.27: *Example of a "Fourier transform" photosensor for a fixed spatial frequency. In the upper part of the chip micrograph, four photodiodes shaped as Hanning-weighted sine and cosine functions are visible. In the lower part, a conventional linear array of photodiodes is visible. This photosensor is the heart of an absolute, very precise optical position encoder. Chip size is around* $0.8 \times 1.2\,mm^2$.

building blocks—electronic and photosensitive devices—with which custom functionality for a specific application can be obtained. The symbolic representations of the capabilities in Fig. 7.26 are briefly summarized in what follows; more details can be found in Seitz [7].

The generation of a photocurrent proportional to the incident light is not restricted to rectangular pixel geometry as employed traditionally. Applications exist wherein a suitable choice of geometry serves as a linear or nonlinear transformation of the incident light distribution. It is possible, for example, to "calculate" the (complex) Fourier transform of a 1-D light distribution with a suitable sensor shape (see also Section 19.4. This is illustrated in Fig. 7.27, with a photosensor optimized for an absolute optical position encoder [52]: while the lower part of the sensor is a conventional linear array of photodiodes, the upper part consists of two harmonic photosensors in quadrature (sine and cosine), weighted with a Hanning (cosine) window. Using such a Fourier photosensor, the position of a 1-D periodic light pattern can be measured with an accuracy of better than 1/1000th of the pattern period, and the speed of such measurements easily surpasses the MHz range.

It is even possible to make the effective shape of the photosensors programmable, that is, electrically adaptable to changed conditions in real-time. This is realized by adding together photogenerated charge packets in the charge domain using the CCD principle, under control of a digital CCD sequencer; for example, a microcontroller [53]. Spatially

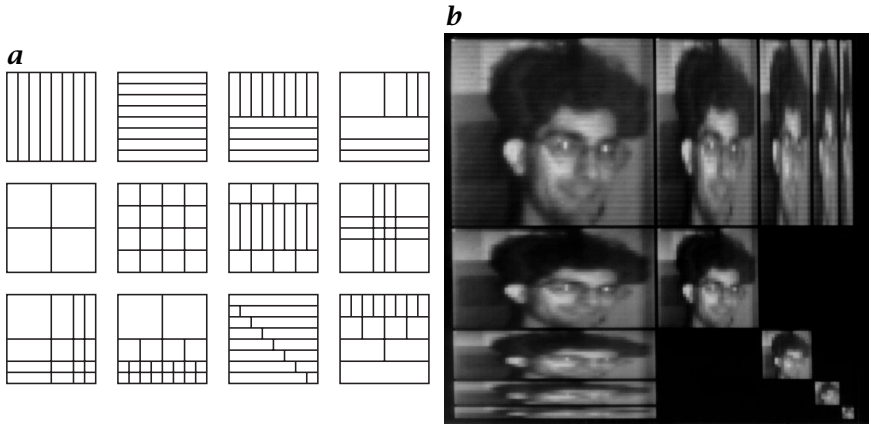


Figure 7.28: *Using the CCD principle, an image sensor with dynamic pixel form and size can be realized:* ***a*** *examples of some practical pixel shapes and sizes, emphasizing that nonuniform resolution and space-variant pixel patterns are also possible, adaptable to a given measurement problem in real-time;* ***b*** *collection of images taken with the dynamic CCD image sensor described in Seitz et al. [53], illustrating different pixel aspect ratios and sizes.*

variant pixel patterns can also be implemented, as shown by a few examples of achievable dynamic pixel patterns illustrated in Fig. 7.28a. This property can be used to adapt the form and size of the pixels for an optimized image data acquisition strategy, where, for example, the resolution is chosen so that a minimum amount of image data is acquired and processed; this is illustrated in the image collection of Fig. 7.28b, taken with different pixel aspect ratios.

The spectral sensitivity of a detector can be changed with an electrical signal, or photosensors with different spectral sensitivity can be stacked on top of each other for a solid-state color pixel without filters. This is illustrated in Fig. 7.29 with the three quantum efficiency curves of three overlaying p-n junctions realized with a standard CMOS process. Shallow junctions are more sensitive to blue light, while deep junctions are predominantly sensitive to red light. Such a simple color sensor already has a CIE general color rendering index of $R_A = 70$, corresponding to a low-quality color video camera [53].

As discussed in Section 7.4, offset currents can be added to or subtracted from the photocurrents, as used for example for nonuniformity or background cancellation. There are also several ways in which multiplication and division can be implemented. Multiplication factors can even be made programmable using a similar voltage-storing technique as described in Section 7.4. These capabilities are still experimental, however, and they have not yet been developed into commercially available image sensor products.

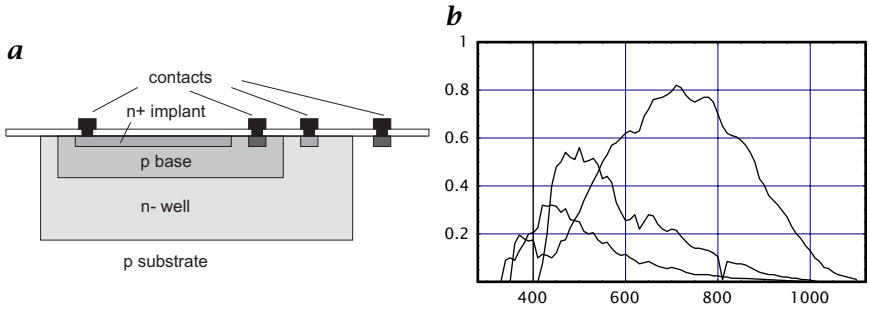


Figure 7.29: *Color pixels can be realized without color filters by employing the wavelength-dependent absorption properties of silicon (Fig. 7.5):* ***a*** *cross section of three overlaying p-n junctions, realized with a commercially available CMOS process;* ***b*** *quantum efficiency curves of the three p-n junctions, showing pronounced blue, green and red sensitivity. A CIE general color rendering index of* $R_A = 70$ *is achieved in practice.*

Another property is that photocurrents can be redirected very quickly, with sub-microsecond switching times, to different electronic circuits for further processing. An example of this capability is the realization of a so-called "lock-in CCD" [54]. Each pixel of this image sensor is capable of synchronously detecting the local phase, amplitude, and offset of a 2-D temporally modulated wave field. In this way the well-known "lock-in" detection for periodic signals can be implemented locally within each pixel, combined with the detection of the light, as used, for example, in optical range cameras based on the time-of-flight principle, described for example in Section 18.5. A micrograph of such an experimental lock-in CCD image sensor offering eight taps (sampling values per signal period) in each pixel is shown in Fig. 7.30a. By using more than three taps per pixel, higher-order moments or higher-order Fourier coefficients can be determined of the modulated light. An application of this is the discrimination of temporal code patterns for the differentiation between various modulated light sources.

The dynamic range of optical signal detection can be increased with a circuit or a device exhibiting nonlinear transfer characteristics, such as the logarithmic pixel described in Section 7.3.5. As mentioned, it is of course also possible to supply each pixel with its own amplifying circuit, as was the original motivation for the development of APS technology. When combined with suitable, exposure-time based multiplication and charge summation capabilities, 2-D photocharge shifts make it possible to realize a CCD image sensor that can acquire and convolve an optical scene simultaneously, with a freely programmable convolution kernel of any extent [55]. Results obtained with an experimental convolution CCD offering 43×32 pixels show that the accuracy of the individual

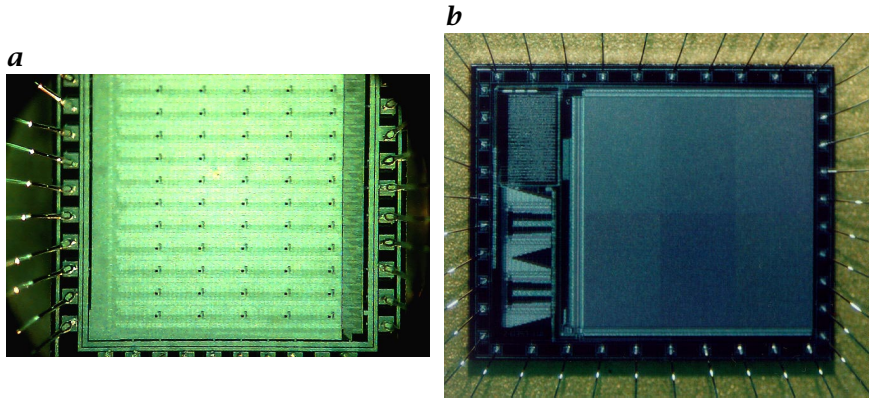


Figure 7.30: ***a*** *Chip micrograph of an experimental lock-in CCD image sensor offering* 5×12 *pixels. Each pixel is provided with eight taps (sampling points per period), with which higher-order moments and Fourier coefficients can be determined from temporally modulated wave fields. A prime application of the lock-in imager is in range imaging without moving parts based on the time-of-flight distance measurement principle. Chip size is around* $2.4\times 2.4\,\mathrm{mm}^2$. ***b*** *Chip micrograph of a single-chip digital camera with* 256×256 *pixels, fast analog-to-digital converters for 60 frames/s conversion rate, all analog and digital electronics for the timing generation, autoexposure, subsampling logic and switchable linear/logarithmic pixel sensitivity. Chip size is around* $5\times 5\,\mathrm{mm}^2$.

taps of a convolution kernel is around 1 % of the largest value, which is sufficient for most applications in image processing and machine vision.

While traditional image sensors have relied on sequential scanning of the individual pixels for readout with a single output amplifier, many types of CCD image sensors developed in the past few years offer several output amplifiers working in parallel. Such multitap image sensors offer a much increased frame rate, albeit at the cost of increased complexity of the external image data acquisition circuitry. It is also possible to preprocess the image data on-chip, by making use of other parallelisms. Analog circuitry, such as comparators, differentiators, maximum finders, etc. can be combined in each pixel or for each column with digital circuitry for controlling the data acquisition and preprocessing functions. Analog-to-digital converters of various precisions can be integrated with each pixel or—with improved performance—they are integrated for each column. An example of this capability is demonstrated with the digital single-chip camera shown in Fig. 7.30b. On one chip, a 256×256 photodiode array is combined with fast analog-to-digital converters, all analog and digital electronics for the timing generation, autoexposure and subsampling logic. Requiring a single 3.3 V supply voltage and a clock signal, this digital camera-on-a-chip produces 60

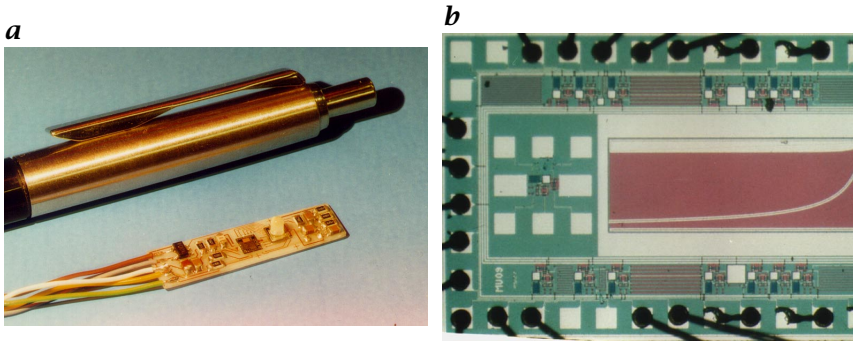


Figure 7.31: ***a*** *Example of a miniaturized, pen-sized video camera. It is realized with a low-power CMOS imager, on top of which a minilens imaging system is placed. The imager is programmed and driven by a single-chip microcontroller right next to it.* ***b*** *Chip micrograph of a photosensor with nonlinear spatial sensitivity, realizing a 1/x function. This photosensor is used in a planar optical distance sensor based on a triangulation setup, offering a distance resolution of 1 %. Chip size is around* 1.5×2 *mm*^{2}.

digital images per second with 10 bits per pixel, while consuming only 10 mW. An additional feature is the switchable sensitivity behavior of the pixels: in one mode, pixels show linear sensitivity with a dynamic range of around 65 dB. In another mode, the pixels exhibit logarithmic sensitivity as described in Section 7.3.5 and in Chapter 8, with a dynamic range exceeding 100 dB. One can switch very fast between the modes, from one frame to the next if necessary.

Obviously, such camera chips are not only the basis of low-cost camera systems for many consumer applications, but such cameras lend themselves quite well to extreme miniaturization. In Fig. 7.31a, a pen camera is shown, consisting of a CMOS image sensor with microlens imaging system on top, controlled by a microcontroller chip. The complete camera can be mounted in the upper third of a conventional pen, with much potential for further miniaturization.

7.10.2 Smart image sensors

As described in the preceding, the individual pixels of a modern, custom-designed image sensor can contain a wide variety of analog and digital circuitry, giving the pixel astonishing levels of functionality. Such "smart pixels" profit directly from the on-going development in semiconductor technology, because the shrinkage of design rules translates directly into more functionality per area in a pixel of a given size. At the same time, analog processing circuits and digital signal processing modules can be integrated monolithically on the image sensor chip, leading to what is called "smart image sensors."

An obvious application is the integration of all components of a video camera on one single chip, as described, for example, in Renshaw et al. [56]. Such single-chip video cameras are commercially available now. Recently, an improved version of a single-chip digital camera has been advertised, combining a 160×160 photodiode pixel array, autoexposure circuitry, all necessary analog and digital control/timing electronics, as well as an on-chip A/D converter with interface to processor-compatible serial and parallel ports. Volume price for such a single-chip digital camera is around $10, making it very attractive for many practical applications such as surveillance, automatic manufacturing, process control, picture telephony, etc. It is not difficult to imagine that the next step can be taken as well, that is, the cointegration of such an electronic camera with a general-purpose digital processor, capable of evaluating the acquired imagery directly on chip. Such camera-processor products, either based on line or area cameras, are already commercially available now [57], with the first successful industrial applications, primarily in automatic manufacturing and process control. Consumer applications with on-chip image compression, on-chip modem for image transmission etc. have also been addressed, and it is expected that such low-cost camera systems find applications in many security and safety applications at home and in public places.

Various types of smart image sensors for different approaches to range imaging have been realized. For variations of the well-established triangulation distance-measuring technique (see Sections 18.4 and 19.4), smart imagers exhibiting the following properties have been described: A predefined 1-D spatial response can be obtained with suitably shaped photosensors, exhibiting for example 1/x spatial sensitivity characteristics, as illustrated in Fig. 7.31b. The resulting planar triangulation distance sensor achieves a distance reproducibility of around 1 % [58]. This simple solution saves a digital processor that would have been necessary for the calculation of this nonlinear transformation, and a robust, easy-to-design single-chip distance sensor result. A 2-D array of "time-to-maximum-light pixels" is the basis of another triangulation setup with swept sheets of light [59]. Stereodepth vision can also be considered to be a (passive) triangulation technique (see also Section 20.2, for which special high-speed stereodepth vision chips have been proposed, see, for example, Hakkarainen et al. [60].

The few examples given here should serve as an indication that machine vision can profit enormously from the developments in the field of smart image sensing. They make it possible to miniaturize, improve, or extend known measurement techniques, while at the same time often reducing the cost and increasing the performance of the system.

7.10.3 On the way to seeing chips?

The rapid development of image sensors with more and more integrated functionality led a prominent researcher in the field to proclaim the imminence of "seeing chips" [61]. A few examples of image sensors with complete, integrated image processing hardware have been reported for certain tasks, such as the fingerprint recognition and identification chip described in Denyer et al. [57]. Various successful smart image sensors have been demonstrated that are capable of carrying out certain important, but still only basic functions for the vision process on a single chip, see, for example, Koch [62]. The suspicion that "vision is difficult" [61] has been fully verified, and it has become obvious that the early expectations of monolithically integrated single-chip vision systems were too high. As demonstrated for example by the fingerprint verification chip [57] it is possible today to co-integrate an image sensor and all the necessary processing circuitry on a single chip for the solution of a given—still not too complex—machine vision problem. However, this would be far removed from the original idea of a seeing chip which visually perceives some aspects of its surroundings, and in most cases it would make no economical sense.

The basic philosophy behind the seeing chip is to distribute the processing power over the photosensitve part. This strategy is inspired by the biological concept of highly parallel, low-speed and low power distributed analog computing, which is the basis of nature's marvelous visual perceptive systems, such as our own highly-developed sense of vision. In contrast to the planar, essentially two-dimensional semiconductor fabrication technology, nature realizes fully three-dimensional processing systems, in which each "pixel" is backed by a tremendous number of nerve cells—more than 10^5 in the human visual system [63]—performing the necessary calculation for the sense of vision. In the near future, it will be unrealistic to expect that each pixel on a solid-state image sensor will contain more than a few ten transistors, while maintaining a useful pixel size of the order of $30 \times 30\,\mu\text{m}^2$ and an optical fill factor of at least 10 %.

As a consequence, recent developments in the area of integrated machine vision also consider architectures based on different planes: an image acquisition plane might be followed by several (analog) preprocessing planes, an (essentially digital) classification plane and an output plane, all connected using suitable high-bandwidth bus schemes with an appropriate software protocol. This guarantees a maximum fill factor for the image sensing part and allows for the use of optimal architectures and technologies for the different parts of the complete system. Such an approach does not necessarily mean that every plane resides on its own chip; different planes can be integrated on the same chip. The technology for stacking and interconnecting silicon chips,

so called 3-D or z-plane technology, has been developed [64], but the appealing idea of a low-cost single-chip vision system, a seeing chip, becomes seriously compromised.

The conclusion is that smart image sensors (offering additional on-chip functionality) and integrated vision systems are certainly trends that will lead to a wide range of practical products, albeit rarely in the form of single, self-contained seeing chips. Instead, it can be expected that smart image sensors with extended capabilities for the dynamic acquisition of images will be part of an integrated vision system. This will consist of an economically sensible combination of imager, analog and digital processing parts. Special properties built into such smart image sensors include lower noise, higher DR, programmable sensitivity, on-chip nonuniformity and shading correction, variable exposure and timing control, region-of-interest capability, dynamic pixel size and shape, and on-chip image preprocessing, which can be carried out for all pixels in parallel, etc. It might well be that "seeing chip" is a misnomer, and that the silicon retina [65], with its less exaggerated expectations and the suggestion of more of a front-end image acquisition/pre-processing module, is a much more appropriate name for the current and future development directions in the field of integrated image acquisition and processing systems.

7.11 Conclusions

It was only about a decade ago that a few researchers started to exploit one of the most exciting capabilities offered by modern silicon-based semiconductor technology, the monolithic integration of photosensitive, analog and digital circuits. Some of the results of these efforts are described in this work, representing just a small fraction of the many applications already demonstrated. They all support the main assertion of this chapter, that today's image sensors are no longer restricted to the acquisition of optical scenes. Image sensors can be supplied with custom integrated functionality, making them key components, application-specific for many types of optical measurement problems. It was argued that it is not always optimal to add the desired custom functionality in the form of highly-complex smart pixels, because an increase in functionality is often coupled with a larger fraction of a pixel's area being used for electronic circuit, at the cost of reduced light sensitivity. For this reason, each new optical measurement problem has to be inspected carefully, taking into account technical and economical issues. For optimum system solutions, not only smart pixels have to be considered. Functionality could also be provided by separate on-chip or off-chip circuits, perhaps by using commercially available electronic components.

Machine vision system architects can no longer ignore the freedom and functionality offered by smart image sensors, while being well aware of the shortcomings of semiconductor photosensing. It may be true that the seeing chips continue to be elusive for quite some time. The smart photosensor toolbox for custom imagers is a reality today, and a multitude of applications in optical metrology, machine vision, and electronic photography can profit from the exciting developments in this area. "Active vision," "integrated machine vision," "electronic eyes," and "artificial retinae" are quickly becoming more than concepts: the technology for their realization is finally here now!

7.12 References

[1] Gonzalez, R. and Wintz, P., (1987). *Digital Image Processing, 2nd edition*. Reading, MA: Addison-Wesley.

[2] Beck, R. (ed.), (1995). *Proc. AAAS Seminar on Fundamental Issues of Imaging Science, Atlanta (GA), February 16-17, 1995*.

[3] Beyer, H., (1992). *Geometric and radiometric analysis for a CCD-camera based photogrammetric close-range system*. PhD thesis No. ETH-9701, Federal Institute of Technology, Zurich, Switzerland.

[4] Chamberlain, S. and Lee, J., (1984). A novel wide dynamic range silicon photodetector and linear imaging array. *IEEE Jour. Solid State Circ.*, **SC-19**: 175–182.

[5] Lenz, R., (1996). *Ein Verfahren zur Schätzung der Parameter geometrischer Bildtransformationen*. Dissertation, Technical University of Munich, Munich, Germany.

[6] Schenker, P. (ed.), (1990). *Conference on Active Vision*, Vol. 1198 of *Proc. SPIE*.

[7] Seitz, P., (1995). Smart image sensors: An emerging key technology for advanced optical measurement and microsystems. In *Proc. SPIE*, Vol. 2783, pp. 244–255.

[8] Saleh, B. and Teich, M., (1991). *Fundamentals of Photonics*. New York: John Wiley and Sons, Inc.

[9] Wong, H., (1996). Technology and device scaling considerations for CMOS imagers. *IEEE Trans. El. Dev.*, **43**:2131–2142.

[10] Sze, S., (1985). *Semiconductor Devices*. New York: John Wiley and Sons.

[11] Spirig, T., (1997). *Smart CCD/CMOS based image sensors with programmable, real-time temporal and spatial convolution capabilities for applications in machine vision and optical metrology*. PhD thesis No. ETH-11993, Federal Institute of Technology, Zurich, Switzerland.

[12] Heath, R., (1972). Application of high-resolution solid-state detectors for X-ray spectrometry—a review. *Advan. X-Ray Anal.*, **15**:1–35.

[13] Bertin, E., (1975). *Principles and Practice of X-Ray Spectrometric Analysis*. New York: Plenum Press.

[14] Budde, W., (1979). Multidecade linearity measurements on Si photodiodes. *Applied Optics*, **18**:1555–1558.

[15] Theuwissen, A., (1995). *Solid-State Imaging with Charge-Coupled Devices*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

[16] Vietze, O. and Seitz, P., (1996). Image sensing with programmable offset pixels for increased dynamic range of more than 150 dB. In *Conference on Solid State Sensor Arrays and CCD Cameras, Jan. 28–Feb. 2, 1996, San Jose, CA*, Vol. 2654A, pp. 93–98.

[17] Vietze, O., (1997). *Active pixel image sensors with application specific performance based on standard silicon CMOS processes*. PhD thesis No. ETH-12038, Federal Institute of Technology, Zurich, Switzerland.

[18] Webb, P., McIntyre, R., and Conradi, J., (1974). Properties of Avalanche Photodiodes. *RCA Review*, **35**:234–277.

[19] Seitz, P., (1997). Image sensing with maximum sensitivity using industrial CMOS technology. In *Conference on Micro-Optical Technologies for Measurement Sensors and Microsystems II, June 16–June 20, 1997, Munich, Germany*, Vol. 3099, pp. 22–33.

[20] Zappa, F., Lacatia, A., Cova, S., and Lovati, P., (1996). Solid-state single-photon detectors. *Optical Engineering*, **35**:938–945.

[21] Mathewson, A., (1995). *Integrated avalanche photo diode arrays*. Ph.D. thesis, National Microelectronics Research Centre, University College, Cork, Ireland.

[22] Mahowald, M., (1991). Silicon retina with adaptive photodetectors. In *Conference on Visual Information Processing: From Neurons to Chips Jan. 4, 1991, Orlando, FL*, Vol. 1473, pp. 52–58.

[23] Graf, H., Höfflinger, B., Seger, Z., and Siggelkow, A., (1995). Elektronisch Sehen. *Elektronik*, **3**:3–7.

[24] Sankaranarayanan, L., Hoekstra, W., Heldens, L., and Kokshoorn, A., (1991). 1 GHz CCD transient detector. In *International Electron Devices Meeting 1991*, Vol. 37, pp. 179–182.

[25] Colbeth, R. and LaRue, R., (1993). A CCD frequency prescaler for broadband applications. *IEEE J. Solid-State Circ.*, **28**:922–930.

[26] Carnes, J. and Kosonocky, W., (1972). Noise sources in charge-coupled devices. *RCA Review*, **33**:327–343.

[27] Allen, P. and Holberg, D., (1987). *CMOS Analog Circuit Design*. Fort Worth: Saunders College Publishing.

[28] Hopkinson, G. and Lumb, H., (1982). Noise reduction techniques for CCD image sensors. *J. Phys. E: Sci. Instrum*, **15**:1214–1222.

[29] Knop, K. and Seitz, P., (1996). Image Sensors. In *Sensors Update*, W. G. Baltes, H. and J. Hesse, eds., pp. 85–103. Weinheim, Germany: VCH-Verlagsgesellschaft.

[30] Chandler, C., Bredthauer, R., Janesick, J., Westphal, J., and Gunn, J., (1990). Sub-electron noise charge coupled devices. In *Conference on Charge-Coupled Devices and Solid State Optical Sensors, Feb. 12–Feb. 14, 1990, Santa Clara, CA*, Vol. 1242, pp. 238–251.

[31] Janesick, J., Elliott, T., Dingizian, A., Bredthauer, R., Chandler, C., Westphal, J., and Gunn, J., (1990). New advancements in charge-coupled device technology. Sub-electron noise and 4096 X 4096 pixel CCDs. In *Confer-*

ence on Charge-Coupled Devices and Solid State Optical Sensors, Feb. 12–Feb. 14, 1990, Santa Clara, CA, Vol. 1242, pp. 223–237.

[32] Matsunaga, Y., Yamashita, H., and Ohsawa, S., (1991). A highly sensitive on-chip charge detector for CCD area image sensor. *IEEE J. Solid State Circ.*, **26**:652–656.

[33] Fossum, E., (1993). Active pixel sensors (APS)—are CCDs dinosaurs? In *Conference on Charge-Coupled Devices and Solid-State Optical Sensors III, Jan. 31–Feb. 2, 1993, San Jose, CA*, Vol. 1900, pp. 2–14.

[34] Mendis, S., Kemeny, S., Gee, R., Pain, B., Staller, C., Kim, Q., and Fossum, E., (1997). CMOS active pixel image sensors for highly integrated imaging systems. *IEEE J. Solid-State Circ.*, **32**:187–197.

[35] Pratt, W., (1991). *Digital Image Processing,* 2nd edition. New York: Wiley.

[36] Benson, K., (1986). *Television Engineering Handbook.* New York: McGraw Hill.

[37] Ramo, S., Whinnery, J. R., and van Duzer, T., (1994). *Fields and waves in communication electronics,* 3rd edition. New York: Wiley.

[38] Jackson, J. D., (1975). *Classical Electrodynamics,* 2nd edition. New York: Wiley.

[39] Gagnaire, M., (1997). An overview of broad-band access technologies. *Proc. IEEE*, **85**:1958–1972.

[40] Pritchard, D. H., (1984). U.S. color television fundamentals — a review. *RCA Engineer*, **29**:15–26.

[41] Hunt, R. W. G., (1991). *Measuring Colour,* 2nd edition. Ellis Horwood.

[42] Bayer, B. E., (1976). Color imaging array, U.S. patent No. 3,971,065.

[43] Knop, K., (1985). Two-dimensional color encoding patterns for use in single chip cameras. *Proc. SPIE*, **594**:283–286.

[44] Aschwanden, F., Gale, M. T., Kieffer, P., and Knop, K., (1985). Single-chip color camera using a frame-transfer CCD. *IEEE Trans. Electron. Devices*, **ED-32**:1396–1401.

[45] Arnold, L., (1992). *Moderne Bildkomminikation.* Heidelberg: Hüthig Verlag.

[46] Davis, A. W., (1997). Where the cameras will fit in. *Advanced Imaging*, **Nov. 97**:43–49.

[47] Kreider, G., Bosiers, J., Dillen, B., van der Heijden, J., Hoekstra, W., Kleinmann, A., Opmeer, P., Oppers, J., Peek, H., Pellens, R., and Theuwissen, A., (1995). An mK $\times$ mK Modular Image Sensor Design. In *International Electron Devices Meeting 1995, Washington, D. C.*, pp. 155–158.

[48] Peek, H. L., Verbugt, D. W., Beenhakkers, M. J., Huininck, W. F., and Kleimann, A. C., (1996). An FT-CCD imager with true $2.4 \times 2.4\,\mu\text{m}^2$ pixels in double mebrane poly-Si technology. In *Proceedings of the IEDM '96, International Electron Devices Meeting, San Francisco, Dec.. 8-Dec. 11, 1996*, pp. 907–910.

[49] MOSIS, (1999). MOSIS VLSI fabrication service, Information Sciences Institute, University of Southern California, USA, Marina del Rey, CA 90292-6695; http://www.mosis.org/.

[50] Flores, J., (1992). An analytical depletion-mode MOSFET model for analysis of CCD output characteristics. In *Conference on High-Resolution Sensors and Hybrid Systems, Feb. 9–Feb. 14, 1992, San Jose, CA*, Vol. 1656, pp. 466–475.

[51] Raynor, J. and Seitz, P., (1997). A linear array of photodetectors with wide dynamic range and near photon quantum noise limit. *Sensors and Actuators A*, **61**:327–330.

[52] Engelhardt, K. and Seitz, P., (1996). Absolute, high-resolution optical position encoder. *Applied Opics*, **35**:201–208.

[53] Seitz, P., Leipold, D., Kramer, J., and Raynor, J. M., (1993). Smart optical and image sensors fabricated with industrial CMOS/CCD semiconductor processes. *Proc. SPIE*, **1900**:21–30.

[54] Spirig, T., Seitz, P., Vietze, O., and Heitger, F., (1995). The lock-in CCD. Two-dimensional synchronous detection of light. *IEEE J. Quantum Electronics*, **31**:1705–1708.

[55] Spirig, T., Seitz, P., Vietze, O., and Heitger, F., (1997). A smart CCD image sensor with real-time programmable parallel convolution capabilities. *IEEE Trans. Circuits and Systems*, **44**:465–468.

[56] Renshaw, D., Denyer, P., Wang, G., and Lu, M., (1990). ASIC vision. In *Proc. of the IEEE 1990 Custom Integrated Circuits Conference, Feb. 14–Feb. 16, 1990, Dan Francisco, CA*, pp. 7.3.1–7.3.4.

[57] Denyer, P., Renshaw, D., and Smith, S., (1995). Intelligent CMOS imaging. In *Conference on Charge-Coupled Devices and Solid-State Optical Sensors V, Feb. 5–Feb. 10, 1995, San Jose, CA*, Vol. 2415, pp. 285–291.

[58] Kramer, J., Seitz, P., and Baltes, H., (1994). Planar distance and velocity sensor. *IEEE Jour. Quantum Electronics*, **30**:2726–2730.

[59] Gruss, A., Carley, L., and Kanade, T., (1991). Integrated sensor and range-finding analog signal processor. *IEEE J. Solid State Circ.*, **26**:184–192.

[60] Hakkarainen, J., Little, J., Lee, H., and Wyatt, J., (1991). Interaction of algorithm and implementation for analog VLSI stereo vision. In *Conference on Visual Information Processing: From Neurons to Chips, Jan. 4, 1991, Orlando, FL*, Vol. 1473, pp. 173–184.

[61] Koch, C., (1989). Seeing chips: analog VLSI circuits for computer vision. *Neural Computation*, **1**:184–200.

[62] Koch, C., (1991). Implementing early vision algorithms in analog hardware—an overview. In *Conference on Visual Information Processing: From Neurons to Chips, Jan. 4, 1991, Orlando, FL*, Vol. 1473, pp. 2–15.

[63] Hubel, D., (1988). *Eye, Brain and Vision*. New York: Scientific American Library.

[64] Carson, J. (ed.), (1989). *Materials, devices, techniques and applications for Z-plane focal plane array technology*, Vol. 1097 of *Proc. SPIE*.

[65] Mahowald, M. and Mead, C., (1991). The silicon retina. *Scientific American*, **264**:40–46.

8 HDRC-Imagers for Natural Visual Perception

Ulrich Seger, Uwe Apel, and Bernd Höfflinger

Institut für Mikroelektronik, Stuttgart, Germany

8.1 Introduction

In the development of electronic cameras, human perception of scenes has been the measure for the camera quality.

While in acoustic signal transmission a level of high fidelity was reached in the late 1960s with logarithmic compression technique, it took nearly 30 yr to recognize that electronic imaging as well as machine vision could benefit from mimicking human visual perception with nonlinear rather than linear image sensors. With the advent of million-transistor VLSI-chips, nonlinear active-pixel imagers have been realized in recent years, among them the logarithmic *high-dynamic range CMOS* (HDRC)-imager and digital cameras, which seem to come close to a high fidelity electronic imaging system. Principles, examples, and trends for logarithmic, high-fidelity image acquisition including innovative color vision approaches are presented.

Because image recovery is seen as a medium to facilitate both documentation and communication, the goal of imaging has been to provide as much detailed information as an individual observer could get in the life scene.

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

For Michelangelo or Rembrandt, their wish was to preserve the beauty of a moment, a person, or a landscape by creating an image that would last forever.

The development of *photography* in the last two centuries allowed a huge advance in that area and today still represents the state of the art in imaging. So far, no electronic system comes close to the photography standard and even the best film materials fall short in situations where the dynamic range is very high.

8.2 Log compression at the pixel site

In the late 1960s and early 1970s Tomita [1] and Cornsweet [2] showed that the *responsivity* of cone receptors has a logarithmic characteristic. Eye-like *logarithmic compression* overcomes most of the dynamic range-related restrictions. Different approaches of log compression have been examined [3] but technological variability of critical parameters foiled the manufacture of large area sensors. Advances in MOS technology and circuit concepts as well as new system structures allow these problems to be overcome today.

To understand the advantage of log signal compression, it is necessary to take a close look at what the signal or information is. The information in images is (with only a few exceptions) the contrast [4, 5]. Intensities in most images, with the possible exception of light sources themselves, are the product of irradiation and the reflectivity/absorptivity of imaged objects as well as the absorptivity of the medium that is in between the light source, the object and the imager. In situations where scattering and light transmission effects can be neglected, the intensities within an image are the product of the *irradiance* E and the *reflectance* ρ of the imaged surface.

Technical surfaces show reflectance values between 5 % (black surface, nearly absorbing all light energy) and 95 % (bright shining surface nearly reflecting all light energy seen at a short distance). So far a *dynamic range* of 20 : 1 is necessary for all the variations within this scenery to have a *signal-to-noise ratio* (SNR) of at least one. High-dynamic range requirements arise from higher SNR demands and are caused by varying illumination. In fact, illumination can vary not only by factors of tens of thousands between moonlight and bright sunshine but also between a bright headlamp and shadowed regions by night and even between bright direct reflections of the sun and shadowed regions under daylight conditions. With technical light sources, apertures and optical filters, it is also possible to generate illumination variations in a scene that span several decades.

The response S of a linear converting device is $S \approx E\rho$. With a (HDRC) sensor the response will be $S \approx \log E + \log \rho$. Two different

Figure 8.1: *HDRC image of* ***a*** *left side;* ***b*** *top;* ***c*** *right side; illuminated newspaper reader.*

Figure 8.2: *CCD image of* ***a*** *left side;* ***b*** *top;* ***c*** *right side; illuminated newspaper reader, showing saturation effects typical for linear responding cameras.*

regions on a surface representing different reflectivities ρ_1 and ρ_2 with a difference of $\Delta\rho$ will produce an output signal different by ΔS. In the linear converting case with $S \approx E\rho$, you will notice a dependency of S on the irradiation E.

In the HDRC case, the signal difference is independent of the absolute irradiation level because $\Delta S \approx \log \rho_1 - \log \rho_2$. For example, look at the newspaper reader under different illumination conditions in Fig. 8.1. Spot lights from the left side, the right side, or from the top are used to illuminate the scenery.

As the reflectance difference of printed letters and paper remain the same, the intelligibility of this "information" is always present in the log image, and changes in illumination translate to an offset in the sensor output that is defined by the log of the illuminance.

Let us assume a black printed character with $\Delta\rho = 5\,\%$ and the white paper in its background with $\Delta\rho = 80\,\%$. Effectively a difference in the reflectance of two areas is represented by a fixed number of intermediate levels (supposing 15 levels) regardless of whether the paper is illuminated by moonlight or bright sunlight. This applies for logarithmic conversion only, whereas with linear converting sensors different numbers of gray levels will result for different illuminations. Hence, under bright illumination, one will be able to distinguish up to 250 gray levels (unless the sensor saturates) between the character and its background while under low illumination, one can hardly distinguish between characters and the background. Normally, this problem is overcome by changing the aperture or integration time settings of the camera and very often this requires extra signal-processing power for

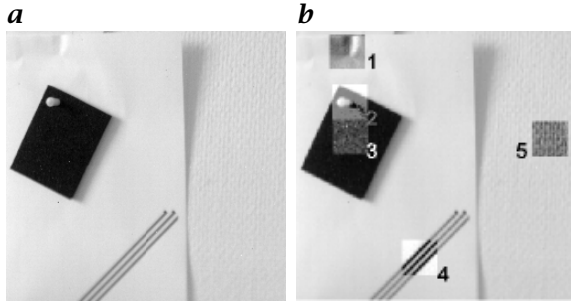


Figure 8.3: *Demonstration of information contents of a log image:* ***a*** *adapted to max. achievable print media dynamic;* ***b*** *five selected portions locally normalized for maximum visibility.*

dynamic thresholding. As can be seen with the images in Fig. 8.2, sometimes the information will not be recoverable at all due to saturation effects.

The potential strength of contrast resolution can best be demonstrated with scenes that show only slight contrast in different areas but a high overall dynamic range.

The pictures in Fig. 8.3 show a sheet of paper on a white wall, with a small part of black foam material pinned on it. The original image contains only 20-30 gray levels in the “white region” and 20–30 gray levels in the “dark region.”

To demonstrate the data content of the HDRC logarithmic image with a depth of only 8 bits, local expansion in five areas has been performed as a postprocessing step. This leads to improved visibility in both dim (2, 3) and bright (1, 5) regions. Notice the structure of the wall paper that can be seen in area 5. Visibility of wrinkles in the paper (see region 1) as well as fine structures of the foam are enhanced by linear operation on the gray values of the original image on the left. High amplification in the dark region does not result in noisy data known from linear sensors. Rather, it allows detection of the black shadow of a pin on black foam. This is information that is within only two gray levels of difference (representing a 3 % difference in brightness)—a difference not even resolvable by a human observer.

In scenes where the dynamic range is in the order of 10.000 : 1 and higher, the dynamic range of any linear operating sensor is exceeded. This happens easily in scenes with light sources or other light emitting objects.

The imaging of a light bulb with its filament and with details beside or behind it is shown in Fig. 8.4. This was shown for the first time in 1993 using the first generation of HDRC-imagers and has become a symbol for high-dynamic range imaging. Figure 8.4a shows the OSRAM

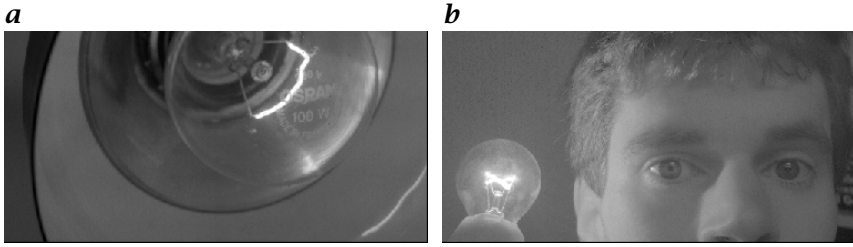


Figure 8.4: *HDRC images of* ***a*** *fully powered 100 W light bulb;* ***b*** *portrait under "impossible" conditions.*

Figure 8.5: *HDRC2 Traffic scenes* ***a*** *in bright sunlight;* ***b*** *at dawn;* ***c*** *by night.*

Figure 8.6: *Effect of different data representations of the same data contents.*

100-W lightbulb fully powered with details both in filament, socket, and even in the background of the lamp. Figure 8.4b shows a portrait that could not be taken with any linear responding system without using multiple exposure techniques.

In Fig. 8.5, the images show the use of HDRC imagers without shutter control or integration time adaptation. Images were taken using a HDRC2-EC. Figure 8.5a was taken at noontime in bright sunshine, Fig. 8.5b was taken at dawn, and Fig. 8.5c was taken at night. Although all images are normalized to minimum and maximum value, not all of the data content can be printed.

Further details of images are present in the digital data, however, printing media cannot provide the dynamic range necessary to represent all the content within one image.

Another representation of the night scene as depicted in Fig. 8.5c is shown in Fig. 8.6. The picture has been modified using a gamma correction with $\gamma = 1.1$ in Fig. 8.6a, $\gamma = 1.2$ in Fig. 8.6b, and linear stretch in Fig. 8.6c.

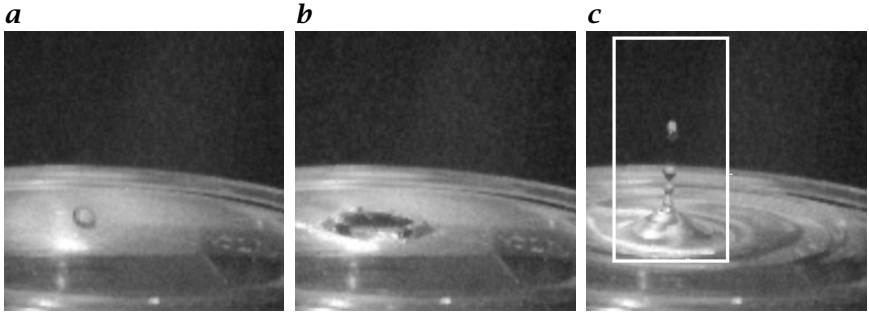


Figure 8.7: *High-speed imaging of 3 images out of 200 images within 1 s.* ***a*** *At t=0;* ***b*** *at t + 5 ms;* ***c*** *at t + 15 ms marked subframe can be imaged with* $>$ 3000 *frames/s, featuring an additional 15 images between the shown images.*

Where illumination control is not possible and illumination conditions are not predictable (as in natural scenes or in some industrial applications), sensors with a high "intrascene dynamic range" are a prerequisite for successful image processing. A safe margin in a detectable dynamic range simplifies applications and will extend the application field for image processing systems. Such systems promise a solution for fast changing unpredictable illumination situations.

8.3 Random pixel access

Active pixel structures allow random access, which in combination with the forementioned fast self-setting bandwidth control allows high-speed imaging of a new dimension. The falling water droplet is imaged with 200 fps, featuring an image every 5 ms. Reducing the image field to the marked subframe, which contains 32×64 pixel only, another 69 images of the falling droplet can be imaged in between each of the foregoing images. Tracking operations with different windows and subframe frequencies are also possible.

8.4 Optimized SNR by bandwidth control per pixel

There are two ways to achieve a log transfer characteristic. One is to implement a log amplifier succeeding a linear-responding sensor element, the other is to attenuate a linear system response by an exponentially increasing attenuation. Using log amplification with feedback, the signal bandwidth remains constant while the amplification factor is variable. A small photo current causes high amplification factors adding noise to the signal. A small photo current affects high gain settings

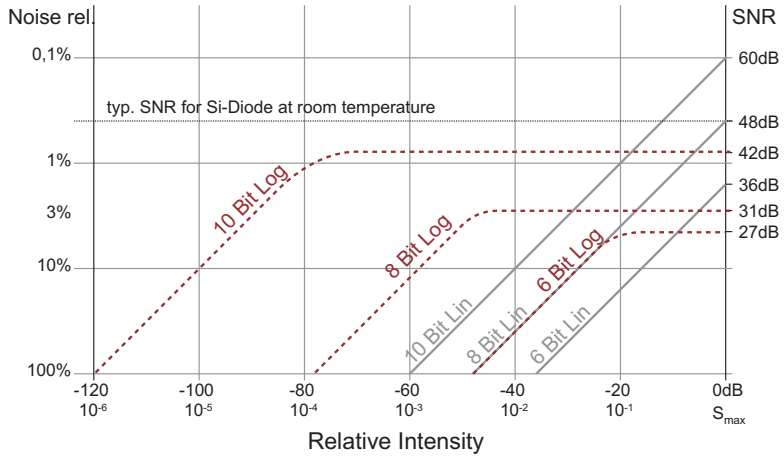


Figure 8.8: *The SNR of real systems using log (dotted line) and linear (solid line) sensors.*

resulting in considerably higher noise. Using the principle of exponential attenuation, the bandwidth is variable. The attenuation factor is increased the higher the photosignal is. The signal bandwidth of each pixel depends on the actual illumination. A small photo current results in a small bandwidth, while a high photocurrent results in a large bandwidth. Thus a high noise potential is accompanied by a low signal bandwidth and a low noise potential is paired with a high bandwidth.

Practically, this means that high signal frequencies are detected at bright illumination while poor illumination results in a low-pass filtering in each pixel. Log amplification inserts an amplification noise proportional to the amplification factor. In contrast, exponential attenuation results in a constant *signal-to-noise ratio* (SNR) over most of the entire operational range because the noise is proportional to $\sqrt{\Delta f}$:

$$\begin{array}{lll} \text{Shot noise} & I_{\text{shot noise}} & = \sqrt{2qI\Delta f} \quad \text{and} \\ \text{Johnson noise} & V_{\text{johnson, rms}} & = \sqrt{4kTR\Delta f} \end{array} \tag{8.1}$$

(For a detailed discussion on noise of CMOS image sensors, see Section 7.5.3.) The resulting SNR is depicted in Fig. 8.8. In the final mathematical description of the form $U_{\text{out}} = U_a + U_b \log(I\phi/I_{\text{dark}})$, it is no longer recognizable which type the sensor is; however, the difference in the SNR is significant as illustrated in Fig. 8.8. The decrease of the SNR in the lower region of the operational range results from the quantization noise in the A/D conversion.

Any other approach in log conversion of images after image sensing and amplification (e. g., by log A/D converters or digital lin-log convert-

ers) also results in the known advantages of high efficient data coding and constant contrast steps (digilog) but can not overcome problems arising from saturation effects in the light sensing or signal amplification process.

8.5 Data density in the log space

In the log domain, one digitization or gray level represents no longer an absolute intensity step but an increase by a constant contrast step or, in other words, the increase by a fixed (percentage) multiplication factor. Each digitized intensity step corresponds to $I_n = I_{n-1}(1 + C)$ where C is the contrast resolution of the imager. (C might reach values of 0.015 or 1.5%.) The resulting *dynamic range* D/R is calculated for a 10-bit system as $D/R = (1 + C)^n$, with $n = 2^{10}$. With 1.5% contrast resolution, the *dynamic range* D/R is 4.180.490 : 1.

This leads to an inherent information coding and allows processing of images with lower bandwidth requirements. Histograms of real images are given in the upper right corners of Fig. 8.9. In this case the setting of the camera resulted in a value of $C = 9\%$. With 64 gray levels, a dynamic range of $1,09^{64} \approx 250 : 1$ is covered. The linear system used 250 gray levels to represent the same scene. Despite a data compression by a factor of 4, the logarithmically compressed image shows a better visibility of details.

The example in Fig. 8.9 shows a comparison of a standard road scene between a linear converting CCD-Camera (lower image) and the logarithmic responding HDRC-Camera (upper image).

8.6 Color constancy in the log space

Figure 8.10 shows log scaled digital color images with an illumination difference of 8 f-stops (a linear factor of $2^8 = 256$). With a given dynamic range of 1:40 in the reflectivity of the chart this produces a total dynamic range of approximately $40 \cdot 256 \geq 10,000$ to be imaged correctly. Using AGC or adaptive background suppression, this dynamic range might be covered in succeeding recordings using different integration time or aperture settings; however, the same dynamic range may be required within one scene (e. g., caused by a shadowed region).

The McBeth color checker card has been taken with different aperture settings; the result is shown Fig. 8.10a. The numbers at the corners indicate the lens stop used. Using f-stop 16 gives only a very poor illumination on the sensor, which results in images close to the noise floor. Despite poor illumination, a linear operation is sufficient to restore the correct color representation (see Fig. 8.10b). Using sensors with nonconstant SNR will result in heavy color noise.

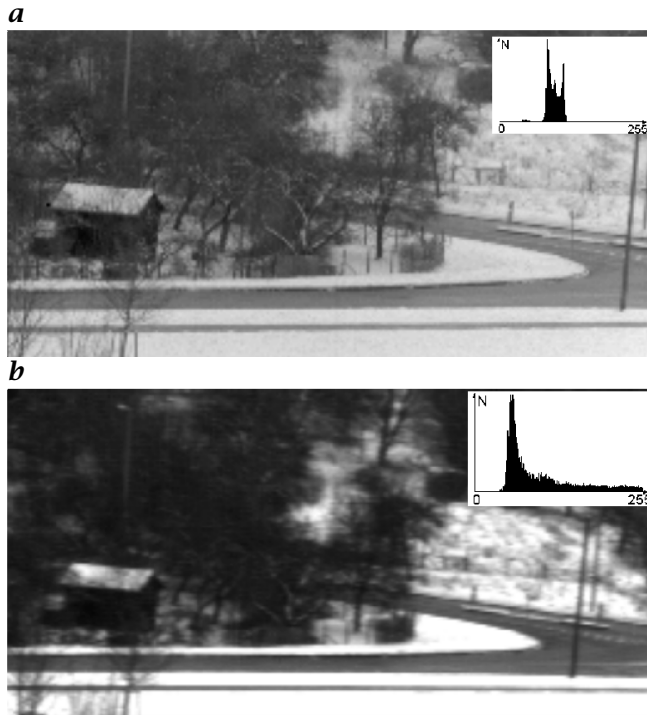


Figure 8.9: *Histogram of the road scene with* ***a*** *HDRC camera;* ***b*** *CCD camera.*

For comparison, Fig. 8.11 shows a McBeth color chart that was synthetically generated.

8.7 Development of functionality and spatial resolution

For a few application areas (such as video phone or videoconferencing), the signal bandwidth for information transfer is rather limited. In those cases an image size of CIF format ($288\text{v} \times 352\text{h}$) is a well-adapted spatial resolution. In contrast, for most other applications proper identification of details in the image is required, for example, traffic signs to be interpreted by driver assistants, mechanical parts to be handled by robots. At the upper end the resolution of fine structures in electronic still video images has to compete with the classical silver halogenide technology. While for the acquisition of video scenes for surveillance applications or industrial automation the VGA resolution will be sufficient, pixel arrays with a size of up to $2\text{k} \times 3\text{k}$ will be required for electronic still video images.

Increasing the numbers of rows and columns while maintaining an image diagonal compatible to a cost-saving lens limits pixel pitch. On

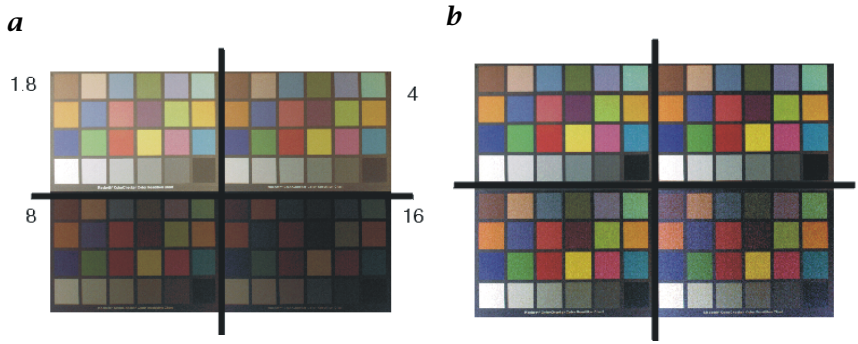


Figure 8.10: ***a*** *Log image of McBeth chart with f-stops as indicated;* ***b*** *same as* ***a*** *but normalized to black and white for each quarter; (see also Plate 1).*

Figure 8.11: *Noiseless McBeth chart generated synthetically for comparison with Fig. 8.10; (see also Plate 2).*

the one hand, the continuous shrinking of structure dimensions in modern CMOS processes supports this endeavor. On the other hand, the improvement of performance criteria of large HDRC sensor arrays (such as readout speed or homogeneity) requires a certain amount of circuitry footprint in each pixel. To retain a *fill factor* of about 40%, a percentage sufficient for avoidance of aliasing effects in the image, the *pixel size* has a lower limit, which is a factor 35 up to 40 above the minimum feature size of the used CMOS process. On the other hand, total chip size is bound by lithography exposure fields and process yield to a chip diagonal of 25 mm [6]. As can be seen in Fig. 8.12, an HDRC sensor with VGA resolution can be fabricated with an economic chip with a diagonal below 10 mm.

The chip diagonal influences the system costs directly, especially as regards chip and optics prices together. Reflections and scattering within the optics have to be minimized because even low intensity ghosts (and reflections) will appear in the image. Further improvement of the effective fill factor can be achieved with microlens arrays.

The HDRC sensors benefit from the evolution of structure sizes in CMOS technology as well as from the increasing number of metalliza-

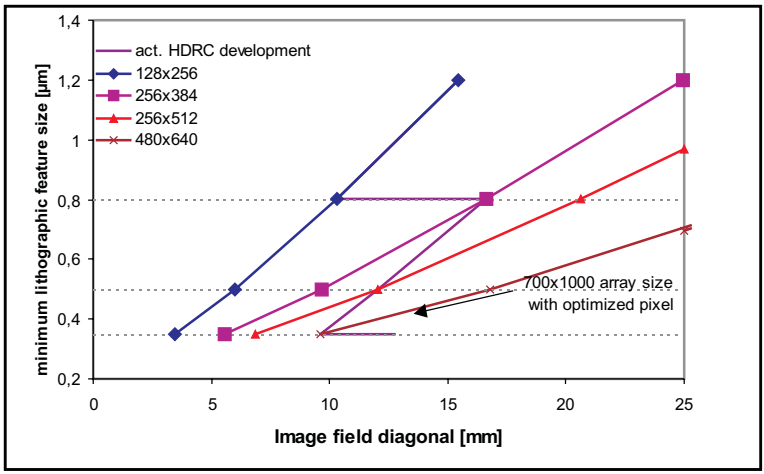


Figure 8.12: *Development of imager diagonals with decreasing feature sizes.*

tion layers and the high level of planarization available on recent technology generations. The active transistor channel area even in pixels with complex architectures covers a fraction of only 3-5 % of the pixel cell, a value that has been constant over several technology generations. The major area (about 50 %) is required for diffusion separations, well contacts, and other interconnects.

Mostly induced by local variations of semiconductor process parameters such as interface state densities, solid-state imagers suffer from so-called *fixed-pattern noise* (FPN). The HDRC sensors with their pixel-internal lin-log conversion exhibit this phenomenon as a pure offset overlay on the image that can be corrected in real-time at the video signal output. The video pad buffers of recently developed HDRC imagers provide an offset cancellation feature that operates on the analog signal path. The correction pattern requests a memory size of 1 byte per pixel to yield a resolution of 500 μV. On-chip D/A conversion and the FPN cancellation on the analog signal path provides a proper adjustment of the operation input range of the external video A/D converter and saves system cost. Image overlay information for the correction of fixed pattern noise is constant over the lifetime of the sensor and can be stored permanently in the system.

As the purpose of acquired images is shifted more and more from documentation and communication towards the control input of autonomous systems, low-level image processing operations are best performed shortly afterwards or at sensor level. Spatial filtering for noise reduction purposes or edge enhancement for image segmentation are the most time-consuming tasks on digital image processing systems.

Resistive grids operating with analog voltages to adapt the functions of biological systems have been placed on the sensor chip to implement the forementioned low-level operations. This approach provides massive parallel processing in real-time with moderate effort by hardware. In the examples of focal plane image processing described by Mead and Mahowald [7], the individual photodiodes are directly connected to the input nodes of the resistive grid. The signal bandwidth at the output of these so-called early-vision systems is drastically reduced, allowing compact system design.

However, it is also obvious that imager performance will be downgraded as a result of integrating these additional functions. On the one hand, pixel size and fill factor will be restricted as well, the high demand for power that results from such a high number of elements also has to be taken into account. Moreover, incident light influences the proper function of analog circuits, which finally limits use of these approaches.

Technology progress in the digital domain and continuously growing processing power of dedicated digital image processors and the independent development cycles of hardware and software are the winning points of pure digital image processing.

The following approach, however, shows a combination of massive parallel analog processing and serialized image data communication. The "*retina chip*" [8], which mimics the function of the already proven retina chip from Mahowald was developed for real-time post-processing HDRC images. The purpose of this device is to level the gray-scale variations that occur over larger distances in the image while simultaneously enhancing the image.

The complexity of a single cell of about 60 transistors occupying an area of $66 \times 100\,\mu\text{m}^2$ shows clearly the mismatch to imager chips. Implementing the retina chip as a stand-alone post-processor offers the opportunity to choose an array size with fewer rows than the imager's vertical extension.

As a result of the short settling time of the net after loading a row of image data and the limited affected area, a small grid can be used repetitively during processing a complete image. Presuming an adequate organization of load and read addresses in the retina, a net as wide as the imager and 32 rows in height is sufficient for full image processing.

The advantages over existing focal plane solutions are: the possibility of gaining the original image data from the imager itself and of being able to process data from different processing levels separately; only 1/10th of the retinal cells need to be used in the focal plane approach; the superb fill factor reached in the sensor part; the possibility of correcting for imager nonuniformity before post-processing image

data; and the independence of development cycles and technologies for imager and retina device.

The HDRC (High-Dynamic Range CMOS) is a registered trademark of Institute for Microelectronics, Stuttgart, Germany. The HDRC-Technology as well as the functional principle of logarithmic sensor cells implementing weak inversion transistors for achieving the log response are patented by the Institute for Microelectronics, Stuttgart.

8.8 References

[1] Tomita, T., (1968). Electrical response of single photo receptors. *Proc. IEEE (Special Issue on Neural Studies)*, **56**:1015–1023.

[2] Cornsweet, T. N., (1970). *Visual Perception*. New York: Academic Press.

[3] Chamberlain, S. and Lee, J. P. Y., (1984). A novel wide dynamic range silicon photodetector and linear imaging array. *IEEE Journal of Solid State Circuits*, **SC-19(1)**:41–48.

[4] Boring, C. G., (1950). *A History of Experimental Psychology*. New York: Appleton-Century-Crofts.

[5] Hurvich, L. M. and Jameson, D., (1966). *The perception of Brightness and Darkness*. Boston: Allyn and Bacon.

[6] Wong, H. S., (1996). Technology and Device Scaling Considerations for CMOS Imagers. *IEEE Trans. ED*, **43(12)**:2131–2142.

[7] Mead, C. A. and Mahowald, M. A., (1988). Silicon Model of Early Visual Processing. *Neural Networks*, **1**:91–97.

[8] Apel, U., Graf, H. G., Höfflinger, B., Regensburger, U., and Seger, U., (1998). Continuous parallel analogue image processing using time discrete sampling. In *Advanced Microsystems for Automotive Applications*, E. Ricken and W. Gessner, eds., pp. 149–156. Berlin: Springer.

9 Image Sensors in TFA (Thin Film on ASIC) Technology

Bernd Schneider[1], Peter Rieve[2], and Markus Böhm[1,2]

[1]Institut für Halbleiterelektronik (IHE), Universität-GH Siegen
[2]Silicon Vision GmbH, Siegen, Germany

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

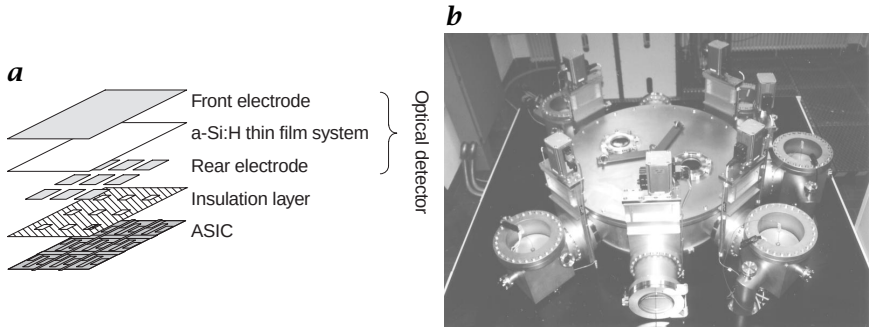


Figure 9.1: ***a*** *Schematic layer sequence of a TFA image sensor;* ***b*** *process chambers of a Plasma Enhanced Chemical Vapor Deposition (PECVD) ultrahigh-vacuum cluster system.*

9.1 Introduction

As computer vision systems become more ambitious, the performance of image sensors has become especially important. Future image sensors are expected not only to provide raw signals, but also to include part of the image processing system on-chip. This approach is suited to lower fabrication costs and improvement of sensor performance. Due to their inflexible function principle and technology, commonly used *charge-coupled devices* (CCDs) suffer from several disadvantages with regard to dynamic range, fill factor and feasibility of on-chip electronics [1]. Lately, CMOS imagers have become competitive by overcoming some of these drawbacks by using CMOS circuitry and a photodiode or photogate employing the same technology as the optical detector [2].

Unlike a CMOS imager, a sensor in ***Thin Film on ASIC*** (TFA) technology is vertically integrated, providing a fill factor close to 100% for both the detector and the circuitry. Another benefit of TFA is flexibility because the technology allows separate design and optimization of either component. An existing ASIC can be supplied with different detector structures for application-specific device optimization. The basic structure of a TFA sensor is depicted in Fig. 9.1a. The detector is formed by an a-Si:H thin-film system that is sandwiched between a metal rear electrode and a transparent front electrode [3, 4]. The crystalline ASIC typically includes identical pixel circuitry underneath each pixel detector and peripheral circuitry outside the light-sensitive area.

This chapter provides a survey of TFA research results to date and outlines future TFA applications and solutions. In Section 9.2, different thin-film detector structures are studied with regard to spectral sensitivity, dark current, temperature behavior and long-term stability. The combination of thin-film detector and ASIC is evaluated in

Section 9.3, and fundamental design approaches are discussed. Section 9.4 presents recent TFA array prototypes designed with regard to different potential applications that have been successfully fabricated and tested. Finally, further TFA concepts that are currently being developed are outlined in Section 9.5.

9.2 Thin-film detectors

9.2.1 Fabrication of thin-film detectors

Image sensors in TFA technology employ thin-film detectors based on multilayer structures of hydrogenated amorphous silicon (a-Si:H) and its alloys. The thin-film system of a TFA sensor is deposited onto the completed ASIC wafer in a *Plasma Enhanced Chemical Vapor Deposition* (PECVD) cluster system (Fig. 9.1b). The PECVD process is based on the decomposition of a gaseous compound near the substrate surface. Amorphous silicon layers are fabricated using the process gas silane (SiH_4) at substrate temperatures between 150 °C and 200 °C, which inherently leads to the formation of a silicon-hydrogen alloy. The hydrogen atoms in a-Si:H prevent the formation of dangling bonds, therefore the mid-gap defect density is decreased. The a-Si:H material properties are considerably better than those of pure amorphous silicon (a-Si), which is indeed useless for electronics because of its extremely low carrier mobility.

Due to its higher absorption coefficient in the relevant spectral range and its maximum spectral response for green light, amorphous silicon is more qualified for visible light detection than crystalline silicon. Moreover, the a-Si:H deposition sequence is adaptable to the specific requirements of an application. With a suitable layer sequence it is possible to distinguish three or more colors within the same pixel. The following sections give a survey of b/w and color detectors that have been fabricated and tested so far. The experimental data concerning both the steady-state and transient device characteristics presented in the following have been obtained on optimized test structures deposited on glass substrates or crystalline silicon wafers. The test device area is $3.14\,mm^2$.

9.2.2 Thin-film detector structures for b/w recognition

The b/w photodiodes can be realized in the form of pin layer sequences or Schottky devices, both of which have been successfully implemented in TFA sensors. A pin diode consists of a light-absorbing intrinsic a-Si:H layer sandwiched between two heavily doped layers that provide the electric field necessary for the collection of photogenerated carriers in the i-layer. Optimization of the device performance resulted in

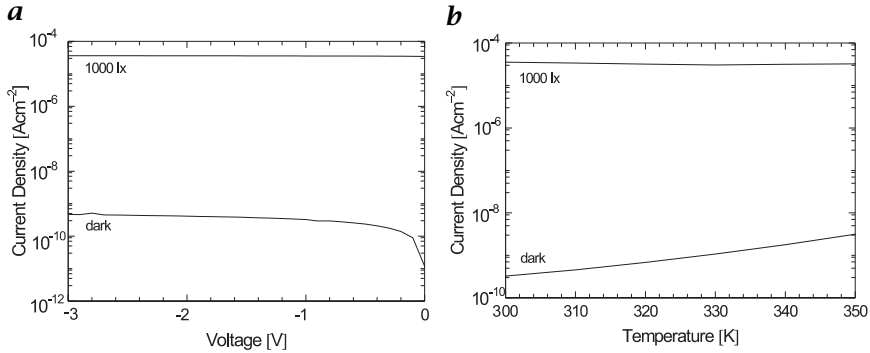


Figure 9.2: *Characteristics of an optimized pin photodiode:* ***a*** *I/V characteristics in the dark and under white light illumination with 1000 lx;* ***b*** *temperature dependence of dark and photocurrent between 300 K and 350 K.*

a configuration in which the light enters through a wide bandgap a-SiC:H p-layer, which is produced by adding methane (CH_4) to the silane (SiH_4). The layer thicknesses of the optimized structure are given by 15 nm (p-layer), 600 nm (i-layer) and 20 nm (n-layer). A semitransparent aluminum layer (12 nm) acts as front contact.

Figure 9.2a shows measured I/V characteristics of an optimized pin photodiode in the dark and under illumination of 1000 lx. The curves demonstrate excellent saturation of the primary photocurrent and a remarkably low dark current in the range of 3×10^{-10} A cm^{-2} for -1 V. The dark current is determined mainly by thermal generation within the i-layer and injection of carriers from the doped layers. The latter causes an increase of the dark current for rising reverse bias voltage and vanishes for high-quality diodes. The gap between photocurrent and dark current defines the dynamic range of the photodiode that amounts to more than 100 dB for low levels of negative bias voltage. Because no upper limitation of linearity was found for the detector current with regard to the incident illumination intensity, the operation range can easily be extended to higher illumination levels. The temperature influence on the diode performance is determined by the dark current that is proven to be thermally activated. Figure 9.2b demonstrates an exponential increase of the dark current with temperature in the range between 300 K and 350 K, whereas the photocurrent is less influenced by temperature. A temperature increase of 50 K decreases the dynamic range of the detector by 20 dB, which yields a dark current doubling temperature of about 15 K.

Usually, for devices made of amorphous silicon, degradation upon light soaking is a crucial problem. The decrease of dark and photoconductivity in connection with the Staebler-Wronsky effect results in a significant reduction of the efficiency of amorphous silicon solar cells.

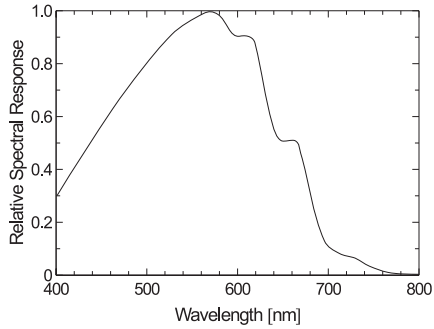


Figure 9.3: *Relative spectral response of a pin photodiode for moderate reverse bias (-1 V).*

However, light soaking experiments with the optimized pin photodetectors revealed almost stable device characteristics under realistic illumination conditions. After daylight exposure (100,000 lx) for an illumination period of 16 h, virtually no variation of the dark and photocurrent was found within the experimental accuracy [5]. One reason for the excellent long-term stability of pin photodetectors is that they are always operated under zero or reverse bias, while the Staebler-Wronsky effect is associated with carrier injection.

Due to the bandgap of amorphous silicon the spectral sensitivity of a-Si:H devices matches the responsivity of the human eye that peaks in the green spectral range. This behavior is verified by the spectral response shown in Fig. 9.3 under slightly negative bias voltage (-1 V). The response curve exhibits a maximum for 580 nm close to the green spectral region. The absolute response values for the test structures are limited to about 0.1 AW-1 due to the poor transparency of the Al front contact. However, by application of a *transparent conductive oxide* (TCO) with considerably higher transparency, the quantum efficiency can be increased to more than 90%. As the photocurrent of the pin device approaches nearly the saturation value even under short-circuit conditions, there is only a slight increase of the response curves for rising negative bias voltages.

Because in many sensor applications readout speed is a very important parameter, the transient behavior of the photodetector is of fundamental interest. The photocurrent rise and decay after switching on and off illumination of a pin diode is demonstrated in Fig. 9.4a,b. Illumination is performed with a pulsed *light-emitting diode* (LED) (pulsewidth: 80 ms) with a broad spectral distribution around 565 nm, approximating a 1000 lx light exposure. The experimental results reveal that steady-state conditions are reached within a few microseconds after switching on illumination irrespective of the incident illumination level. A slight

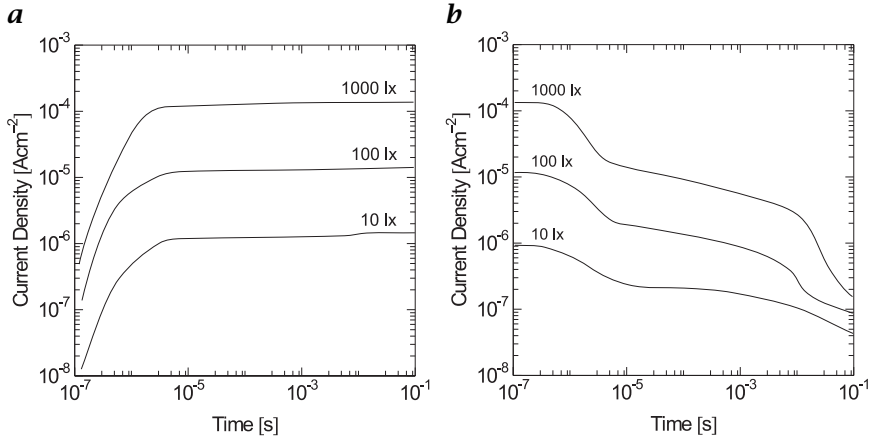


Figure 9.4: *Photocurrent transients of a pin photodiode after:* ***a*** *switching on; and* ***b*** *off illumination for different illumination levels at a bias voltage of* -1.5 *V.*

increase of the photocurrent after this time range is caused by trapping of carriers that occurs after the illumination pulse. In contrast to the very fast photocurrent rise the decay after switching off illumination exhibits a more complicated behavior. After an initial reduction of the current within 10 μs a quasi-stationary plateau occurs during which the transient current decreases only slowly. In the millisecond range the decay exhibits a steeper decrease. The significant intermediate plateau is attributed to thermal emission of trapped carriers into the extended states and subsequent field-assisted extraction and is directly linked to the continuous density of states within the bandgap. Carriers trapped in shallow states are emitted much faster than deeply trapped ones. This effect is responsible for the observed behavior that tends to increase decay times for decreasing illumination intensity.

Another type of b/w detector makes use of a metal-semiconductor contact (Schottky contact). This approach employs a semitransparent metal with high work function (e. g., Pd) or a transparent conductive contact (ITO) on top of an intrinsic a-Si:H layer. The steady-state as well as the transient experimental results of Schottky diodes are quite similar to those for the pin diode shown in the foregoing [6, 7].

9.2.3 Thin-film detector structures for color recognition

Besides simple b/w detection amorphous silicon multilayers have also capabilities in color recognition. In the past, a variety of two terminal color devices have been developed mainly based on a back-to-back diode configuration (nipin, pinip and related device structures) [8, 9, 10, 11, 12]. Common to all of these color detectors is that they ex-

ploit the wavelength dependence of the absorption coefficient in amorphous silicon and the corresponding carrier generation profile inside the device. The absorption coefficient in a-Si:H exhibits a continuous decrease of more than one order of magnitude from the short to the long wavelength end of the visible spectral range. According to this wavelength dependence of the absorption coefficient the absorption length changes from 50 nm for blue light to 250 nm for green light while still longer wavelengths are absorbed more homogeneously and can penetrate deeper into the material. Bandgap engineering by alloying the amorphous silicon with carbon or germanium also affects the generation profile. Because carrier mobility in amorphous silicon is rather poor, a strong electric field has to be applied for efficient carrier collection. The special feature in the color detectors is the voltage-controlled shift of the main collection region that is obtained by appropriate variation of the drift length. Due to the definition of this parameter as product of the carrier mobility and lifetime and the electric field the required shift of the collection parameters can be performed by $\mu\tau$-engineering, or electric field tailoring, or a combination of both.

Two classes of color devices can be distinguished depending on their geometry and operation principle. *Bipolar color detectors* consist of a multilayer in the form of an antiserial diode arrangement (e. g., nipin or pinip) that can be extended by insertion of additional doped layers or by further subdivision of the absorber layers. With this type of color detector the voltage range necessary for full color separation as well as the photocurrent covers both polarities. In contrast to this approach, *unipolar color detectors* based on a simple pin structure with subdivided i-layers have been realized successfully. Here the voltage range and the photocurrent are limited to only one polarity. In the following one device structure representing each type of color detector is described and characterized by experimental data.

Figure 9.5a displays layer sequence and schematic electric field distribution of a nipi^{2}n three color detector consisting of a nip diode with wide bandgap a-SiC:H absorber on top of a heterojunction pi^{2}n diode. Regarding the direction of light penetration, the first i-layer of the bottom diode also uses a-SiC:H material (bandgap 1.8 eV) while the bottom layer is made of pure a-Si:H. The discontinuity at the interface between the two materials with different dielectric constants causes a step in the electric field that is further influenced by the space charge accumulated at the interface. The resulting electric field profile enables carrier collection in the first i-layer where the green photons are absorbed for moderate negative values of bias voltage. Carriers photogenerated in the bottom i-layer are lost due to increased recombination in the low-field region. With decreasing voltage the electric field region extends into the bottom i-layer and allows carrier collection in the complete bottom diode, thus shifting the spectral sensitivity to longer wavelengths.

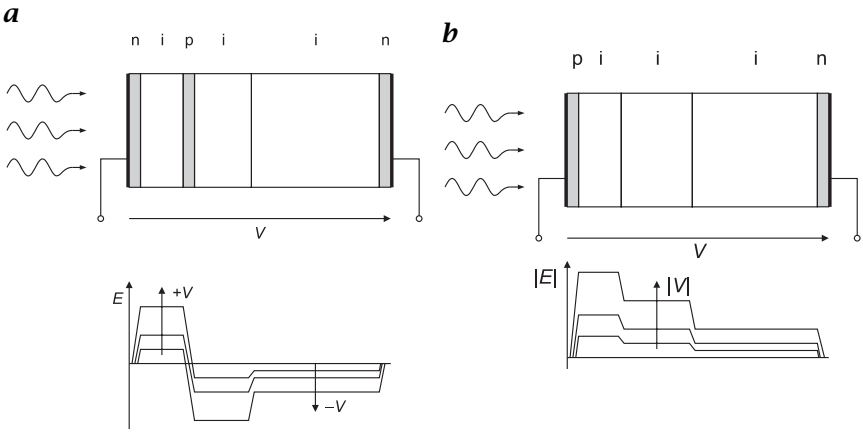


Figure 9.5: *Deposition scheme and schematic electric field profile within:* ***a*** *a nipi*^{2}*n; and* ***b*** *a pi*^{3}*n multilayer.*

For positive bias the top diode is reverse-biased and carriers generated by blue light illumination are extracted.

With a high color separation between red and green as the most important goal of device optimization the drift parameters in the two i-layers of the bottom diode have to be adjusted as different as possible while the higher drift length is required in the first i-layer. Because the difference in the dielectric constants is not sufficient, additional $\mu\tau$-engineering is desired. For enhancement of the $\mu\tau$-product of the a-SiC:H material these layers have been deposited under strong hydrogen dilution of the silane/methane source gas mixture. This has been proven to be necessary because the $\mu\tau$-product of amorphous silicon carbon alloys deposited solely from silane and methane is about a factor of five smaller than in pure a-Si:H. In contrast, the hydrogen-diluted a-SiC:H films exhibit carrier transport parameters that are one order of magnitude better than films with equivalent bandgap deposited without hydrogen dilution as has been verified by photoconductivity measurements on single layers deposited at nominally the same deposition parameters. Another approach to improve the electric field distribution within the bottom diode employs a superthin n-layer between the a-SiC:H and the a-Si:H region that leads to a nipi(δn)in structure [13].

The spectral response curves as a function of bias voltage for an optimized nipi^{2}n multilayer are given in Fig. 9.6a. The data demonstrate blue response for positive applied bias with a peak wavelength of 480 nm while a shift of sensitivity is found in the negative bias range. Down to -0.6 V the sensor is sensitive to green (550 nm) and for decreasing the negative bias below -1.0 V the peak wavelength turns to red (620 nm) along with a significant increase of sensitivity until satu-

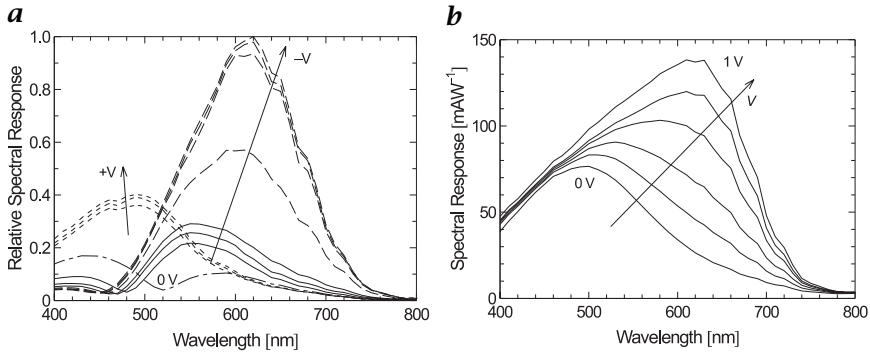


Figure 9.6: ***a*** *Normalized spectral response of a nipi^{2}n three color detector at bias voltages ranging from -3 V to +2 V;* ***b*** *spectral response of a ni^{2}p two color detector at bias voltages ranging from 0 V to 1 V.*

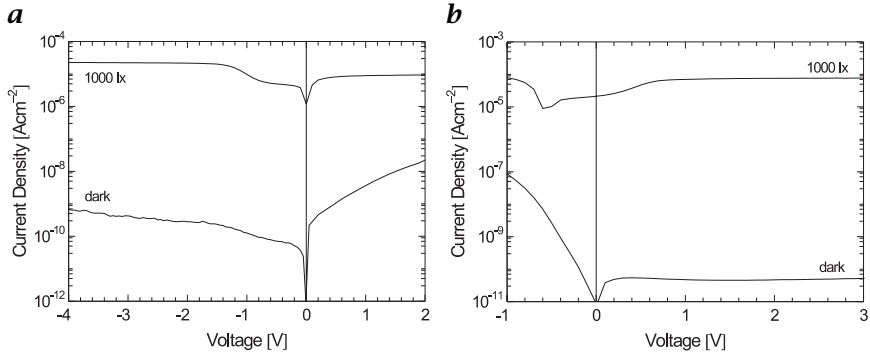


Figure 9.7: *I/V characteristics in the dark and under white light illumination with 1000 lx:* ***a*** *nipi^{2}n three-color; and* ***b*** *ni^{2}p two-color detector.*

ration is reached for voltages lower than -2.0 V. The nipi^{2}n multilayer structure clearly emerges as three color detector with three linearly independent sensitivity ranges.

The I/V curves plotted in Fig. 9.7a represent the dynamic range of the nipi^{2}n device. The photocurrent shows a distinct increase in the negative voltage range interval between -0.5 V and -1.5 V, indicating the transition from green to red sensitivity. The ratio of photocurrent to dark current amounts to 70 dB in the voltage interval required for color separation (-2.0 V to +1.0 V). This value can be increased by further 10 dB employing a highly transparent front contact instead of a semitransparent metal electrode. While the dark current for reverse-biased bottom diode remains below 10^{-9} A cm^{-2}, the dynamic range is limited by the strong increase of the dark current in the positive voltage

range. This behavior is supported by the thin i-layer of the top diode (80 nm) and has also been observed for nipin two color detectors [13].

One of the most severe drawbacks common to all color detectors with bipolar operation, including the forementioned nipi^{2}n color detector, is the transient behavior that suffers from long-term trap recharging currents caused by the back-to-back configuration of two diodes. The transients after changes of bias or illumination show a significant dependence of the illumination level, and steady-state condition is reached later the lower the illumination intensity. This disadvantageous transient behavior has been observed for nipin two color detectors [14, 15] as well as for nipi^{2}n three color devices [16].

The speed limitations for color detectors based on antiserial diode structures are overcome by unipolar devices explained in the following. This type of color detector is based on a simple pin or nip structure with the intrinsic absorber subdivided into two or more i-layers by abruptly or continuously changing the source gas mixture during deposition, resulting in pi^{3}n or ni^{2}p layer sequences, for instance. These multilayers contain at least one heterojunction that causes different collection parameters in the adjacent i-layers in analogy to the bottom diode of the nipi^{2}n structure discussed in the foregoing. In an optimized three color detector the bandgap as well as the $\mu\tau$-product decrease in the i-layers when passing through the device in light penetration direction. Due to the dielectric constant that is correlated with the bandgap, the electric field drops from the top to the bottom i-layer as sketched in Fig. 9.5b. The electric field profile in conjunction with the generation profile allows collection of carriers generated by strongly absorbed radiation for low values of reverse bias or in short-circuit conditions. With increasing reverse bias the collection of carriers from deeper regions in the device is enhanced, resulting in a red shift of the spectral sensitivity. The main advantage of unipolar color diodes with respect to the bipolar ones consists in the operation mode that ensures that the device is permanently operated in reverse bias direction, thus avoiding time-consuming trap recharging effects occurring in forward-biased diodes. Furthermore, the bias voltages of unipolar detectors are in the same range as usual ASIC supply voltages.

Figure 9.6b demonstrates the voltage-controlled variation of the spectral response of a ni^{2}p two color detector consisting of a wide bandgap (135 nm) a-SiC:H i-layer in front of a normal bandgap (1000 nm) a-Si:H layer sandwiched between a top n- and bottom p-layer providing the built-in electric field. The maximum of the response curves shifts continuously from 490 nm for short circuit to 620 nm for 1 V reverse bias while the absolute sensitivity increases as well. It becomes obvious that the response curves measured for higher value of reverse bias always include the curves obtained for lower bias, thus indicating that carrier collection proceeds towards deeper regions for increasing reverse bias.

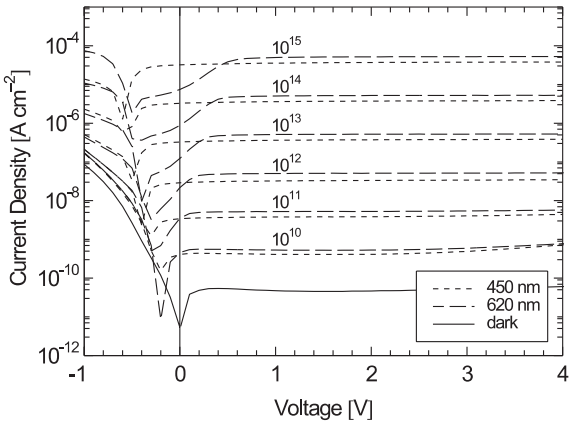


Figure 9.8: *I/V characteristics of a ni^{2}p two color detector in the dark and under various monochromatic illumination levels.*

This effect is responsible for the less pronounced color separation of unipolar photodiodes in comparison to bipolar color detectors.

The unipolar a-SiC:H/a-Si:H color detectors exhibit outstanding dynamic range of more than 120 dB (Fig. 9.7b) due to an extremely low dark current around 3×10^{-11} A cm^{-2} that is nearly independent of bias voltage for reverse bias up to 3 V. The mentioned value is close to the limit of the thermal generation current for the i-layer given by Street [17] so that the absence of contact injection and local shunt effects can be concluded. The photocurrent shows a slight increase in the voltage range between short circuit and 1 V reverse bias correlating with the onset of red sensitivity and perfect saturation for higher amounts of reverse bias.

In order to get an impression of linearity in Fig. 9.8 the I/V characteristics of the ni^{2}p detector are plotted for blue (450 nm) and red (620 nm) monochromatic illumination conditions at various illumination levels (ranging from 10^{10} - 10^{15} photons cm^{-2}s^{-1}). The currents exhibit exact linearity with regard to the illumination level. The measured values j_{ph} obey the commonly found photoconductivity relation

$$j_{ph} \propto \Phi^{\gamma} \tag{9.1}$$

while the exponent γ is almost identical to one. Furthermore, the I/V curves demonstrate an interesting feature concerning the blue/red color separation. The crossing points between the blue and red illumination curves show a remarkable shift to open-circuit conditions by about 0.12 V per decade for decreasing illumination intensity. This effect is caused mainly by free and trapped carriers that have major influ-

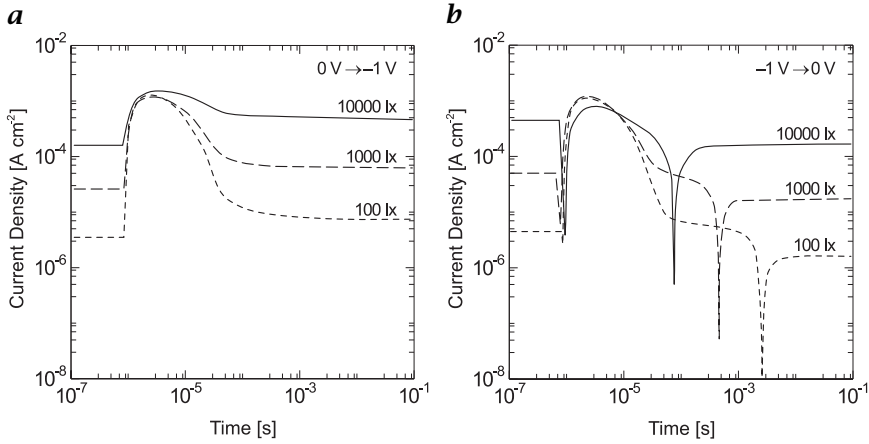


Figure 9.9: *Photocurrent transients of a ni^{2}p two color detector photodiode after switching* ***a*** *from 0 V to -1 V; and* ***b*** *from -1 V to 0 V for different illumination levels.*

ence on the electric field profile for higher generation rates. A similar trend has been found for other device structures by numerical simulations [18]. This result clearly indicates that not only the physical absorption and transport parameters of the intrinsic layers determine the device functionality, but also the charge carriers generated during operation. As a consequence, the applicable illumination range of the device is limited to a value smaller than the dynamic range mentioned in the foregoing when full color recognition is required.

The most important advantage of the unipolar color detectors over the bipolar ones is the superior transient behavior. Due to the simpler device geometry and the operation principle, the unipolar detectors never need to be operated in forward bias, resulting in faster current transients. Figure 9.9 displays the transient photocurrent measured after switching of bias between the two operation voltages for the sample illuminated with white light of variable intensity. The two switching transients show significant differences. For the reverse-bias pulse (0 V → 1 V, blue to red sensitivity) after an initial capacitive peak the steady-state current is reached within a 10% tolerance interval after about 600 μs independent of the illumination level. However, if switching is performed from reverse bias into short-circuit conditions (red to blue sensitivity) a different behavior can be noticed. Here the current transients show a remarkable illumination dependence that manifests in a longer delay before the steady state is reached for lower light intensity. The measured time required for recharging the traps ranges from 200 μs for 10,000 lx to 5 ms for 100 lx. Moreover, the peaks in the logarithmic current scale indicate a change in the direction of the

photocurrent during the transient. Partial field reversal in the bottom i-layer and injection of holes from the rear p-layer are assumed to be responsible for the current flowing opposite to the steady-state value during the transient until the charge state of the traps has rearranged. The observed effect is coupled to the low-field region in the bottom i-layer and seems to be inherent to devices containing regions with low electric field. To much less extent the same tendency can also be found for simple pin diodes. With regard to application of the color detector in two-dimensional image sensor arrays the prolonged duration of the transient current that occurs only for one switching process does not define a general drawback. For example, this delay time can be used for readout of the color sensor array that takes several milliseconds depending on the dimensions of the sensor matrix.

9.3 TFA properties and design considerations

9.3.1 Noise in TFA sensors

Important characteristics such as the sensor's *dynamic range* and the *signal-to-noise ratio* (SNR) are affected by thermal- and transport-related noise sources. According to the TFA concept, the noise sources (see also Section 7.5.3) of the amorphous pin diode and the crystalline pixel and peripheral electronics are treated separately.

Noise in the photodetector consists of shot noise and flicker noise. The *shot noise* power spectral density of the pin diode is constant for frequencies smaller than the reciprocal transit time. In the relevant frequency range shot noise is white noise and determined by the current. The power spectral density is proportional to the sum of dark and photocurrent.

$$W_{i\_shot}(f) = 2eI_0 \qquad I_0 = I_d + I_{ph} \tag{9.2}$$

Flicker noise dominates the noise power spectral density at low frequencies. The flicker noise power spectral density of pin diodes is almost proportional to the square of the dc current and to the reciprocal of the frequency. The flicker noise can be described with the measured dependence from current and frequency.

$$W_{i\_l/f}(f) = \frac{c}{A} I^{2\beta} \frac{1}{f^{\gamma}} \tag{9.3}$$

The parameter γ is close to unity; β equals one for reverse-bias dark current and is lower at forward bias and illumination [19]. Equation (9.3) is similar to Hooge's law for homogenous materials, however, it was found that Hooge's law is not valid for pin diodes [20]. Measurements sustain this result, since the flicker parameter c is not constant

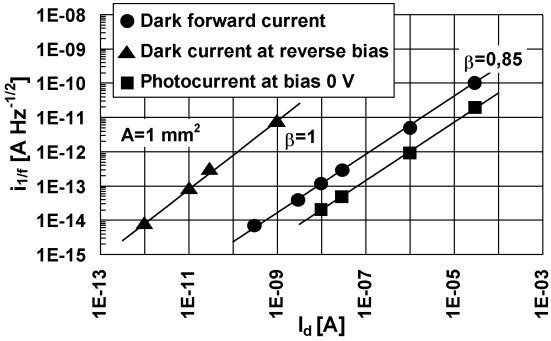


Figure 9.10: *Flicker noise current spectral density of 1 µm thick pin diode at 1 Hz at different operation points.*

as predicted by Hooge's law. It strongly depends on the operating point of the sensor, that is, illuminated or dark and reverse or forward bias. The measured flicker noise power spectral density of the pin diode photocurrent in Fig. 9.10 is several orders of magnitude lower than that of the dark current.

The flicker noise of the photocurrent has to be distinguished from the flicker noise of the dark current. The principle of superposition is valid for dark current and photocurrent noise components as well as for the currents themselves [21].

Equation (9.3) shows the increase of flicker noise with decrease of the pixel area. Boudry and Antonuk [22] confirmed this trend by noise measurements on reverse-biased pin diodes. They found that the data do not scale with A^{-1} and is better approximated as scaling with $A^{-1/2}$. With this scaling the noise of the dark current and the noise of the photocurrent are dominated by shot noise. However, the influence of pixel edge leakage currents on the forementioned measurements and scaling should be further investigated.

Fixed-pattern noise (FPN) is caused by differences of pixel dark currents, pixel coupling due to pixel bias differences and differences of the pixel and periphery circuit offset voltage. Offset differences can be removed by *correlated double sampling* (CDS); however, this increases the circuit complexity. In the CDS mode the reset level is subtracted from the signal in order to eliminate ASIC offset voltages and *kTC noise* and to reduce *1/f noise* of the pin diode and the ASIC.

Reset noise is produced by thermal noise of the reset transistor and the pin diode series resistor in interaction with the capacitance of the diode. This kTC noise is determined only by the capacitance and temperature:

$$\overline{u_{kTC}} = \sqrt{kT/C} \tag{9.4}$$

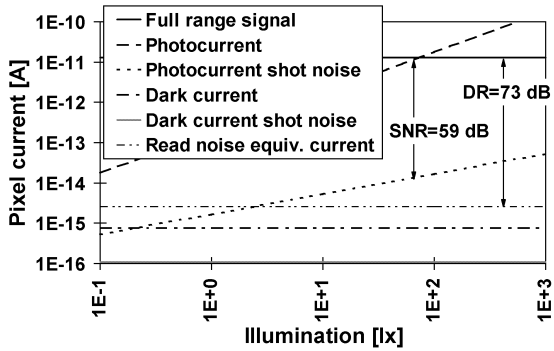


Figure 9.11: *Signal-to-noise ratio (SNR) and dynamic range (DR) of the VALID image sensor (pixel size: 16 µm × 16 µm, dark current density: 3×10^{-10} A/cm^2, photocurrent density: 7×10^{-8} A/L × cm^2).*

It results in a diode voltage uncertainty at the end of the reset cycle. With a specific capacitance of 16 nF/cm^2 the kTC noise is 112 µV for the *locally adaptive sensor* (LAS) (Section 9.4.3) and 175 µV for the *Varactor AnaLog Image Detector* (VALID) sensor (Section 9.4.1). A capacitance connected in parallel to the pixel capacitance limits the kTC noise in sensors with minimized pixel area. This effect is realized by a varactor in the VALID sensor.

Read noise is the total dark signal output noise, including shot and flicker noise of the diode, reset noise and thermal noise sources of the ASIC. The read noise is found to be 200 µV to 300 µV for the described sensors and thus is dominated by reset noise and thermal channel noise.

The *signal-to-noise ratio* (SNR) is limited by the shot noise of the photocurrent. The full-range SNR is about 60 dB. The *dynamic range* (DR) is limited by the read noise and exceeds 70 dB.

9.3.2 TFA design for high local contrast

Coupling effects between TFA pixels are quite different from those in CCD and CMOS sensors due to the different material properties and geometry of the a-Si:H detector. Pixel coupling in TFA sensors is mainly attributed to lateral balance currents flowing through the basically unpatterned thin-film system. In the usual TFA configuration, the front electrode is common to all pixel detectors, whereas the rear electrode potentials are floating in order to allow integration of the photocurrent. Therefore lateral balance through the common thin-film system occurs if two neighboring pixels have different illumination intensities. This leads to a reduction of local contrast, that is, contrast between neighboring pixels. The highly doped (usually n-type) bottom layer of a pin

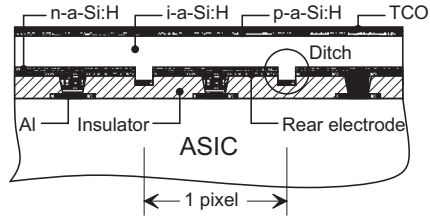


Figure 9.12: *Thin-film system "self structuring" for high local contrast.*

or similar structure mainly contributes to the balance currents, while the conductivity of the i-layer is negligible. Furthermore, local contrast decreases with increasing integration time, as the balance currents are a parasitic contribution to the integrated signal.

A simple measure to suppress lateral balance currents is a self-structured thin-film system as depicted in Fig. 9.12. After completion of the ASIC and before thin-film deposition ditches are etched into the insulator between detector and ASIC that define the pixel borders. The ditch geometry is identical to the detector rear electrode shape, so no additional mask is required for this step. During the PECVD process the thin n-layer is torn at the edges of the ditches, thus lateral balance currents between the pixels are efficiently suppressed. An alternative electronic method for local contrast enhancement is presented in Section 9.4.2.

9.3.3 TFA design for high dynamic range

The range for a linear pixel signal in an image sensor is limited to less than 80 dB, as was demonstrated in Section 9.3.1. This range turns out to be insufficient for applications under real world illumination, for example, automotive vision systems. By means of global sensitivity control, that is, adjusting equal sensitivity for all pixels, the dynamic range can be extended to 100 dB or more that is required for a lane tracking system to handle real world illumination situations. However, if this entire range is covered throughout a single frame, global sensitivity control is ineffective, because saturation as well as signals below the noise level may occur simultaneously. A strict demand apart from blooming prevention is therefore that any portion of the image—within the specified dynamic range—can be recognized any time [23].

In order to achieve high dynamic range with a given linear signal range below 80 dB, the photoresponse has to be compressed or split. A common concept for compression exploits the logarithmic voltage-current response of diodes or MOSFETs in subthreshold operation, as described in Section 7.3.5. A pixel concept with logarithmic compression fits 120 dB of intensities into a voltage range of a few hundred

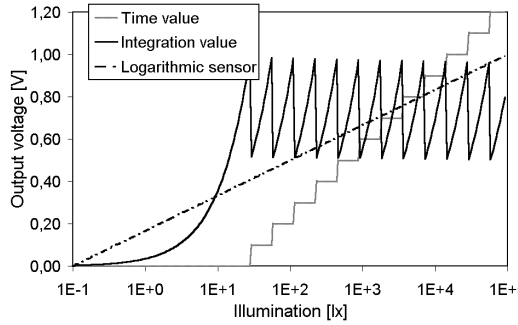


Figure 9.13: *Comparison of logarithmic and linear autoadaptive output voltages.*

millivolts. In contrast to the logarithmic sensors, the actual working ranges for the TFA LAS (Section 9.4.3) and *locally autoadaptive sensors* (LARS) (Section 9.4.4) are determined by the individual pixel integration time control. The complete illumination information is included in two signals with moderate dynamic ranges, the integration value and the time value that are both read out from the LARS pixel. External timing control allows switching between fixed and adaptable integration times whenever necessary. The LARS concept allows dynamic ranges of 150 dB or more for the photosignal.

If an actual logarithmic characteristic is involved, pixel-to-pixel variations in the circuit offset and gain (i. e., *fixed-pattern noise* FPN) lead to exponentially amplified differences in the reconstruction of the original photosignal. An exponential timing of LARS also leads to a quasi-logarithmic compression; however, it is significantly less sensitive to FPN. This becomes apparent in Figs. 9.13 and 9.14, where the characteristics of LARS for exponential timing are compared to those of a logarithmic sensor. In Fig. 9.13 the voltage outputs of the two types are plotted over intensity. In a logarithmic sensor the voltage output is inherently proportional to the logarithm of illumination intensity. The integration value of LARS, however, rises linearly within the integration intervals given by the time value. Figure 9.14a shows the slopes of the output characteristics in units of output voltage per photocurrent. The slope of the LARS output is steeper for most of the dynamic range, thus it exhibits lower sensitivity to temporal noise as well as FPN. This is demonstrated in Fig. 9.14b, where the ratio of photocurrent to noise equivalent current (NEC) in the detector is depicted. The noise values are based on a noise floor of 500 μV for usual ASIC technologies. The input referred SNR is 20 dB higher for LARS except for very low illumination intensities ($<$ 10 lx in this example), where a logarithmic sensor is advantageous, whereas its transient response becomes unsatisfac-

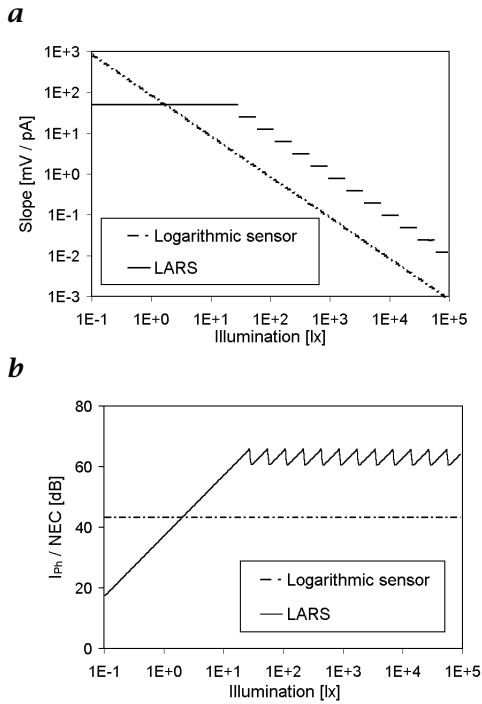


Figure 9.14: *Comparison of logarithmic and linear autoadaptive:* ***a*** *signal conversion slope; and* ***b*** *input referred SNR.*

tory (Section 7.3.5). Considerations similar to those for FPN show that the timing-driven range compression is also immune to temperature drift.

Basically, the autoadaptive concept is applicable to mere CMOS as well as TFA technology. However, in the first case the extensive pixel electronics would lead to a poor fill factor or unacceptably large pixels.

9.3.4 Effects of CMOS device downscaling

In contrast to CCDs, so far TFA and CMOS sensors benefit directly from the decreasing feature sizes of CMOS technologies, because smaller structures enable increased resolution or lower fabrication costs (see also Chapter 7.8.1). However, standard technologies are not optimized for imaging devices; the effects of further downscaling have to be given serious consideration [24]. The minimum useful pixel size is given by the spatial resolution of conventional optical systems, which is about 4 μm. Thus CMOS processes with feature sizes below 0.25 μm will lead to higher fill factors of CMOS sensors with equal functionality, whereas in TFA sensors with their inherently high fill factor the transistor areas

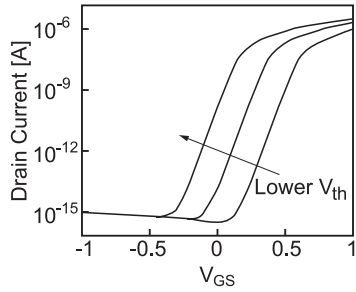


Figure 9.15: *Off current increase for decreasing threshold voltages of MOS transistors.*

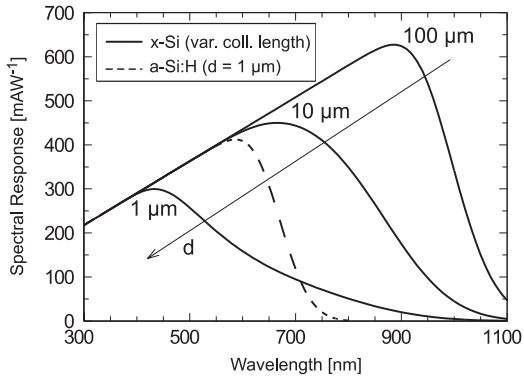


Figure 9.16: *Spectral responses of x-Si devices for decreasing depletion depths and of a-Si:H.*

can be kept comparatively large in order to optimize yield and matching.

Below the threshold voltage of a MOSFET, the drain current drops exponentially with decreasing gate-source voltage. As the threshold voltage is lowered in future CMOS processes, the off current at 0 V will significantly increase, as it is demonstrated in Fig. 9.15. As a consequence, a reverse bias has to be applied to a MOSFET in the off state in order to maintain low off currents, which requires additional design efforts. Furthermore, as the supply voltage is reduced to about 1.8 V in a 0.18 µm process, it is obvious that the available voltage swing and therefore the dynamic range is noticeably reduced for both CMOS and TFA sensors.

Increased doping concentrations and shallower implantations are prerequisites for smaller feature sizes. Both measures lead to shallower depletion depths of photodiodes and photogates in CMOS sensors, while the first also decreases carrier lifetime in the photosensitive

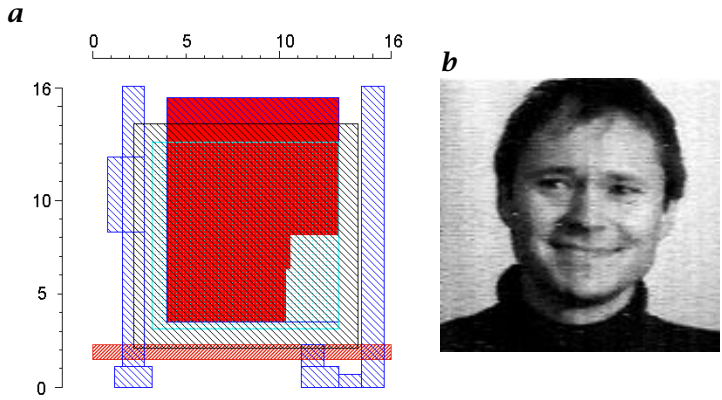


Figure 9.17: *The VALID sensor:* ***a*** *layout of a VALID pixel;* ***b*** *example image.*

region. As a result, fewer photogenerated carriers can be collected in these detector types, and the photocurrent decreases in future CMOS sensors. For feature sizes below 1 μm with a depletion depth of 1 μm the quantum efficiency of an x-Si is already poor compared to an a-Si:H thin-film detector, and will further decrease for still smaller feature sizes. This effect will be a decisive obstacle in using future technologies for CMOS imagers, whereas it does not affect the ASIC independent thin-film detectors of TFA sensors.

9.4 TFA array prototypes

9.4.1 TFA sensor with one-transistor pixel

With an a-Si:H photodiode, a MOS capacitor and a transfer transistor as the only elements, the **V**aractor **A**na**L**og **I**mage **D**etector (VALID) provides the smallest possible pixel size for TFA sensors. The photogenerated charge is stored in the blocking capacitance of the photodiode and transferred to the readout column when the transfer transistor is activated [6, 25]. The additional capacitor serves to increase the saturation illumination and the column output voltage and reduces kTC noise (see Section 9.3.1). The pixel matrix exhibits very low *fixed-pattern noise* and high linearity, because no active pixel concept is implemented. However, the lack of column drivers leads to a limitation of the maximum line number or the output amplitude. The dynamic range is limited to about 60 dB. The VALID concept is suited for low-cost fabrication of image sensors for less demanding applications as, for example, motion control devices.

The current VALID prototype consists of 128×128 pixels with an area of $16\,\mu\text{m} \times 16\,\mu\text{m}$ each. The layout of a single pixel is depicted in

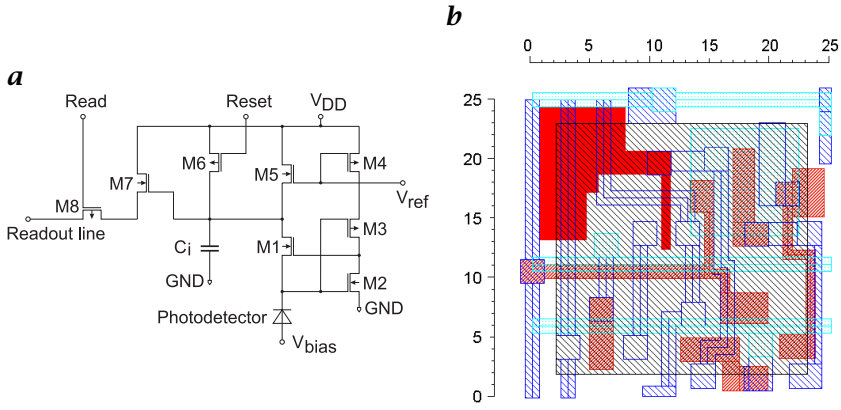


Figure 9.18: *The AIDA sensor:* ***a*** *circuit diagram; and* ***b*** *layout of a pixel.*

Fig. 9.17a. The detector rear electrode is given by a rectangular hatching, the varactor area is shaded. In a 0.5 µm process the pixel size will be reduced to less than 100 μm^2. Figure 9.17b depicts an image taken with the VALID array.

9.4.2 TFA sensor with constant voltage circuit

In order to achieve the highest possible yield and to lower fabrication costs, the thin-film system of a TFA sensor is fabricated in a PECVD cluster tool without temporarily being taken out of the vacuum for lithography. Therefore the pixel is simply defined by the size of its rear electrode. The continuous thin-film layer, however, permits lateral balance currents between adjacent pixel detectors, resulting in a reduced local contrast (Section 9.3.2).

The **A**nalog **I**mage **D**etector **A**rray (AIDA) overcomes the coupling effect by electronic means. A circuit inside each pixel provides a constant rear electrode potential, whereby the local contrast is significantly enhanced compared to VALID. The pixel schematic is given in Fig. 9.18a. The photocurrent is fed into the capacitor C_i that is part of the ASIC; M1 ... M5 form the constant voltage circuit, M6 and M7/M8 serve for reset and readout of C_i, respectively. The integration time and therefore the sensitivity of the pixels is controlled globally, thus a dynamic range of far more than 60 dB can be covered [26].

The prototype consists of a 128×128 pixel array with a pixel size of $25\,\mu\text{m} \times 25\,\mu\text{m}$. The pixel layout is depicted in Fig. 9.18b. In a 0.5 μm ASIC process the pixel will shrink to about $18\,\mu\text{m} \times 18\,\mu\text{m}$. Images taken with the sensor array are given in Fig. 9.19a,b. The device has been tested for illumination levels as high as 80,000 lx and proved to be virtually free of blooming effects or image lag.

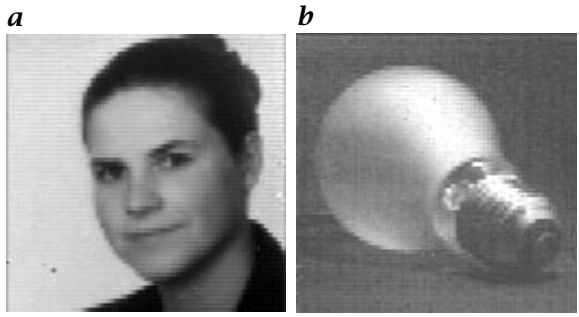


Figure 9.19: *Images taken with AIDA.*

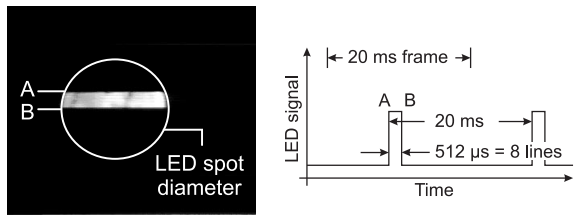


Figure 9.20: *Measurement of AIDA transient response with pulsed LED.*

To evaluate the largest possible portion of each frame, it is desirable to have a fast transient behavior. To estimate the transient response of AIDA, the sensor was illuminated by a 512 μs *light-emitting diode* (LED) pulse synchronized to the 40 ms frame (Fig. 9.19). The effective integration time was 64 μs and equal to the duration of one line. Thus eight lines were illuminated by the LED. Since the ninth and the following lines show no visible signal, the response time of the pixels is below 64 μs.

9.4.3 Locally adaptive TFA sensor

The TFA sensor, ***Locally Adaptive Sensor*** (LAS) for automotive applications has been developed in order to overcome the problems of global sensitivity control discussed in Section 9.3.3. The underlying concept of locally adaptive integration control allows the sensitivity of each single pixel to be adapted to the illumination condition at its respective location in the image. In this way a dynamic range of over 100 dB can be covered throughout the chip at any time [26, 27]. A similar functionality has been demonstrated for CCDs by Chen and Ginosar [28].

The block diagram of a locally adaptive pixel is depicted in Fig. 9.21a. Basically, the sensitivity of a pixel is controlled by determining the time during which the photocurrent from the a-Si:H multilayer is integrated

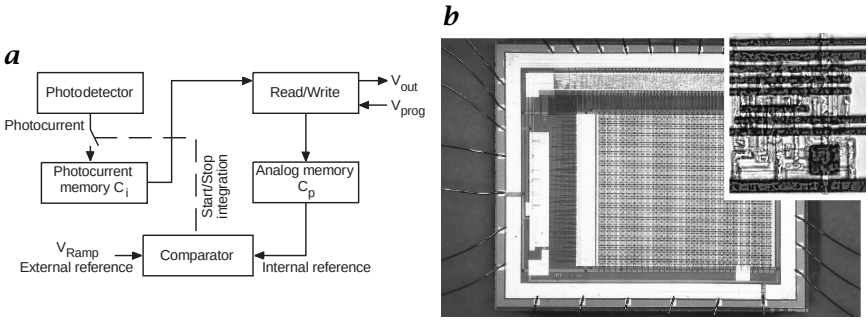


Figure 9.21: *The LAS sensor:* ***a*** *block diagram of a pixel; and* ***b*** *photograph of LAS array and pixel.*

into an on-chip capacitance C_i. A second capacitor C_p contains the programmed timing information represented by a voltage. In the first phase of every frame, C_p is precharged to a value corresponding to the individual illumination intensity of the pixel. Second, a voltage ramp is applied to the pixel and compared to the voltage across C_p. As soon as the ramp exceeds the programmed value, integration of the photocurrent starts. With the falling slope of the ramp the integration is stopped, C_i is read out and afterwards reset to its starting value.

A die photograph of the 64×64 pixel LAS prototype array is depicted in Fig. 9.21b, the inset shows a photograph of a $50\,\mu\text{m} \times 40\,\mu\text{m}$ LAS pixel. The prototype chip includes line and column decoders for pixel programming and for reading out the integrated pixel voltage. The required voltage ramp is generated on-chip and is applied to every line of the sensor array. Finally, a sequencer and timing unit for providing the peripheral circuitry and the pixel array with clock signals is implemented.

Figure 9.22a,b demonstrates the behavior of the pixel circuit for two different illumination conditions with three different integration times each. The large switching offset of the output voltage at the beginning of each integration period does not limit the dynamic range of the pixel. It merely contributes a constant offset that has to be accounted for when designing the readout stages. Optimum sensitivity of the pixel is indicated by a linear increase of the integrated signal over several hundred millivolts without saturation. The measurements prove that the pixel works at 6 lx with a few milliseconds as well as at 80,000 lx with about $10\,\mu$s maximum integration time before saturation occurs.

9.4.4 Locally autoadaptive TFA sensor

The *locally autoadaptive sensor* LARS provides a very high global dynamic range by adapting the integration time for each individual pixel

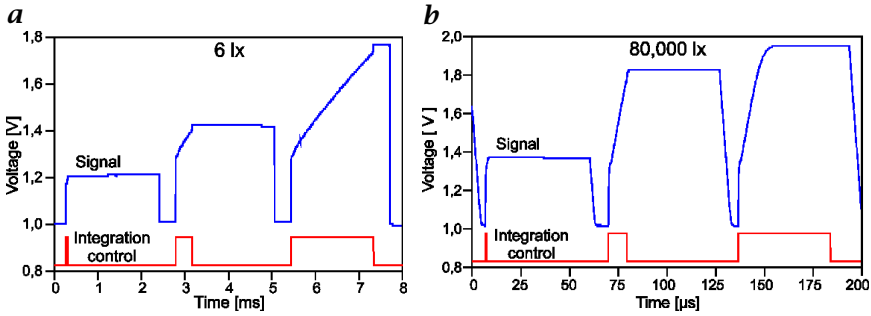


Figure 9.22: *LAS signal voltage for arbitrarily chosen integration durations:* ***a*** *at 6 lx; and* ***b*** *at 80,000 lx .*

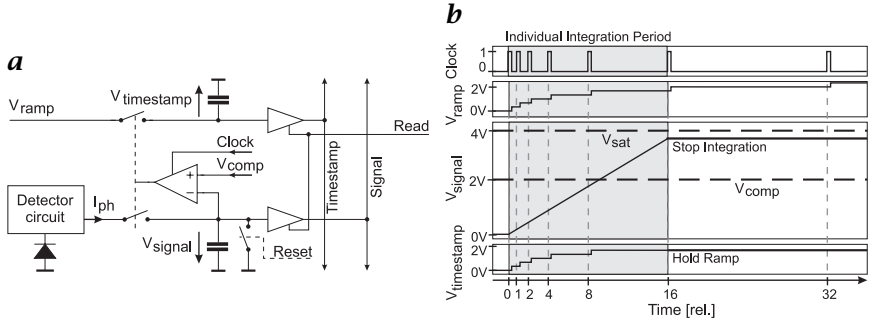


Figure 9.23: *Block diagram of LARS pixel.*

according to the local illumination intensity [23, 29]. Unlike the LAS (Section 9.4.3), the integration time control takes place in the pixel itself in real time. Therefore, off-chip circuitry and additional time for pixel programming are not required. Furthermore, a sudden change to a high illumination intensity is detected immediately; thus the integration of the photocurrent is stopped before the integration capacitor is saturated.

Figure 9.23a shows the schematic of a locally autoadaptive pixel, and Fig. 9.23b shows the corresponding timing diagram. The current of the photodiode is integrated on the integration capacitance to a signal voltage V_{signal}. On every rising edge of the clock input this voltage is compared to a reference voltage V_{comp} that is slightly below half the saturation value of V_{signal}. If the integrated signal is still below V_{comp} the integration time is doubled, whereas the comparator terminates the integration via the switches if the signal exceeds the reference level. With every clock the timestamp input V_{ramp} climbs up one step and is sampled and held in the timestamp capacitance at the moment the integration is terminated. At the end of the integration phase the in-

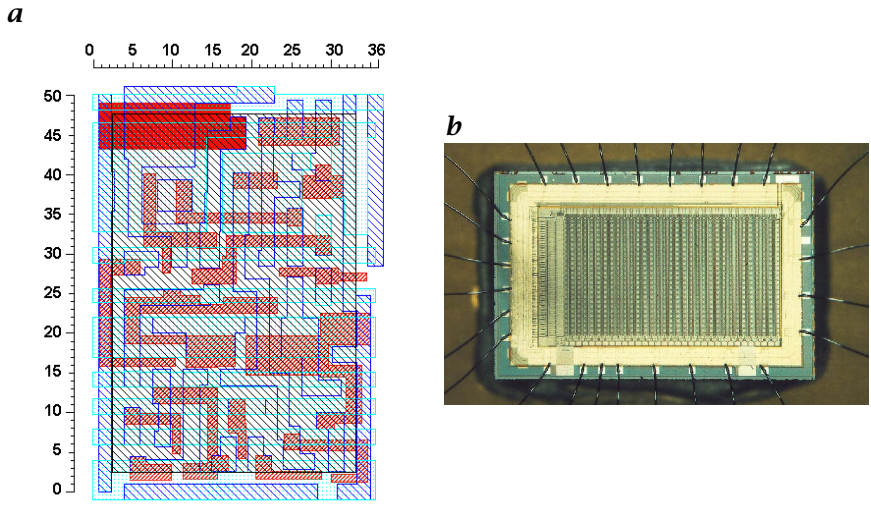


Figure 9.24: ***a*** *Layout of 35 µm × 49 µm LARS pixel;* ***b*** *photograph of a 64 × 64 pixel LARS array.*

formation stored in every pixel consists of the integrated signal and the timestamp, and the latter clearly defines the integration duration of the corresponding pixel. The binary exponential increase of the integration time steps in the forementioned example corresponds with $V_{comp} \leq 1/2\ V_{sat}$. In this way it is ensured that the range for the signal voltage at the end of the integration time is $1/2\ V_{sat} \leq V_{signal} \leq V_{sat}$.

Figure 9.24a shows a complete pixel layout in a 0.7 µm low-power CMOS technology. The autoadaptive functionality is realized with 24 transistors and two capacitors covering an area of 35 µm × 49 µm. The first 64 × 48 pixel prototype shown in Fig. 9.24b includes the readout periphery and line and column address generators. Further integration and improved performance will be accomplished by on-chip A/D conversion of the output signals on future TFA prototypes.

The illumination range (global dynamic) basically is limited only by the detector if any integration time is allowed. With an output swing of 1.5 V and a noise level of some 400 μV_{rms}, the range of the signal voltage is about 71 dB. The additional integration time range depends on the timing and is, for example, 54 dB (5 µs ... 2.56 µs). Thus the global dynamic range included in the signal and timestamp amounts to 125 dB. An exponential timing leads to a quasi-logarithmic compression without the drawbacks of conventional logarithmic sensors discussed in Section 9.3.2.

To evaluate the dynamic range the sensor was illuminated by four spots that cover an illumination range of 94 dB altogether (Fig. 9.25a).

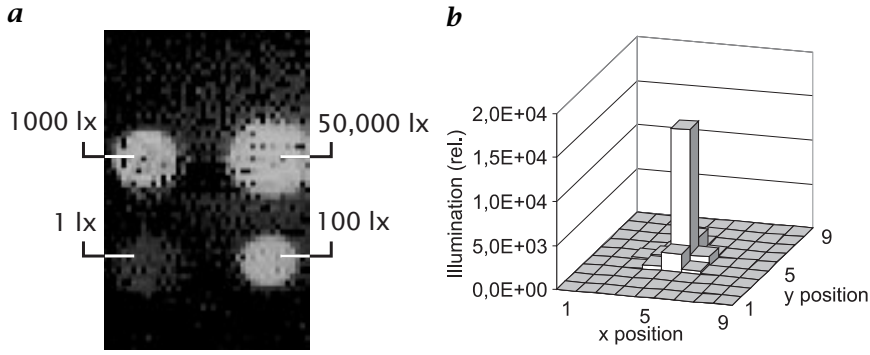


Figure 9.25: ***a*** *Evaluation of dynamic range of LARS array;* ***b*** *evaluation of local contrast.*

For this measurement timing was designed to allow adaptation to one of nine integration times, ranging from 5 μs to 2.56 ms separated by factors of two. The pixels under the 1 lx and 100 lx spots selected the longest integration time of 2.56 ms, whereas the pixels under the 1000 lx and 50,000 lx spot adapted to 320 μs and 10 μs, respectively. The image shown in the figure reproduces the integrated signal only so that the spots show approximately equal brightness except for the 1 lx spot.

A simple method to evaluate blooming effects in an image sensor array is to illuminate a single pixel with high intensity through an optical fiber and to chart the photoresponse of the pixels. Figure 9.25b depicts the result for the LARS array, for which both the integrated signal and timestamp have been taken into account. The chart demonstrates that the array is virtually free of blooming, because the photoresponse drops significantly outside the illuminated central pixel, which is saturated at an intensity of over 300,000 lx. The slightly raised signals of the adjacent pixels are mainly attributed to light scattering from the fiber cladding, thus the actual local contrast is still higher.

9.5 TFA array concepts

9.5.1 TFA color sensor for single flash illumination

The possibility of depositing thin-film detectors with adjustable spectral sensitivity will preferably lead to a 3-color-pixel design in TFA technology. This inherently allows smaller pixels even if every pixel is equipped with three separate information storage and readout units. Such a pixel architecture is well suited for the identification of areas of the same or similar color in automotive systems (color tracking) as well as for single-shot flash exposure in still cameras.

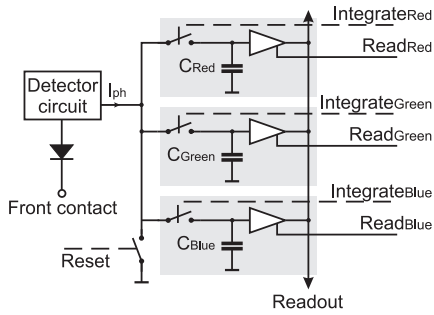


Figure 9.26: *Block diagram of CAESAR pixel.*

Table 9.1: *Timing of CAESAR pixel*

Event	Expected	Future perspective
Switching delay blue	10 ms, during preceding readout phase	
Delay after switching on illumination	300 μs	200 μs
Integrate blue	300 μs	150 μs
Switching delay green	300 μs	200 μs
Integrate green	200 μs	200 μs
Switching delay red	300 μs	200 μs
Integrate red	150 μs	750 μs
Total	1550 μs	925 μs

Figure 9.26 shows the pixel block diagram of CAESAR. The photocurrent is fed into one of the color integration circuits, one at a time, during the integration phase. Thus, the circuitry is able to generate and store the complete RGB information inside each pixel without intermediate readout operation. For readout the integrated color voltages are applied sequentially to the column output line. Simulations of the CAESAR pixel circuitry show excellent results with a high linearity over more than three decades. Control of spectral sensitivity is carried out globally by varying the front contact voltage.

The minimum time required to integrate all three colors is an important design issue. Table 9.1 gives an overview of the expected values based on current measurements and simulations. The most time-consuming step of switching from red to blue sensitivity is done during the readout phase such that the frame time is not unnecessarily extended. After further research on the thin-film system and deeper

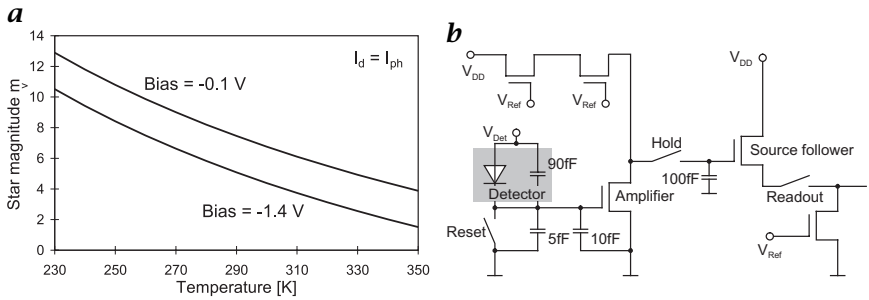


Figure 9.27: *The star tracker:* ***a*** *detectable star magnitudes; and* ***b*** *pixel circuit.*

optimization of the pixel circuitry a total time of down to 925 μs will probably be achieved, which is sufficient for single-shot flash operation.

9.5.2 TFA star tracker

Imagers for space conditions inherently require special qualities such as radiation hardness. Furthermore, in case of satellite attitude determination, stars with illumination intensities in the millilux or microlux range have to be detected. Nevertheless, in contrast to terrestrial observatories, short irradiation times due to the satellite movement must be taken into account. A TFA concept for a star tracker has been developed, employing a radiation hard *silicon on insulator* (SOI) technology for the ASIC and a detector made of a-Si:H which proves to be more radiation resistant than crystalline materials [30, 31, 32].

A suitable design for the star tracker has to handle extremely low photocurrents in the femtoampere range and to convert the integrated charge to a voltage [29]. The dark current can be minimized by operating the detector close to short-circuit condition and by cooling, which becomes obvious from Fig. 9.2a and b, respectively. Figure 9.27a demonstrates the visual star magnitudes that can be detected with the proposed star tracker, taking into account limitations due to dark current. The specified minimum detectable magnitude is 4.75, which is achieved at -0.1 V bias without or with moderate cooling.

As can be seen in the circuit diagram in Fig. 9.27b, early amplification within the pixel is employed in order to minimize noise and leakage current influence. The amplifier is designed as a simple cascode inverter, thus its input capacitance is minimized. The effective integration capacitance therefore is determined mainly by the detector blocking capacitance. More sophisticated pixel circuitries such as a current mirror similar to the one depicted in Fig. 7.9 provide higher gain, but also aggravate leakage current effects. A source follower with a common load for each column serves as driver element of the active

pixel. In the depicted pixel variant, the voltage at the inverter output is sampled and held until the following integration period, hence the integration time is not decreased by the time required for readout. Alternatively, the sample circuit is omitted if correlated double sampling is employed. In this case the reset level and the subsequent signal are read out, and the difference of these two signals is generated outside the pixel. In this way reset noise and *fixed-pattern noise* are efficiently suppressed, while flicker noise with its low cut-off frequency is significantly reduced due to the high-pass characteristic of the CDS procedure.

9.5.3 Hybrid a-Si:H/x-Si detector

The sensitivity of a space-based star sensor has to be as high as possible in order to obtain a sufficient signal voltage. In a TFA sensor such as the one discussed in Section 9.5.2 the photogenerated charge results in a gate voltage V of the pixel amplifier:

$$V = \frac{Q}{C_{\mathrm{det}} + C_{\mathrm{in}}} \tag{9.5}$$

For usual pixel sizes the detector capacitance C_{det} is significantly larger than the input capacitance C_{in} of the pixel circuit. However, as the detector area is decreased in order to minimize C_{det}, the sensitivity and, therefore, Q decrease by the same factor. As a result, the signal voltage is inherently limited by the detector technology.

Unlike the star sensor outlined in Section 9.5.2, the ***HY**brid **DE**tector* (HYDE) employs a charge storage principle similar to a CCD. Thus the charge/voltage conversion is determined only by the readout stage that can be designed to provide a low capacitance [29, 33]. As can be seen in Fig. 9.28 the a-Si:H thin-film system is deposited on the rear side of a thinned ASIC wafer that faces the incident light. The charge carriers are collected in a MOS capacitance, which works in the same way as a photogate in a CCD pixel. Photogeneration takes place in the a-Si:H as well as—for longer wavelength photons—in the x-Si material. Therefore, the advantages of both materials are combined to some degree. It is obvious that this principle can also be employed to realize a hybrid color detector, where the x-Si provides sensitivity for red light.

The HYDE offers the advantages of an increased charge amplification and additional x-Si photosensitivity compared to conventional TFA sensors. Moreover, the HYDE concept includes the substantial benefits of TFA over CCD, because it allows CMOS circuit design and manufacturing in a standard technology. A similar pixel design as in Fig. 9.27b can be employed for the HYDE. However, among other performance restrictions of a standard process, the ASIC wafer has to be thinned in order to achieve a depletion region extending down to the amorphous

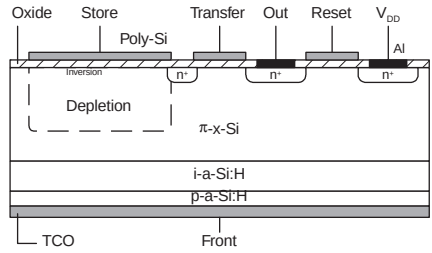


Figure 9.28: *Cross section of the hybrid sensor HYDE.*

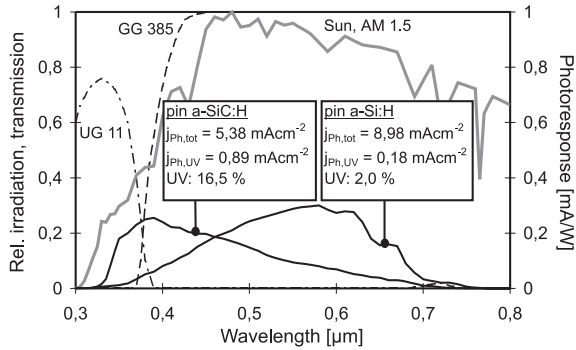


Figure 9.29: *Spectral responses of conventional and carbonized pin photodiodes in comparison to transmission of UV and visible light filters.*

p-layer. Further research is required to determine whether the higher sensitivity justifies the increased technological expenditure.

9.5.4 UV detector

Due to its wide bandgap, amorphous hydrogenated silicon is an appropriate material for the fabrication of *ultraviolet radiation* (UV) detectors. Compared to conventional pin diodes, carbonization and layer thickness variation allow greater UV sensitivity. Figure 9.29 depicts the photoresponses of a conventional a-Si:H pin photodiode and an a-SiC:H pin device for near UV detection. To exclude visible light, which is inherently dominant for natural illumination, a combination of filters is required. In this example, a UG11 is employed to suppress visible light; however, the filter exhibits some transmittivity for near IR radiation. A reference measurement with the UG11 and GG385 filters is necessary to also eliminate this IR portion. Further optimization leads to a device with a maximum spectral response for UV radiation in the range of 200 nm to 300 nm and no measurable sensitivity for visible light [34]. In this way no filters are required to suppress visible light in a UV

detector. The single UV detector can easily be assembled to an array in TFA technology. Possible applications range from solar UV irradiance monitoring to chemical analysis and medical applications.

9.6 Conclusions

Advanced imaging applications create a need for image sensor arrays with improved performance and specialized functions. It is obvious that these requirements can be fulfilled by versatile CMOS based imagers rather than by CCDs. TFA image sensors can be regarded as members of the CMOS imager family, as the circuitry of both types is based on conventional CMOS design. However, TFA sensors have the advantages of independent detector and circuitry optimization and almost 100 % fill factor. Moreover, the application range of TFA is wider due to the a-Si:H based thin-film detectors that can be optimized for visible light, UV or IR detection and three color recognition in one pixel. A disadvantage of a-Si:H detectors is their transient behavior, which is too slow for applications with frame periods in the microsecond range. As the feature sizes of CMOS technologies decrease, the advantages of TFA become more pronounced, because the a-Si:H detectors are not affected by device scaling. In this way it is ensured that TFA keeps pace with the development of ASIC technology.

Cost is important for both series production of multipurpose devices and small-scale production for highly specialized applications. Due to their higher fill factor, TFA sensors are significantly smaller than comparable CMOS devices. As the additional expenses for thin-film deposition are far lower than the ASIC costs, TFA fabrication is less expensive overall than CMOS. Enhanced ASIC yield due to smaller die sizes leads to an additional cost reduction, while the yield of thin-film deposition is close to 100 % in an ultrahigh-vacuum PECVD cluster system.

Several TFA prototypes with increasing complexity and different optimization criteria have been fabricated and tested so far. While the simpler devices may be produced in a large number for multipurpose use, highly complex pixel circuitries can be designed in order to implement application-specific functions. As the most advanced functions to date, locally autoadaptive sensitivity control serves to expand the dynamic range beyond the limitations of other circuit concepts or technologies. Problems due to exponential amplification of fixed-pattern noise and temperature differences such as in logarithmic sensors do not arise because the autoadaptivity is determined by the timing.

The TFA concepts described in the foregoing are being pursued further with regard to a series production. It is expected that TFA market introduction is little impaired by existing CCD and emerging CMOS

sensors, as TFA employs standard technologies. Moreover, TFA technology may be an incentive for novel applications of image sensors. The basic TFA concept provides two independent design flows for the thin-film detector and the ASIC. Based on the customer's specification, circuit design, thin-film optimization and processing are performed by the TFA manufacturer, while the ASIC fabrication is done by an ASIC supplier. Market-ready TFA products are expected to be released within two years.

Acknowledgment

The authors appreciate the cooperation of F. Blecher, A. Eckhardt, K. Seibel and J. Sterzel of the Institute for Semiconductor Electronics, University of Siegen and S. Benthien, H. Keller, T. Lulé and M. Sommer of Silicon Vision GmbH. The authors also wish to thank R. C. Lind, L. Humm, M. Daniels, N. Wu and H. Yen of Delphi Delco Electronics Systems, U. Efron of Hughes Research Laboratories, F. Librecht and B. van Uffel of AGFA-Gevaert N.V., C.-D. Hamann and B. Zerbe of Adam Opel AG and E. Roth of Daimler-Benz Aerospace Jena Optronik GmbH for useful discussions and technical support.

9.7 References

[1] Kemeny, S. E., Eid, E.-S., Mendis, S., and Fossum, E. R., (1991). Update on Focal-Plane Image Processing Research, Charge-Coupled Devices and Solid-State Optical Sensors II. *Proc. SPIE*, **1447**:243–250.

[2] Mendis, S., Kemeny, S., Gee, R., Pain, B., Staller, C., Kim, Q., and Fossum, E., (1997). CMOS active pixel image sensors for highly integrated imaging systems. *IEEE J. Solid-State Circ.*, **32**:187–197.

[3] Fischer, H., Schulte, J., Giehl, J., Böhm, M., and Schmitt, J. P. M., (1992). Thin Film on ASIC—a Novel Concept for Intelligent Image Sensors. *Mat. Res. Soc. Symp. Proc.*, **285**:1139–1145.

[4] Giehl, J., Stiebig, H., Rieve, P., and Böhm, M., (1994). Thin film on ASIC (TFA)-color sensors. In *New Applications of Optical Thin Film Detectors*, G. Hecht and J. Hahn, eds., pp. 560–563. Oberursel: DGM Informationsgesellschaft Oberursel mbH.

[5] Schulte, J., (1996). *Intelligente Bildsensoren in TFA-Technologie am Beispiel eines Äquidensitenextraktors*. PhD thesis, Universität-GH Siegen.

[6] Fischer, H., (1996). *Ein analoger Bildsensor in TFA (Thin Film on ASIC)-Technologie*. PhD thesis, Universität-GH Siegen.

[7] Fischer, H., Schulte, J., Rieve, P., and Böhm, M., (1994). Technology and performance of TFA (Thin Film on ASIC)-sensors. *Mat. Res. Soc. Symp. Proc.*, **336**:867–872.

[8] de Cesare, G., Irrera, F., Lemmi, F., and Palma, F., (1995). Amorphous Si/SiC three-color detector with adjustable threshold. *Appl. Phys. Lett.*, **66 (10)**:1178–1180.

[9] Eberhardt, K., Neidlinger, T., and Schubert, M. B., (1995). Three-color sensor based on amorphous n-i-p-i-n layer sequence. *IEEE Trans. Electron Devices*, **42 (10)**:1763–1768.

[10] Stiebig, H., Giehl, J., Knipp, D., Rieve, P., and Böhm, M., (1995). Amorphous silicon three color detector. *Mat. Res. Soc. Symp. Proc.*, **377**:815–826.

[11] Tsai, H.-K. and Lee, S.-C., (1988). Amorphous SiC/SiC three-color detector. *Appl. Phys. Lett.*, **52 (4)**:275–277.

[12] Zhu, Q., Stiebig, H., Rieve, P., Fischer, H., and Böhm, M., (1994). A novel a-Si(C):H nolor sensor array. *Mat. Res. Soc. Symp. Proc.*, **336**:843–848.

[13] Zhu, Q., Coors, S., Schneider, B., Rieve, P., and Böhm, M., (1998). Bias sensitive a-Si(C):H multispectral detectors. *IEEE Trans. Electron Devices*, **45(7)**:1393–1398.

[14] Giehl, J., Zhu, Q., Rieve, P., and Böhm, M., (1996). Transient behavior of color diodes. *Mat. Res. Soc. Symp. Proc.*, **420**:159–164.

[15] Rieve, P., Giehl, J., Zhu, Q., and Böhm, M., (1996). a-Si:H photo diode with variable spectral sensitivity. *Mat. Res. Soc. Symp. Proc.*, **420**:135–140.

[16] Zhu, Q., Sterzel, J., Schneider, B., Coors, S., and Böhm, M., (1998). Transient behavior of a-Si(C):H bulk barrier color detectors. *Jour. Applied Physics*, **83(7)**:3906–3910.

[17] Street, R. A., (1990). Thermal generation currents in hydrogenated amorphous silicon p-i-n structures. *Appl. Phys. Lett.*, **57 (13)**:1334–1336.

[18] Zhu, Q., Stiebig, H., Rieve, P., Giehl, J., Sommer, M., and Böhm, M., (1994). New type of thin film color image sensor, sensors and control for automation. *SPIE Proc.*, **2247**:301–310.

[19] Blecher, F., Seibel, K., and Böhm, M., (1998). Photo- and Dark Current Noise in a-Si:H pin Diodes at Forward and Reverse Bias. Presented at MRS Spring Meeting, San Francisco.

[20] Wieczorek, H., (1995). 1/f noise in amorphous silicon nip and pin diodes. *J. Appl. Phys.*, **77 (7)**:3300.

[21] Blecher, F. and Seibel, K., (1997). Simulation und experimentelle Verifikation von statistischen Kenngrößen und Rauschmodellen a-Si:H basierter optischer Sensoren. DFG-Abschlußbericht Bo 772/3-1.

[22] Boudry, J. M. and Antonuk, L. E., (1993). Current-noise-power spectra for amorphous silicon photodiode sensors. *Mat. Res. Soc. Symp. Proc.*, **297**: 975–980.

[23] Böhm, M., Blecher, F., Eckhardt, A., Schneider, B., Benthien, S., Keller, H., Lulé, T., Rieve, P., Sommer, M., Lind, R. C., Humm, L., Daniels, M., Wu, N., and Yen, H., (1998). High Dynamic Range Image Sensors in Thin Film on ASIC-Technology for Automotive Applications. Presented at Advanced Microsystems for Automotive Applications, Berlin.

[24] Wong, H.-S. P., (1997). CMOS image sensors—recent advances and device scaling considerations. *Tech. Digest IEDM*, **97**:201–204.

[25] Schulte, J., Fischer, H., Lulé, T., Zhu, Q., and Böhm, M., (1994). Properties of TFA (Thin Film on ASIC) sensors. In *Micro System Technologies ´94*, A. H. H. Reichl, ed., pp. 783–790. Berlin: VDE-Verlag.

[26] Schneider, B., Fischer, H., Benthien, S., Keller, H., Lulé, T., Rieve, P., Sommer, M., Schulte, J., and Böhm, M., (1997). TFA image sensors: From the one transistor cell to a locally adaptive high dynamic range sensor. *Tech. Digest. IEDM*, **97**:209–212.

[27] Lulé, T., Fischer, H., Benthien, S., Keller, H., Sommer, M., Schulte, J., Rieve, P., and Böhm, M., (1996). Image sensor with per-pixel programmable sensitivity in TFA technology. In *Micro System Technologies ´96*, A. H. H. Reichl, ed., pp. 675–680. Berlin: VDE-Verlag.

[28] Chen, S. and Ginosar, R., (1995). Adaptive sensitivity CCD image sensor. *Proc. SPIE*, **2415**:303–309.

[29] Böhm, M., Lulé, T., Fischer, H., Schulte, J., Schneider, B., Benthien, S., Blecher, F., Coors, S., Eckhardt, A., Keller, H., Rieve, P., Seibel, K., Sommer, M., and Sterzel, J., (1998). Design and Fabrication of a High Dynamic Range Image Sensor in TFA Technology, to be presented at 1998 VLSI Circuits Symposium, Honolulu.

[30] Hollingworth, R. E. and J. Xi, A. M., (1989). Proton and neutron damage in thick amorphous silicon diodes. *Mat. Res. Soc. Symp. Proc.*, **149**:655–659.

[31] Woodyard, J. R. and Landis, G. A., (1991). Radiation resistance of thin-film solar cells for space photovoltaic power. *Solar Cells*, **31**:297–329.

[32] Boudry, J. M. and Antonuk, L. E., (1994). Radiation damage of amorphous silicon photodiode sensors. *IEEE Trans. Nuclear Science*, **41(4)**:703–707.

[33] Schneider, B., Blecher, F., Eckhardt, A., Seibel, K., Sterzel, J., Böhm, M., Benthien, S., Keller, H., Lulé, T., Rieve, P., Sommer, M., Librecht, F., and van Uffel, B., (1998). TFA Image Sensors—A Survey with Regard to Possible Applications, presented at OPTO 98, Erfurt, Germany.

[34] Caputo, D., de Cesare, G., Irrera, F., and Palma, F., (1996). Solar-blind UV photodetectors for large area applications. *IEEE Trans. Electron Devices*, **43 (9)**:1351–1356.

10 Poly SiGe Bolometers

S. Sedky[1,2] and P. Fiorini[3]

[1]IMEC, Leuven, Belgium
[2]Faculty of Engineering, Cairo University, Giza, Egypt
[3]Dep. of Physics, III University of Rome, Italy

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

10.1 Overview

The evolution of *infrared* (IR) sensors started during World War II, when they were used mainly for *night vision* [1]. This application pushed thermal imaging technology towards high spatial and temporal resolution and its use was extended to other fields such as fire control and search track. Later, the spatial resolution of the detector was further improved for scientific applications that included remote sensing of earth resources [2] and astronomical exploration.

It is of some interest to follow the different development stages of IR technology with time, as this sheds light on the limitations of different technologies and on how they were overcome. We consider, for instance, early IR detectors, such as Golay pneumatic detectors [3], radiation thermopiles [4], bolometers [5], and pyroelectric detectors [6]. Because the intensity of the IR radiation is deduced from the temperature increase that it generates in the detector active element, these devices are known as *thermal detectors*. The frequency response of these detectors was limited by the large thermal mass of the active element; thus, they could not be used to produce a usable image with raster scanning techniques. Hence, it appeared that the performance of such detectors was limited by the laws of physics. Moreover, the attachment of electrical leads, used to transfer the signal generated by the detector to the electrical detection circuits, formed a highly conductive thermal path that seriously degraded the sensitivity of the detector. As a consequence, the use of thermal detectors was limited.

The development of semiconductor materials in the 1950s and 1960s provided photoconductor quantum detectors capable of sensing photons with high electrical bandwidth. These detectors were used mainly in military applications and, hence, they were designed specifically for *near IR* (1 to 3 μm), *midwave IR* (MWIR) (3 to 5 μm) and *longwave IR* (LWIR) (8 to 12 μm) regions. The response time of photo conductor quantum detectors is determined by the free carrier lifetime and is usually of the order of microseconds. Thus, it was possible to integrate these detectors in arrays used for parallel scanned imaging systems. Such arrays need cryogenic cooling to improve the threshold sensitivity and to reduce the fundamental noise of the detector. Moreover, high spatial resolution, which requires a large number of detectors, was not achievable as the power dissipated in each detector, and the heat load due to electrical wiring was too large. Later, development of PN junction diode devices noticeably reduced the bias power dissipation, as these devices act as a photovoltaic IR detector having high impedance. In spite of this improvement, the number of detector elements was still limited by the number of interconnection leads that could be used. This problem was solved by the development of *charge-coupled devices* (CCD) and *charge injection devices* (CID), which

provided the possibility of multiplexing the IR detectors of a *focal plane array* (FPA) and allowed packaging a large number of detectors in a small practical sensor. Sensors, based on this technology are now available to cover the IR spectrum, with Si-based devices for the near IR (e. g., *platinum silicide* (Pt-Si) [7], *indium antimonide* (InSb) for MWIR [8] and *mercury cadmium telluride* (HgCdTe) for LWIR [9]). Such detectors have quantum efficiencies approaching the theoretical limit.

As quantum detectors must operate at low temperature, they are inserted in a radiative cooler, which is bulky, heavy, and delicate. This fact causes important logistic problems, especially in military and space applications. Furthermore, a radiative cooler is expensive and this limits the amplitude of the market for detection systems. These drawbacks motivated the development of *uncooled detectors*.

Both quantum and thermal uncooled detectors have been realized. Examples of the first type include *lead sulfide* (PbS) and *lead selenide* (PbSe) detectors. Although they are somewhat slow, when combined with modern readout electronics they have a sufficient bandwidth for imaging operation with high sensitivity in the 1–5 μm region [10]. The most important advances in the field of IR uncooled detectors have been achieved with thermal detectors. The development of micromachining techniques [11] has allowed realization of detectors of small size (50 μm and below), with low thermal capacity and large thermal insulation. The combination of these latter two features gives high responsivity and reasonably large cut off frequency. Several types of thermal detectors operating at room temperature have been realized. *Pyroelectric* and *ferroelectric detectors* have been developed by Texas Instruments [12, 13, 14, 15, 16]. These detectors are essentially thermally modulated electrical capacitors and can be operated uncooled or in a temperature-stabilized mode. Their large detectivity compensates for some drawbacks, such as the need of chopping IR radiation and the difficulty of integration with the driving electronics on a single chip (detectors must be bump-bonded to the silicon substrate [17]). These drawbacks are not present in FPAs of resistor *microbolometers* developed by Honeywell [18] and by the Australian organization DSTO [19, 20, 21, 22, 23, 24]. They do not require modulated IR radiation and can be integrated monolithically with the driving electronics. The high frame rate capability of these microbolometer FPAs makes them well-suited for automated moving target detection and tracking. Their low cost and the low maintenance that they require enlarge the market of IR cameras to a wide variety of civilian applications, including security, automotive areas, rescue, and fire control.

In the following sections we shall focus on *bolometers*. In Section 10.2, a general description of the principle of operation of bolometers will be presented. This will include the different factors contributing to the temperature rise of the bolometer, the derivation of the re-

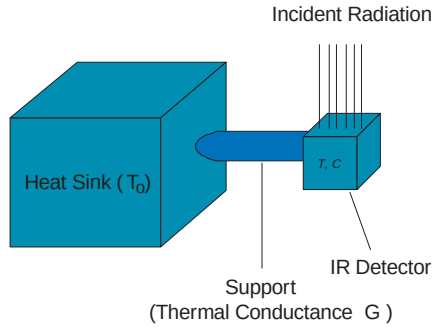


Figure 10.1: *Basic structure of an IR thermal detector.*

sponsivity, and the discussion of the different sources of noise affecting the performance of the device. A model describing the performance of micro bolometer arrays will be presented in Section 10.3. In Section 10.4, we shall discuss the different materials used for fabricating bolometers. *Poly SiGe* will be introduced as a new material suitable for this application and its advantages will be discussed in detail. The detailed process for realizing poly SiGe bolometers will be given in Section 10.5. As the active elements of bolometers are, in general, transparent to IR, an absorbing layer must be used. Different types of absorbers will be discussed in Section 10.5.2. In Section 10.6 a complete characterization of the realized bolometers will be given. This will include the electrical, thermal, mechanical, and optical properties of the device. Moreover, the performance of poly SiGe microbolometer FPAs will be modeled and compared to those obtained with the available technologies and materials.

10.2 Principle of operation of bolometers

A bolometer is a thermistor IR detector. The active element is a resistor with a very small thermal capacity C and a large *temperature coefficient of resistance* (TCR) and, hence, a fast and significant change occurs in the resistance when the detector is heated by the incident radiation. The resistor must be thermally insulated to obtain large temperature variations, even with small incident power. The basic structure of a bolometer, or more generally of a thermal IR detector, is shown in Fig. 10.1. The bolometer is represented by a thermal capacity (C) and it is connected to an infinite heat sink by a support having a thermal conductance (G). In the absence of both external radiation and applied bias, the temperature of the bolometer is the same as that of the heat sink.

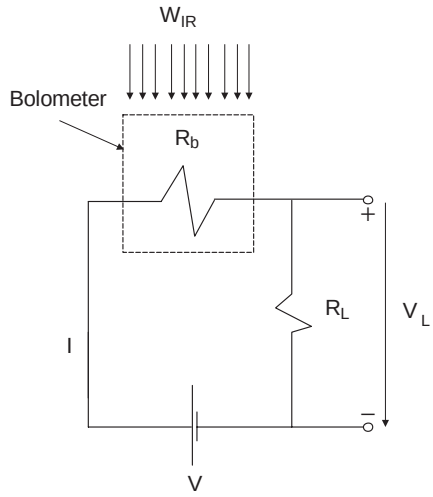


Figure 10.2: *Biasing circuit of a bolometer.*

10.2.1 Thermal behavior

In this section, we analyze briefly the factors that make the bolometer temperature larger than that of the heat sink. In general, the circuit shown in Fig. 10.2 is used to bias the bolometer. The voltage source V generates a current I that flows through the circuit; as a result, there is power dissipation and the bolometer heats up. Moreover, the absorbed IR radiation will also change the temperature of the bolometer. The thermal balance is described by

$$W = C\frac{\mathrm{d}T}{\mathrm{d}t} + G(T - T_0) \tag{10.1}$$

This equation simply states that a part of the power W (absorbed IR plus dissipated electrical power) is used to heat the bolometer (first term in RHS) and the rest flows towards the thermal sink (second term in RHS).

First, we shall discuss the dependence of temperature on the incident radiation. In general, the incident radiation varies with time and can be expressed as $W = W_0 + W_\omega e^{j\omega t}$, which is a superposition of a constant component and a time-dependent component. The power absorbed by the bolometer is $\tilde{\epsilon}W$, where $\tilde{\epsilon}$ is the emissivity of the bolometer. The temperature increase is $\Delta T = W_0/G + \Delta T_{ac}$. The first term is due to the constant power W_0 while the second term is due to the time dependent power and is expressed as [25]

$$\Delta T_{ac} = \frac{\tilde{\epsilon} W_\omega e^{j\omega t}}{\sqrt{G^2 + \omega^2 C^2}} = \frac{\tilde{\epsilon} W_\omega e^{j\omega t}}{G\sqrt{1 + (\omega\tau)^2}} \tag{10.2}$$

where $\tau = C/G$ is the *time constant* of the bolometer. It should be noted that in most cases the frequency ω is relatively small and consequently $\omega C \ll G$. It is clear from Eq. (10.2), that for a given radiation, the temperature increase of the detector can be maximized by reducing the thermal conductance G. The minimum possible value of G is that obtained when the only thermal coupling of the detecting element to the heat sink is via radiative exchange. This means that the detector is kept in vacuum and does not have any contact with the surrounding media. Such value can be estimated from Stefan-Boltzmann total radiation law and it is of the order of 10^{-8} W/K for a 50 μm $\times$ 50 μm detector. In practical situations, the thermal conductance is higher than this value as the detector must be connected to an infinite heat sink to transfer the signal to the driving electronics and to be supported. This contribution to the thermal conductance is about one order of magnitude larger than the above "radiative limit" in optimized detectors. It can be minimized by using, for the supports, low thermal conductivity materials, by increasing the length of the supports, and by reducing their thickness and width. The last contribution to the thermal conductance is the heat lost to the surrounding media by conduction and convection. This component can be eliminated by operating the device under vacuum. While minimizing thermal conductance, the thermal time constant increases, so care must be taken to keep it below an upper limit, determined by the frame rate.

The second contribution to the temperature increase of the bolometer is due to biasing. The electrical power dissipated in the bolometer is $W_{ele} = V^2 R_b(T)/(R_b(T) + R_L)^2$, where $R_b(T)$ is the bolometer resistance, which as a function of the temperature T is expressed as $R_b(T) = R_b(T_0)e^{-\alpha T/T_0(T-T_0)}$, where α is the temperature coefficient of resistance (TCR) and $R_b(T_0)$ is the bolometer resistance at room temperature.

In general, a bolometer can be biased in two ways: either by using a short ($T \ll \tau$) voltage pulse or a dc voltage. In most applications pulsed bias is used and hence we shall consider it first. In this case, the bolometer will experience a slight temperature rise ($T \cong T_0$) and, consequently, the second term in Eq. (10.1) can be neglected. Furthermore, the term $\alpha T/T_0$ can be considered constant. Using these approximations, Eq. (10.1) can be reduced to

$$\Delta T = T - T_0 = \frac{V^2 t R(T_0)}{C\,(R(T_0) + R_L)^2} = \frac{W_{ele} t}{C} \tag{10.3}$$

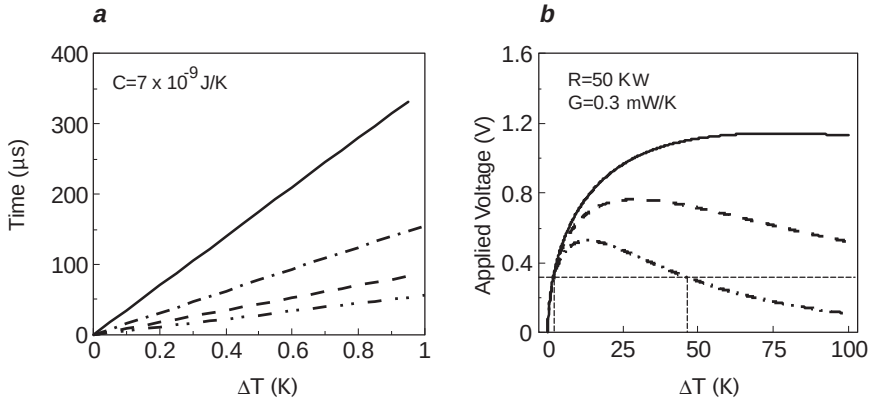


Figure 10.3: ***a*** *Time required to obtain a temperature increase ΔT at different pulsed bias (Solid line 2 V, dashed-dotted line 3 V, dotted line 4 V, dashed-double-dotted line 5 V). The value of the thermal capacitance is also reported in the figure.* ***b*** *Applied dc bias vs the temperature increase that it generates in the bolometer. Different curves refer to different activation energies. (Solid line $E_a = 0.16\,eV$, dotted line $E_a = 0.32\,eV$, dashed dotted line $E_a = 0.64\,eV$.) Values of the resistance and thermal conductance used in the calculations are reported in the figure.*

Fig. 10.3a displays the time interval required to increase the temperature of the bolometer by 1 K for different biasing voltages. The different parameters in this plot are $\alpha = 0.02\,\mathrm{K}^{-1}$, $R_b(T_0) = R_L = 50\,k\Omega$, and $C = 7\times10^{-9}$ J/K. As in actual operation the duration of the voltage pulse is few μs ($T < 10\,\mu$s), it is obvious that the use of pulsed bias allows applying large voltages without appreciable heating.

On the other hand, when using dc bias, the temperature increase is large and the change of TCR with temperature must be taken into account. It is then more convenient to express the TCR in terms of the activation energy E_a, which is temperature independent [26] and is related to the TCR by $E_a = kT^2\alpha$. The associated temperature increase can be evaluated directly from Eq. (10.1) at steady state. This yields the following expression

$$\Delta T = \frac{W}{G} = \frac{V^2 R_b}{G\left(R_b(T) + R_L\right)^2} \tag{10.4}$$

To maximize the responsivity of the bolometer, the load resistor R_L must be equal to the bolometer resistance at the operation temperature, hence, we assume $R_L = R_b(T)$. Equation (10.4) is plotted in Fig. 10.3b for $G = 3\times10^{-7}W/K$, $R_b(T_0) = 50\,k\Omega$ and for different activation energies. It is evident that for high activation energies, there is a maximum

voltage that can be applied to the bolometer. It is also clear that the two temperatures T_1 and T_2 correspond to the same voltage. These two temperatures also correspond to two different resistances R_1 and R_2. T_1 (or T_2) is reached depending on whether $R_L = R_1$ (or R_2). Applying a dc bias voltage, the bolometer will heat up and the TCR will decrease (we recall that $\alpha = E_a/kT^2$) with a consequent reduction of the bolometer sensitivity. Thus, it is always recommended to use pulsed bias.

10.2.2 Responsivity

The performance of IR detectors is expressed in terms of *responsivity*, noise, and *signal-to-noise ratio* (SNR). The responsivity is the signal generated per unit incident power. To derive the responsivity of a bolometer, we refer again to Fig. 10.2. We suppose first that no IR power impinges the detector (dark conditions). $R_{b(d)}$ will indicate the bolometer resistance at the temperature determined by the bias and without IR power. In this case, the voltage drop across R_L will be

$$V_{\text{dark}} = \frac{V}{(R_{b(d)} + R_L)} \tag{10.5}$$

When a steady infrared power, W_{light}, impinges the detector, the bolometer temperature will increase by $\Delta T_{\text{light}} = \tilde{\epsilon} W_{\text{light}}/G$. The time required to reach this temperature increase is the bolometer time constant (τ) and falls typically in the range of few milliseconds. This temperature rise results in changing the resistance of the bolometer to $R_{b(l)}$, which in turn changes the voltage measured across the terminals of the load resistance; this voltage, indicated by V_{light}, is

$$V_{\text{light}} = \frac{V}{(R_{b(l)} + R_L)} R_L \tag{10.6}$$

We define the signal generated by the incident radiation as

$$S = V_{\text{dark}} - V_{\text{light}} \tag{10.7}$$

Taking into account that the bolometer resistance with and without radiation can be related to each other by

$$R_{b(l)} = R_{b(d)} + \frac{\mathrm{d}R}{\mathrm{d}T} \Delta T_{\text{light}} \tag{10.8}$$

the responsivity can be expressed as

$$R = \frac{S}{W} = \left(\frac{(\mathrm{d}R_b/\mathrm{d}T)\tilde{\epsilon}}{G(R_b + R_L)^2} \right) V R_l = \frac{\alpha \tilde{\epsilon} V R_b R_L}{G(R_b + R_L)^2} \tag{10.9}$$

where we have used the relation $\alpha = (1/R_b)(dR_b/dT)$. α is a decreasing function of temperature and its value at the operating temperature must be used. It can be concluded from Eq. (10.9) that the responsivity varies linearly with the biasing voltage; this is true only for pulsed bias, when the TCR is nearly constant. Using dc bias, the increase of voltage will be compensated by the decrease in TCR and the performance will have an upper limit.

10.2.3 Sources of noise

Random noise plays an important role in the performance of bolometers as it determines the minimum power that can be detected. The noise sources may arise in the bolometer, in the incident radiation, or in the electronic circuitry associated with the detection system. In this section we shall discuss only the noise associated with the bolometer. The major noise sources in bolometers are the thermal conductance noise, the Johnson noise, and the low frequency $(1/f)$ noise.

We start by discussing fluctuations in the thermal conductance. As described in the previous section, in absence of an external radiation the temperature of the bolometer is the same as that of the heat sink (T_0). This means that on average the heat flow from the bolometer to the heat sink is completely balanced by the heat flow in the opposite direction. This is true only on the average, as there are instantaneous fluctuations of the power flowing through the support into the detector. Their *root mean square* (RMS) is given by $\Delta W_{th} = \sqrt{4kT^2G\Delta f_{th}}$ [27], where Δf_{th} is the thermal noise bandwidth and is determined by the inverse of the thermal time constant τ. These fluctuations in the power generate fluctuations in the voltage V_{th}, the RMS value of which is given by

$$V_{th} = R\Delta W_{th} = R\sqrt{4kT^2G\Delta f_{th}} \tag{10.10}$$

The Johnson noise arises from the random motion of free carriers within any resistive material. The RMS of the voltage fluctuation associated with the Johnson noise is [28]

$$V_J = \sqrt{4kTR_b\Delta f_e} \tag{10.11}$$

where Δf_e is the electric bandwidth and is determined by the time interval allocated to read the signal.

In any resistor, it is possible to measure as noise a component that decreases as $1/f^{\gamma}$ (where $0 \le \gamma \le 0.5$) and which adds up to the Johnson noise. Thus, as its origin is still a matter for debate, it is well described by an empirical expression due to Hooge [29]

$$V_{n,1/f} = KV_b\sqrt{\frac{\rho}{WLtf}} \tag{10.12}$$

where K is a constant that depends on the type of material, V_b is the voltage applied to the bolometer, W, L, and t are, respectively, the width, length, and thickness of the active part of the bolometer.

The total noise of the bolometer V_n is the RMS of these three noise components and it is given by

$$V_n = \sqrt{V_{th}^2 + V_f^2 + V_{1/f}^2} \tag{10.13}$$

Normally the noise component due to fluctuations in thermal conductance is negligible. Material properties, frequency bandwidth, and bias voltage determine whether the $1/f$ noise or the Johnson noise dominates.

The performances of the detector are characterized by the ratio of voltage noise to responsivity, or more precisely by the *noise equivalent power* (NEP), which is the power required to produce a unity signal to noise ratio and is given by

$$\text{NEP} = \frac{V_n}{R} \tag{10.14}$$

10.3 Microbolometer focal plane arrays

The development of silicon-based, uncooled, *focal plane arrays* of microbolometers started after the Gulf War, as the current imaging employing visible light cameras, low-light level, image intensified, or conventional IR cameras had serious deficiencies. Visible TV cameras require well-lighted areas and can not image in darkness. Low-level light TV cameras have difficulties when operating in bright sunlight or in total darkness. Image intensifiers require some ambient light. Conventional IR cameras are costly, require an initial cool-down period and need additional power for the cooling pump or a periodic gas replenishment for long-term operation. Meanwhile, uncooled FPA operating in the 8-12-μm range combine the capability of operation in bright sunlight or total darkness, typical to IR cameras, of low cost, light weight, and ease of use.

Depending on the technology used for integrating the detectors with the readout electronics, FPA are called hybrid or monolithic. Hybrid arrays are realized by bump-bonding the detectors to a silicon microcircuit; this is the case of ferroelectric detectors [17], where monolithic

arrays are obtained by directly preparing the detector and the readout microcircuit on the same silicon chip [30]. The two principle routes of monolithic technology development are thin pyroelectric or dielectric bolometers and resistance bolometers. The former technology is less mature due to the need to develop materials fully compatible with silicon wafer processing. Meanwhile, there are many materials suitable for resistance bolometers, which can be readily integrated with VLSI microcircuits. Moreover, ferroelectric arrays have only ac response and require choppers to operate in a slowly changing scene [31]. This feature increases system complexity and decreases the SNR by a factor of 1.4, as the thermal radiation signal is blocked on every other frame basis. Also, the sensor internal frame rate is augmented by a factor of two, which yields a decrease of the SNR of an additional factor of 1.4. Resistance bolometers can be used in dc mode and do not imply choppers.

Focal plan arrays of microbolometers were first developed by Honeywell [30]. *Micromachining* techniques were used to realize large arrays (240×336) of detectors operating in the range 8–12 μm. The active element is made of *vanadium oxide*. The FPA readout is at 30 Hz and does not utilize a chopper. The array can detect a temperature difference of 0.1 K between two elements of the scene. The thermal insulation between pixels is less than -142 dB, which means that a 1000 °C target directed toward a certain pixel will have in adjacent pixels an effect that is 8×10^{-8} less. Such a feature makes blooming virtually nonexistent. Each FPA is a single chip monolithically integrated with multiplexers for simplicity and affordability. The driving electronics for such arrays are realized by bipolar transistor technology. Recent uncooled arrays developed by Honeywell and Rockwell have been integrated with CMOS circuitry [32]. These arrays have a frame rate of 60 Hz and the minimum detectable temperature difference on the scene is 0.07 K. Similar arrays, using vanadium oxide as the temperature-sensitive element, have been developed by Amber [33] and Loral Infrared and Imaging Systems [34].

10.3.1 Model describing the performance of FPA

In this subsection we briefly analyze the different quantities affecting the performance of FPA. To achieve this goal, we shall present a model describing the performance of an IR camera consisting of 640×480 pixels and operated at a frame rate of 20 Hz. The biasing circuit for each pixel is shown in Fig. 10.4; all noise sources are indicated in the figure. Here $V_{b,th}$, $V_{b,J}$ and $V_{b,1/f}$ represent, respectively, the thermal conductance noise, the Johnson noise, and the $1/f$ noise of the bolometer; in addition, $V_{L,J}$ is the Johnson noise of the load resistor and V_a the noise of the amplifier.

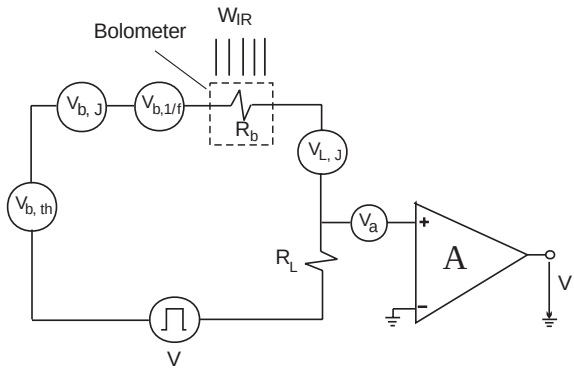


Figure 10.4: *Biasing circuit of a camera pixel. The different sources of noise are also reported in the figure.*

We first concentrate on the noise introduced by the amplifier, which has two main components: $1/f$ and Johnson. The $1/f$ noise can be expressed as [28]

$$V_{a,1/f} = S_{a,1/f} \left(ln(4 f_{\max}/f_{\min}) \right)^{1/2} \tag{10.15}$$

where $S_{a,1/f}$ depends on the type of amplifier used, f_{max} and f_{min} are the extremes of the bandwidth. The Johnson noise is given by Laker and Sansen [28]

$$V_{a,J} = S_{a,J} \sqrt{\Delta f_e} \tag{10.16}$$

where $S_{a,J} = 8kT/3g_m$, g_m is the transconductance of the amplifier and $\Delta f_e = f_{max} - f_{min}$ is the bandwidth of the amplifier.

Two solutions are available for amplifying the signal: to use a built-in amplifier for each pixel or to use one or more external amplifiers. We will briefly discuss advantages and disadvantages of the two. For a good external commercial amplifier, for example, OP-27, the value of $S_{1/f}^{1/2}$ is on the average 5 nV, while $S_{a,J}^{1/2} = 3$ nV.

An estimate of the noise characteristics of an internal amplifier can be obtained by assuming that the amplifiers are built by CMOS transistors having the following properties:

channel width $W = 10\,\mu$m,
channel length $L = 5\,\mu$m,
oxide thickness $t_{ox} = 0.2\,\mu$m,
oxide permittivity $\epsilon_{ox} = 3.45 \times 10^{-11}$ F/m, and
carrier mobility $\mu_n = 0.08$ m^2/Vs.

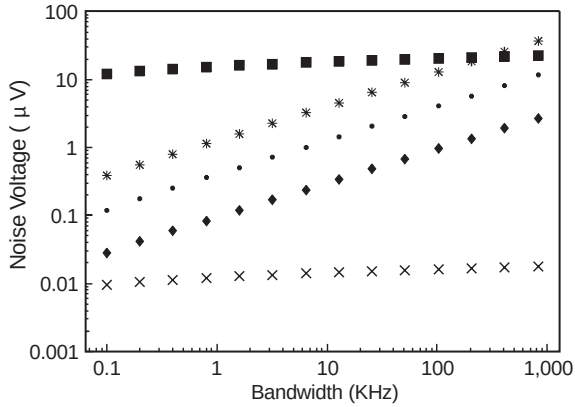


Figure 10.5: *Amplifier and bolometer noise vs system bandwidth: ♦ Johnson noise of external amplifier, × $1/f$ noise of external amplifier, • Johnson noise of internal amplifier, ■ $1/f$ noise of internal amplifier, and ∗ Johnson noise of a 100 kΩ bolometer.*

In terms of these quantities, $S_{a,1/f}$ and g_m are given by:

$$S_{a,1/f} = \left(2 \times 10^{-21}/WL\right) \quad \text{and} \quad g_m = \mu_n \frac{W}{L} \frac{\epsilon_{ox}}{t_{ox}} (V_G - V_T)$$

Inserting the numerical values, we obtain

$$S_{a,J}^{1/2} = 8.7\,\text{nV} \quad \text{and} \quad S_{a,1/f}^{1/2} = 6.3\,\mu\text{V}$$

In Fig. 10.5, the noise components of the internal and external amplifier are reported as a function of the system bandwidth. Although the internal amplifier has a large total noise, it requires a small bandwidth (around 10^4 Hz) for operation. The external amplifier has a smaller noise but it requires a larger bandwidth, given by the inverse of the time allocated to read the signal of each pixel. In order to evaluate this time, we assume that the signal is multiplexed to 16 external amplifiers and to refresh the image 20 times per s; this gives an access time of $\approx 3\,\mu$s, which corresponds to a bandwidth of about 4×10^5 Hz. From the analyses of Fig. 10.5, it is clear that in spite of the increase in bandwidth, the use of external amplifiers still gives lower noise. It must be pointed out that using external amplifiers requires that the time necessary to transfer the signal is shorter than the access time. As the access time is equal to the product of the line capacitance ($\approx 10^{-12}$ F) times the bolometer resistance, it is found that the resistance of the bolometer must be lower than $10^5\,\Omega$. This condition is not difficult to fulfil.

In Fig. 10.5, the Johnson noise of the bolometer is also reported for $R_b = 100\,\text{k}\Omega$. We see that it dominates over the noise of the amplifier if,

as suggested by the previous discussion, an external amplifier is used. We conclude then that the performances of the FPA are dominated by the noise of the bolometer more than from that of the amplifier.

Finally, the total noise generated at the input of the amplifier is expressed as

$$V_n = \left(\xi^2 (V_{b,J}^2 + V_{L,J}^2 + V_{b,1/f}^2) + V_{b,th}^2 + V_a^2\right)^{1/2} \tag{10.17}$$

where $\xi = R_L/(R_b + R_L)$ represents the fraction of the noise at the input of the amplifier.

10.3.2 Noise equivalent temperature difference

The performance of the camera is evaluated by its ability to detect slight temperature changes of the scene. If the temperature of the scene is changed from T_0 to $T_0 + \Delta T$, the power reaching the detector will be changed by [35]

$$\Delta W_{\text{det}} = \frac{A_{\text{det}}(\partial M/\partial T)\Delta T}{4f_{\#}^2} \tag{10.18}$$

where A_{det} is the area of the detecting pixel, $\partial M/\partial T$ is the thermal contrast, and $f_{\#}$ is the f-number of the imaging system defined by

$$f_{\#} = \frac{\text{focal length}}{\text{entrance of the pupil diameter}} \tag{10.19}$$

The *noise equivalent temperature difference* (NETD) is defined as the temperature change of the scene that results in a change in the detected power equal to the noise equivalent power (NEP), thus,

$$\text{NETD} = \frac{4f_{\#}^2 \text{NEP}}{A_{\text{det}}(\partial M/\partial T)} \tag{10.20}$$

This model will be used in Section 10.6 to evaluate the performance of poly SiGe FPAs and to compare them to those of vanadium oxide and metal bolometers.

10.4 Bolometer materials

In this section, we describe which physical properties make a material suitable for IR bolometers. We review the materials most commonly used and analyze the advantages of using poly SiGe.

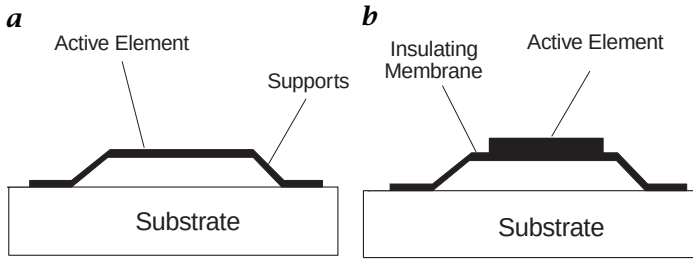


Figure 10.6: *Two possible approaches for realizing thermal insulation:* ***a*** *self suspended;* ***b*** *supported by an insulating membrane.*

10.4.1 Properties of bolometer materials

Any bolometer material should have high TCR and low $1/f$ noise. For the same TCR, materials with low resistivity must be preferred as they minimize the Johnson noise. The requirement of low thermal conductance also influences the choice of the material, with choices dependent upon the technology used for realizing the bolometer. Thermal insulation can be achieved either by micromachining the active element in the form of a self-sustained, suspended membrane, or by depositing the active element on top of a thermally insulating membrane. These two possibilities are shown schematically in Fig. 10.6.

In case of a self-sustained membrane, the supports and the active elements are made of the same material, which, besides the properties described herein, must also have low thermal conductance. Furthermore, the stress in the active element must be carefully controlled as it dramatically affects the mechanical stability of the device. High compressive stress results in buckling, the active element goes in direct contact with the substrate, and loses the thermal insulation. High tensile stress might break the active element.

For the structure shown in Fig. 10.6b, the requirements of high TCR, low noise on one side, and low thermal conductance, low stress on the other side, refer to different materials. This gives more freedom in the choice of materials, but results in a process involving a larger number of steps. It is worth noticing at this stage that deposition conditions also influence the choice of the material. Active elements that can be prepared at low temperature, and whose deposition methods are compatible with standard IC technology are preferred. These characteristics allow post-processing the bolometer on wafers already containing the driving and readout electronics.

10.4.2 Materials used for bolometers

Both metals and semiconductors have been used as active element deposited on an insulating membrane. In spite of their low TCR, metals such as gold ($TCR_0 = 0.15\,\%$)[1] [36], platinum ($TCR_0 = 0.25\,\%$) [37, 38], titanium ($TCR = 0.2\,\%$) [39, 40] have been used to provide low cost thermal imaging in industrial automation and in security and safety systems. These applications usually do not require the high infrared sensitivity demanded by military purposes. Metal bolometers are characterized by having low $1/f$ noise [39] and low thermal capacity, which means low thermal time constant ($\tau < 1$ ms [19]). The responsivity of titanium bolometers is of the order of 10^4 V/W and the maximum detectivity is 6.3×10^9 cm $\sqrt{\text{Hz}}$/W [39].

The performance of bolometers can be improved by using semiconductor materials. With respect to metals, they have a TCR about one order of magnitude larger but also larger $1/f$ noise and resistivity. To date, the best results have been obtained using *vanadium oxide*, which is an amorphous film deposited by the ion beam sputtering process where tight control of the oxygen content is maintained. High TCR can be obtained at relatively low resistivity ($-2\,\%$ for a sheet resistance of $13.5\,\text{K}\Omega$/sq for typically used thickness [41]). Noise in vanadium oxide can be reduced by reducing the void content [42]. Using an optimized material, an NETD of about 75 mK for a 324×240 array made of $50\,\mu\text{m} \times 50\,\mu\text{m}$ pixels [43] has been achieved. Such low noise level made vanadium oxide widely used in resistance bolometers and other applications requiring high TCR [44, 45, 46, 47, 48]. It should be noted that noise and TCR in vanadium oxide depend strongly on the preparation conditions, thus, a complicated optimization process is necessary to achieve the required performance level. Furthermore, the material, thus being compatible with post-processing of wafers already containing the readout electronics, is certainly not standard in IC technology and foundries with the capability of depositing vanadium oxides are not easily found. Other semiconductor materials, such as plasma chemical vapor-phase deposited amorphous silicon [49], amorphous germanium [50] and amorphous silicon carbide [51], have been used as the active element of bolometers deposited on top of an insulating membrane. These materials have TCR around $-2\,\%$, but they have a high $1/f$ noise component.

10.4.3 Poly SiGe as a bolometer material

As already mentioned, the use of self-sustained suspended bolometers provides a simpler technology, but is more demanding from the point of view of material properties. The common material used for such

[1] TCR_0 is the Temperature Coefficient of Resistance at 20 °C.

technology is poly Si [52]. The performances of realized devices are not comparable with those obtained using vanadium oxide, mainly because of the high *thermal conductance* of poly Si [53]. Furthermore, the control of the stress in poly-Si requires high-temperature annealing [54], which does not allow post-processing of the bolometer.

In this work, we introduce *poly SiGe* as a new material for bolometers. The most attractive point in using poly SiGe is that its thermal conductivity is at least a factor of five lower than that of poly Si [53]. To clarify this issue, we consider the heat transport mechanism in lowly doped semiconductors, which is due mainly to phonons, as the electronic contribution is negligible. The observation of a finite thermal conductivity is due to the existence of different of phonon scattering processes, the most important of which are:

1. Phonon-phonon scattering (two phonons interact to give a third phonon, the total momentum is either conserved or changed by a reciprocal lattice vector) [55];
2. Phonon scattering due to interaction with electron (or holes) [55];
3. Point-defect scattering [56]; and
4. Grain boundary scattering [57].

Mechanisms (1) and (2) are always present and their discussion is not relevant for our purposes. Mechanism (3) is very important. This kind of scattering is due to the presence of foreign atoms in substitutional positions in the lattice. They are capable of hindering the propagation of elastic waves in the solid. As an example, this mechanism is responsible for the decrease of thermal conductivity when dopants are added (notice that at high doping levels, the thermal conductivity increases again due to the increase of the electronic contribution). Moreover, this mechanism reduces the thermal conductivity of alloys, which is the case for silicon germanium (germanium atoms can be considered as point defects in the silicon lattice). The minimum thermal conductance of poly SiGe is obtained at a germanium content of 30% [58]. Thus, we shall use this germanium concentration in preparing poly SiGe bolometers. Mechanism (4) is also very important as it is thought to be responsible for the lower thermal conductivity of poly Si with respect to c-Si [53]. Preliminary data on poly SiGe shows that its thermal conductivity is a factor of two lower than that of c-SiGe (see Section 10.6.2). Low thermal conductivity, when combined with thin supports, will greatly improve the performance of self-suspended bolometers.

Poly SiGe has been prepared by *chemical vapor deposition* (CVD) from a mixture of germane and diclorosilane at atmospheric or reduced (40 torr) pressure. These two types of material will be indicated, respectively, as APCVD and RPCVD. The chemical vapor deposition, especially

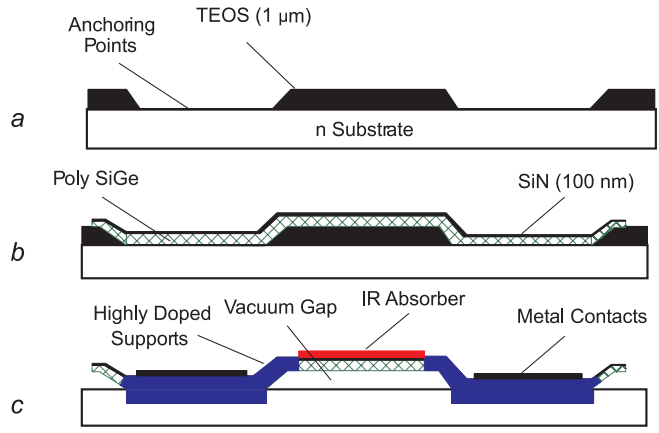


Figure 10.7: *Process flow for the fabrication of poly SiGe bolometers (**a**, **b**, **c** are explained in the accompanying text; see Section 10.5.1).*

at low pressure (LPCVD), is a technique largely used in microelectronics; poly SiGe is a material compatible with standard IC processes. Moreover, it has been shown that the stress can be tuned to the required value at relatively low temperatures (650 °C) [59], meanwhile similar stress can be realized in poly Si at temperatures higher than 850 °C [59]. Unlike vanadium oxide, the electrical properties of poly SiGe are less sensitive to the deposition conditions, which means a simpler and more easily transferable technology.

10.5 Poly SiGe bolometers

In this section, we will describe the process for realizing poly SiGe bolometers. The different types of IR absorbers will be discussed. It will be shown that selecting the best type of absorbers is a compromise between high emissivity and low thermal capacity. Finally, we will describe in detail the absorber used for poly SiGe bolometers.

10.5.1 Process for fabrication of poly SiGe bolometers

In this section we shall demonstrate the process required for realizing poly SiGe bolometers using surface micromachining techniques. In brief, it consists of depositing the active layer onto a sacrificial layer, which is etched away at the end of the process. In this way, the active layer will be suspended and connected to the substrate only through

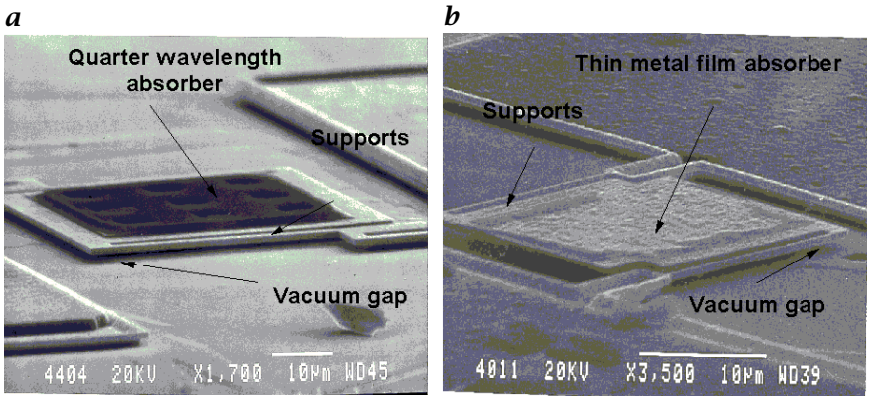


Figure 10.8: *An SEM picture of poly SiGe bolometers:* ***a*** *50 μm × 50 μm pixel, poly SiGe layer is 1 μm thick and the supports are 1 μm wide.* ***b*** *25 μm × 25 μm pixel, poly SiGe layer is 0.5 μm thick and the supports are 0.6 μm wide.*

thin supports. The sacrificial layer that we have chosen is TEOS, as it can be etched selectively with a high etch rate with respect to poly SiGe.

The different steps of the process are presented in Fig. 10.7. First a TEOS layer having a thickness of 1 μm is deposited on top of an N-type epi substrate. After TEOS deposition, the anchor points of the active element to the substrate are patterned (refer to Fig. 10.7a). This is followed by the deposition of poly SiGe (see Fig. 10.7b). The required TCR is obtained by ion implanting poly SiGe with the appropriate dose (in the range 1.5 $\times 10^{13}$ to 9 $\times 10^{13}$ boron/cm^2). For a given TCR, the thickness of poly SiGe is an important factor that decides the electrical resistance and thermal conductance of the structures. The thicker the layer, the lower the electrical resistance and the higher the thermal conductance. In our process the thickness of the active element varied from 0.25 μm to 1 μm.

As poly SiGe is transparent to IR radiation, an absorber layer must be deposited. We will see later that this layer is electrically conductive. In order to insulate the absorber electrically from the active element, a thin insulator layer is deposited on top of poly SiGe (see Fig. 10.7b). This layer is selected to be SiN having a thickness of 100 nm.

To transfer the electrical signal generated by the bolometer to the driving electronics, the support must be highly doped. This is achieved by a dose of 10^{16} boron/cm^2. This doping also forms a p-n junction with the substrate providing electrical insulation. Metal contacts are then deposited. Finally, the sacrificial layer is etched away using standard "freeze out" techniques [60, 61]. This step is shown schematically in Fig. 10.7c.

We have designed several structures having lateral dimensions varying from $50\,\mu\text{m} \times 50\,\mu\text{m}$ down to $25\,\mu\text{m} \times 25\,\mu\text{m}$. An SEM picture of some devices is displayed in Fig. 10.8. The thicknesses of poly SiGe layers for these devices are, respectively, $1\,\mu\text{m}$ and $0.5\,\mu\text{m}$. The support width is $1\,\mu\text{m}$ for $50\,\mu\text{m} \times 50\,\mu\text{m}$ and $0.6\,\mu\text{m}$ for the $25\,\mu\text{m} \times 25\,\mu\text{m}$. The holes on top of the structures are introduced to enhance etching of the sacrificial layer. The absorbers shown in the figure will be explained in detail in the next section.

10.5.2 IR absorbers

The function of the absorber is to convert the incident IR radiation into heat. It must have high absorption efficiency and high reproducibility and must be compatible with standard processes. Furthermore, its thermal mass must be low, compared to the thermal mass of the active element. Different materials and different structures can be designed to achieve 100% absorption. Examples are metal black coatings, very thin metal coatings, and quarter-wavelength structures.

Metal black coating has a significant absorption in the visible and near IR regions. It has been shown that 80% absorption, at $1\,\mu\text{m}$, can be achieved for a thick bismuth black deposited on top of glass [62]. A 92–98% absorption efficiency, in the range 8–14 μm, has been claimed for platinum black absorbers, few μm thick [63]. The main disadvantages of this type of absorber are its large thermal capacity, and the low compatibility of the method used for coating with the fabrication of the other elements of the array.

Thin metal film absorbers has been used in thermal detectors for a long time [64]. The sheet resistance of the metal film is adjusted to the value $R = Z_0/2$, where $Z_0 = \sqrt{\mu_0/\epsilon_0} = 377\,\Omega$ is the free space impedance. Such an adjustment of R gives maximum absorption (50%) at long wavelengths. It should be noted that such low values for the sheet resistance requires the deposition of a very thin layer (few tens of Angstrom), which is difficult to realize in a uniform and controllable way. This type of absorber is more feasible if metals having relatively high resistivity are used. An example of such metals is NiCr, where the desired sheet resistance is achieved for a layer 5-nm thick. An alternative is to use a quarter-wavelength structure, which improves the absorption of thin films from 50% to 100%, by backing the film with a perfect reflector at a $\lambda/4$ optical distance. When using this absorber, the different materials composing the absorber should have relatively low thermal mass so as not to slow down the performance of the detector.

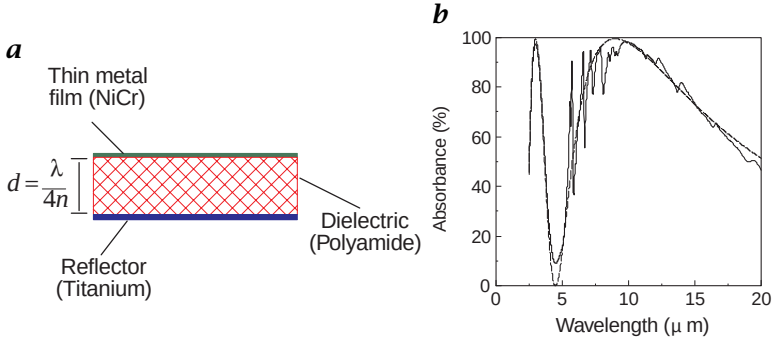


Figure 10.9: ***a*** *Schematic of the quarter-wavelength absorber;* ***b*** *Dependence of the IR absorber emissivity on wavelength. Full line: experimental data, dotted line: fit according to Eq. (10.21).*

10.5.3 Quarter-wavelength absorber

We shall discuss in detail the materials and characteristics of quarter-wavelength absorbers [65] as this is the one that we used in poly SiGe bolometers. The absorber is composed of a thermally evaporated titanium layer (0.2 μm thick), a polyamide layer of thickness d and a resistive nickel chromium (NiCr) absorber film having a sheet resistance R_f. A cross section of this absorber is shown schematically in Fig. 10.9a. The absorbance $\tilde{\epsilon}$ of such structure is given by [65]

$$\tilde{\epsilon}(d,\alpha,R_f) = \frac{4f}{[(f+1)^2 + n^2\cot^2(2\pi nd/\lambda)]} \tag{10.21}$$

where $f = 377/R_f$, n is the refractive index, λ is the wavelength of the incident radiation. It is clear from Eq. (10.20) that the absorbance is controlled mainly by the sheet resistance of the absorbing film and by the thickness of the dielectric layer. It is possible to achieve 100% absorbance for certain wavelength ($\lambda = 4nd$), if the absorbing film is matched to the free space ($R_f = 377\,\Omega/\text{sq}$).

For bolometer applications, we are interested in the wavelength region from 8 μm to 14 μm, which corresponds both to the maximum of the emission of a blackbody at 300 K and to a transmission window of the atmosphere [66]. As the refractive index of polyamide is 1.8, then its thickness should be around 1.4 μm to achieve maximum absorbance at 10 μm. The measured emissivity of the realized absorber is given by the solid line in Fig. 10.9. The behavior of the absorbance, calculated from Eq. (10.21) using $R_f = 320\,\Omega/\text{sq}$ and dielectric thickness of 1.25 μm, is reported in Fig. 10.9 (see the dotted line), which gives good agreement with experimental data. The thermal capacity of this absorber is 2.4 J/m^{2}K, which is reasonably good as compared

to that of other alternatives as platinum black has a thermal capacity varying from 1.6–9.9 J/m^{2}K. Moreover, the process of realizing the quarter-wavelength absorber is simpler and more reproducible as compared to metal black coatings.

10.6 Characterization of poly SiGe bolometers

In this section, we present a complete characterization of poly-SiGe-based bolometers. This includes the electrical, thermal, mechanical, and optical properties. In Section 10.6.1, the effect of boron doping on TCR and resistivity will be investigated over a doping range extending from 1.5×10^{13} boron/cm^2 to 9×10^{13} boron/cm^2. The procedure for measuring the thermal conductance of fabricated bolometers will be demonstrated in Section 10.6.2. The value of thermal conductivity of poly SiGe will be deduced and compared to that of poly Si. The dependence of stress in poly SiGe films on both the annealing temperature and deposition conditions will be presented in Section 10.6.3. It will be shown that the optimum value of stress can be obtained at 650 °C by adjusting the deposition conditions. These values will be compared to those typically found for poly Si. In Section 10.6.4, the procedure for measuring both the responsivity and noise of the device will be discussed. It will be shown that it is possible to achieve a responsivity of about 10^5 V/W. It will be demonstrated that the $1/f$ noise is the dominant noise component. The effect of the deposition conditions of poly SiGe on the $1/f$ noise will be investigated and possibilities for reducing this noise component will be discussed. It will be shown that an average detectivity of about 2×10^9 cm$\sqrt{\text{Hz}}/W$ can be achieved.

10.6.1 TCR of poly SiGe

In general, conduction in poly crystalline materials is a thermally activated process [26] and, hence, the resistivity can be expressed as

$$\rho(T) = \rho_0 e^{E_a/kT} \tag{10.22}$$

where E_a is the activation energy. The performance of the bolometer depends mainly on the TCR, which measures the change in the resistance corresponding to a temperature increase of one degree. The TCR is related to the activation energy by

$$TCR = -\frac{E_a}{KT^2} \tag{10.23}$$

It is clear from Eq. (10.23) that high sensitivity corresponds to large activation energies. The activation energy depends on the doping con-

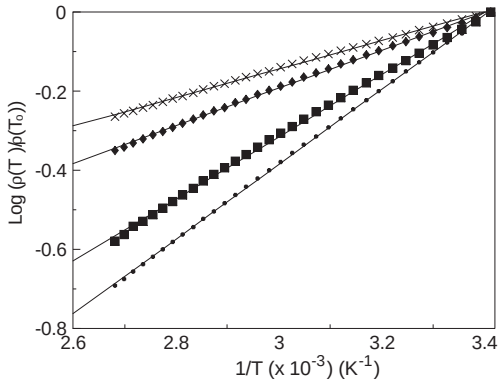


Figure 10.10: *Logarithm of the ratio between the resistance at temperature T, and the room temperature resistance vs the inverse of the absolute temperature for samples having different resistivity. Symbols: experimental data, solid line: fit according to Eq. (10.22)). (• $\rho(T_0) = 17.45\Omega\,cm$, $\alpha_0 = -2.45\,\%$; ■ $\rho/T_0) = 8.64\Omega\,cm$, $\alpha_0 = -2.11\,\%$; ♦ $\rho(T_0) = 1.97\Omega\,cm$, $\alpha_0 = -1.28 \times \rho(T_0) = 0.97\Omega\,cm$, $\alpha_0 = -0.96\,\%$).*

centration, the grain size, and the density of defects at the grain boundaries [26].

The activation energy of poly SiGe was experimentally determined by measuring the electrical current as a function of temperature, for constant voltage. Figure 10.10 displays the temperature dependence of the resistance of 1 μm thick poly SiGe layers ion implanted with 1.5 $\times 10^{13}$, 3×10^{13}, 6×10^{13}, and 9×10^{13} boron atoms / cm^2. The closed circles represent the measured data, while the straight lines represent a linear fit based on Eq. (10.22). From the slope of the line we can determine the activation energy, and hence, the TCR at any temperature can be computed from Eq. (10.23). The resistivity varies from 17.45 to 0.973 Ω cm, meanwhile, the TCR varies from $-2.54\,\%$/K to $-1\,\%$/K. The selection of the optimal doping dose, stems from a compromise between high TCR (low dose) and low noise (high dose), and it is the value that minimizes the NETD of the array. This will be clarified in Section 10.6.5.

It is interesting to compare the forementioned electrical properties with those of poly Si. Also, in poly Si the TCR and resistivity depend on the deposition conditions. For micromachining applications, as is the one of bolometers, deposition conditions are set by the requirement of low stress. The stress in LPCVD poly Si can be reduced by reducing the deposition temperature to 590 °C. In this case poly Si is deposited in the amorphous state and it crystallizes in the furnace during deposition. The TCR for this material (when doped at 3×10^{13} boron/cm^2) has been found to be $-2.87\,\%$ and the resistivity is 39.2 Ω cm. Comparing this

result to those obtained for poly SiGe, it can be deduced that poly SiGe has higher activation energy than poly Si, for the same resistivity. This feature allows reducing the resistivity of the material, thus maintaining a reasonable sensitivity.

10.6.2 Thermal conductivity of poly SiGe

Due to the role that thermal insulation plays in the performance of the device, it is important to measure exactly its value. To this aim two different procedures can be used. The first is based on applying a wide voltage pulse ($T \gg \tau$) and on measuring the power dissipated and the corresponding temperature rise of the bolometer, at steady state. The other approach is based on determining the time constant of the device from the dependence of the detected IR signal on frequency. This approach requires the knowledge of the thermal capacity of the structure and it will be discussed in Section 10.6.4.

In this section, we shall describe the first method. The experiment is performed in vacuum, and without IR radiation. The bolometer is biased at different V_b and the current I_b flowing through it is measured. A power $W = I_b V_b$ is dissipated in the bolometer. At steady state, the thermal conductance G and the power W are related by

$$W = G(T - T_0) \tag{10.24}$$

where T_0 is the room temperature and T is the temperature of the bolometer. The temperature T can be easily evaluated by manipulating Eq. (10.22) if the activation energy, the resistance at room temperature, $R(T_0)$, and the resistance at the temperature T, $R(T)$, are known. It is expressed as

$$T = \frac{T_0}{\left(1 + \frac{kT_0}{E_a} \ln\left(\frac{R(T)}{R(T_0)}\right)\right)} \tag{10.25}$$

The resistance of the bolometer $R(T)$ is simply computed as V_b/I_b. The value of $R(T_0)$ is computed in the same way but a narrow voltage pulse ($T \ll \tau$) is applied, so that the bolometer is not heated. In Fig. 10.11, the values of W vs $(T - T_0)$, obtained for different bias are plotted for $25\,\mu\text{m} \times 25\,\mu\text{m}$ pixels realized using $1\,\mu$m poly SiGe films, having support width of $0.6\,\mu$m and support length of $10\,\mu$m or $20\,\mu$m. From the slope of the straight line fitting these data we can compute the thermal conductance of the structures. It is clear from the figure that as the supports become longer the thermal insulation of the device is improved.

It should also be noted that the thermal conductance in this case is comparable to that achieved by insulating membranes (10^{-7} W/K [41]).

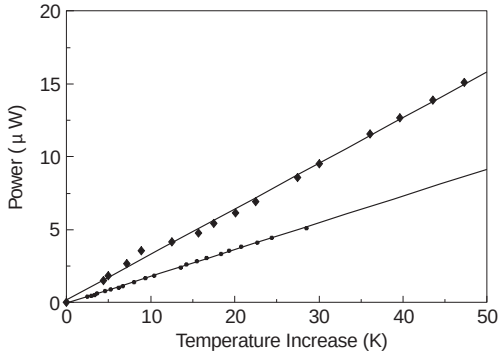


Figure 10.11: *Power dissipated in the bolometer vs the temperature rise that it generates in the bolometer itself. The two curves refer to bolometers having thermal conductance* $G = 1.81 \times 10^{-7}$ *W/K (•) and* $G = 3.2 \times 10^{-7}$ *W/K (♦).*

This means that using poly SiGe together with advanced lithographic techniques allows for achievement of high thermal insulation by means of a simpler technology. In order to relate the thermal conductance of the device to the thermal conductivity g of the material and to the geometry of the supports, we have performed a finite element analysis using the ANSYS simulator. Simulations were done by considering a uniform heat generation over the active area of the bolometer. Agreement between the simulated and experimental data was obtained for a value of $g = 2.7$ W/mK. Thermal simulations have shown that for wide (2 μm) and short (5 μm) supports, there is a significant temperature drop at the active area. When the supports become thin and long the thermal conductivity g of the material is, as expected, related to the thermal conductance of the structure by $G = gA/l$, where A and l are, respectively, the cross-sectional area and the length of the supports. It should be noted also that the value of thermal conductivity of poly SiGe, which we found, is lower than that of crystalline SiGe (5 W/mK [58]). As previously discussed, we believe that this effect is due to grain boundary scattering.

10.6.3 Mechanical properties of poly SiGe

The mechanical stability of the device is affected by the total stress of the suspended structure. This includes the stress in the active element and in the absorber layers. In general, the absorber could have either tensile or compressive stress, which should be compensated by the active element. This means that it should be possible to tune the stress of the active element to be either compressive or tensile. It is advisable to carry out this tuning at relatively low temperatures. To clarify this issue, the effect of annealing on stress induced in poly SiGe deposited

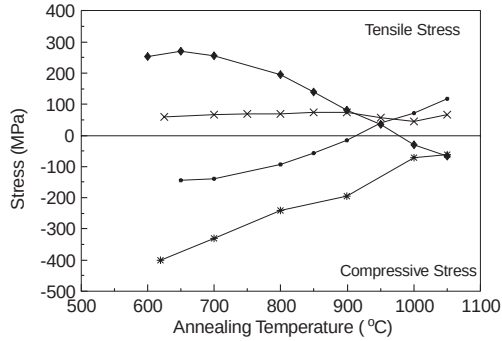


Figure 10.12: *Dependence of stress on annealing temperature: ♦ LPCVD poly Si grown at 590 ℃, ∗ LPCVD poly Si grown at 620 ℃, • APCVD poly SiGe grown at 650 ℃, and ■ RPCVD poly SiGe grown at 625 ℃.*

by APCVD and by RPCVD is displayed in Fig. 10.12, together with the stress of poly Si deposited at 590 ℃ (diamonds) and 620 ℃ (stars). It is evident that as-grown poly SiGe has in general lower stress than as-grown poly Si. Moreover, changing the deposition pressure of poly SiGe reverts the sign of stress from compressive to tensile. This illustrates that the stress can be easily tuned at low temperatures. It should be noted also that stress in RPCVD poly SiGe is insensitive to the annealing temperature, meanwhile, stress in APCVD poly SiGe is reduced and changes from compressive to tensile at about 900 ℃. On the other hand, the stress in as-grown poly Si is relatively high, and annealing is always necessary to reduce stress. Reducing the deposition temperature of poly Si reduces the stress induced in the as-grown material and at the same time results in tensile stress. The reason for this is that at 590 ℃, poly Si is deposited in the amorphous state and crystallizes in the furnace during deposition; tensile stress results from contraction against the grain boundaries.

It is interesting also to note that that the stress obtained from annealing poly SiGe at 650 ℃ is similar to that obtained for annealing poly Si at nearly 900 ℃ (compare the curve of RPCVD to that of poly Si deposited at 590 ℃). This means that using poly SiGe reduces the processing temperature by more than 200 ℃.

10.6.4 Responsivity of poly SiGe bolometers

The *responsivity* of the different devices is measured by mounting the bolometer inside a vacuum chamber and chopping the incident radiation of a blackbody. The effect of ambient light is eliminated by means of a germanium filter placed in front of the bolometer. The signal generated by the bolometer is detected by a lock-in amplifier. The responsiv-

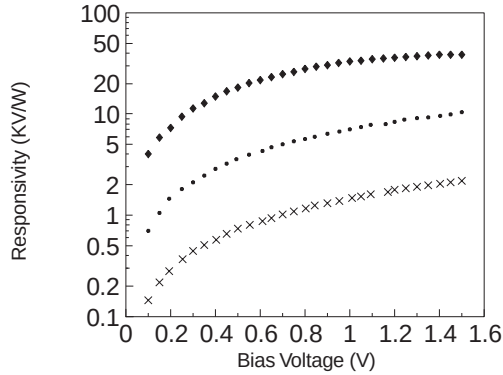


Figure 10.13: *Dependence of responsivity on bias voltage for devices having different thermal conductance:* ♦ $G = 1.56 \times 10^{-7}$ *w/K,* • $G = 8 \times 10^{-7}$ *W/K, and* ∗ $G = 4.13 \times 10^{-6}$ *W/K.*

ity of the device was measured at a chopper frequency of 9 Hz, which is smaller than the inverse of the time constant. We measured the responsivity of devices realized by using 1 μm thick, poly SiGe layers, and having the quarter-wavelength absorber described in Section 10.5.2. The dimensions of the device were varied from 25 μm × 25 μm to 50 μm × 50 μm, and the support width and length were varied, respectively, in the range 5–0.6 μm and 5–50 μm.

Figure 10.13 displays the measured responsivity as a function of the applied bias for devices with different thermal conductance. This figure clarify the role of the thermal insulation on the performance of the device. It is evident that the responsivity increases when the thermal conductance decreases (see Fig. 10.14, where it can be seen that the responsivity varies linearly with the thermal resistance).

As the devices are biased using a dc source, the responsivity does not increase linearly with voltage and will be limited. On the other hand, for pulsed bias the device temperature does not increase and the responsivity will vary linearly with the bias voltage. If data of Fig. 10.13 are linearly extrapolated to 5 V, a responsivity of more than 10^5 V/W is obtained.

The signal generated by the IR has been measured as a function of the chopper frequency. It is plotted in Fig. 10.15 for 25 μm × 25 μm pixels. For small temperature variations, the detected signal is proportional to the temperature increase of the bolometer, which can be expressed by Eq. (10.2). By fitting the measured data, using Eq. (10.2) (see the solid line in Fig. 10.15), we can determine the thermal time constant which is 10.35 ms in this case. Using the thermal capacity of SiGe (1.7 J cm^{-3}K^{-1}, for a 30% germanium content [57]) and that of the absorber (2.4 J m^{-2}K^{-1} [67]), we can compute the thermal conductance

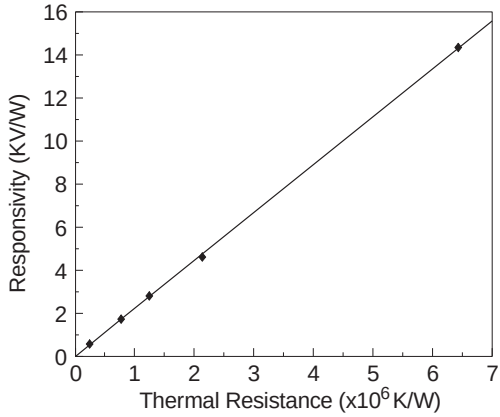


Figure 10.14: *Dependence of responsivity, measured at 0.1 V, on thermal resistance.*

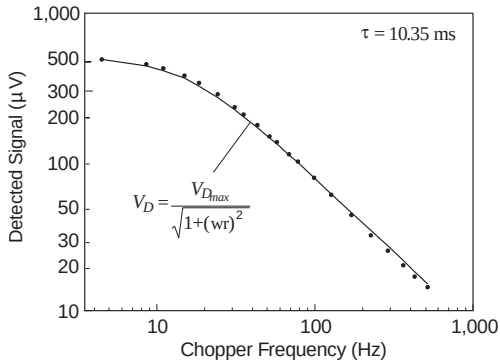


Figure 10.15: *Dependence of the detected signal on the chopper frequency.*

$(G = C/\tau = 1.56 \times 10^{-7}\text{W/K})$. This value is close to the value computed in Section 10.6.2 (see Fig. 10.11).

The thermal time constant of the device is an important parameter that decides whether or not the bolometer can fit to specific applications. If we consider integrating the bolometer into an infrared (IR) camera, having a frame rate of 20 Hz, this means that the time constant of the device should not exceed 40 ms. This is the case for the 25 μm $\times$ 25 μm devices. As the noise of small devices is large (see Eq. (10.12)) it might be necessary to increase the dimension of the device to reduce noise. If we consider, for instance, 50 μm $\times$ 50 μm devices having support widths of about 0.6 μm, the thermal time constant will be a factor of 8 higher than that of the 25 μm $\times$ 25 μm pixel. Reducing the time constant can be achieved by using as absorbers, a thin metal film,

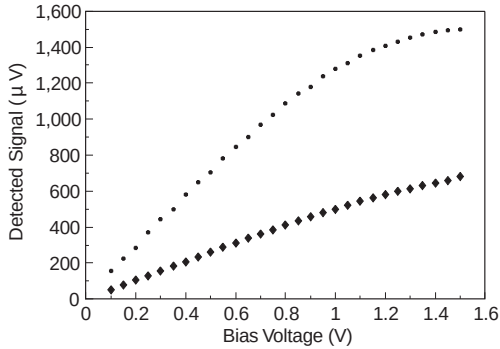


Figure 10.16: *Dependence of the detected signal on supply voltage for two bolometers differing only in the type of absorber. • quarter-wavelength absorber, ♦ thin metal film absorber. The time constant for the two detectors is, respectively, 10.35 ms and 3.5 ms.*

which has a negligible thermal mass as compared to that of the active element. In Fig. 10.16, we demonstrate the impact of using NiCr (a thin film metal absorber) on both the thermal time constant and the signal of a $25\,\mu\text{m} \times 25\,\mu\text{m}$ pixel. It is clear from the figure that the thermal time constant is nearly reduced by a factor of 2.5 and the detected signal is decreased by a factor of 3 as compared to the quarter-wavelength absorber. A better control of the NiCr thickness is expected to improve the level of the detected signal.

10.6.5 Noise in poly SiGe bolometers

To measure the noise, a load resistor having the same resistance as the bolometer and the bolometer itself are connected in series and biased by a 1.5 V battery. The voltage across the load resistor is sent to the input of a dynamic signal parameter analyzer, which is used to measure the power spectrum of the noise. The reliability of measurements is checked by verifying that the expected value of the Johnson noise is obtained if the bolometer is replaced by a resistor.

Figure 10.17a and b displays the dependence of the noise on frequency for bolometers prepared, respectively, as RPCVD and as APCVD. Different curves refer to different resistivities. It is possible to note that the noise scales with the resistivity according to Eq. (10.12). The comparison of Fig. 10.17a and b also shows that the noise is larger for materials deposited at reduced pressure, probably because of a difference in grain structure, which as previously demonstrated also influences the stress.

Although the $1/f$ noise is large, its effect on camera performance is not dramatic. In order to explain this point, we consider a $50\,\mu\text{m}$

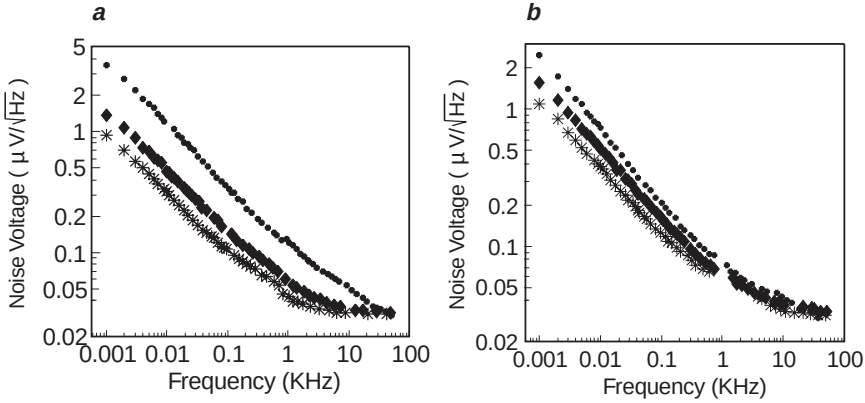


Figure 10.17: *Dependence of total noise on frequency for samples having different resistivity.* ***a*** *RPCVD poly SiGe (• ρ = 4.43Ωcm, ♦ ρ = 0.9Ωcm, ∗ ρ = 0.45Ωcm).* ***b*** *APCVD poly SiGe (• ρ = 8.64Ωcm, ♦ ρ = 1.9Ωcm, ∗ ρ = 0.97Ωcm).*

$\times$ 50 μm bolometer, having a thermal conductance of 8×10^{-7} W/K and realized by using APCVD poly SiGe with a resistivity of 8.64 Ωcm. Its responsivity as a function of bias is given by the middle curve of Fig. 10.13. Its noise spectrum is instead given by the upper curve of Fig. 10.17. If we assume a 10 to 10^5 bandwidth, which corresponds to an array of 240 $\times$ 320 pixels, we can compute the total contribution of the $1/f$ noise by integrating the noise power spectrum over this range. This gives, at 1.5 V of bias, 6 μV. The contribution of the Johnson noise is 15.2 μV, which is a factor of 2.5 larger. These values result in an NEP of 1.6×10^{-9}, which corresponds to an average detectivity of $D = \sqrt{A_d \Delta f}/NEP = 10^9\,\mathrm{cm}\sqrt{Hz}/W$. The value of the detectivity is doubled if noise and responsivity at 5 V of bias are used. The detectivity can be further improved by reducing the thermal conductance of the device, but due to mechanical stability reasons, the structure should be realized by RPCVD poly SiGe. The increase in $1/f$ noise, resulting from using RPCVD, as compared to APCVD, is compensated by reducing the thermal conductance. In this case detectivities larger than $4 \times 10^9\,\mathrm{cm}\sqrt{\mathrm{Hz}}/W$ can be achieved.

10.6.6 Noise in poly SiGe FPA

Based on the results presented in the previous subsection, we will discuss the structure of an optimized, poly SiGe based bolometer to be used in FPA. Optimization is performed by varying the geometry of the detector (which means changing the thermal conductance) and the doping level of the active area (which affects the TCR and the noise).

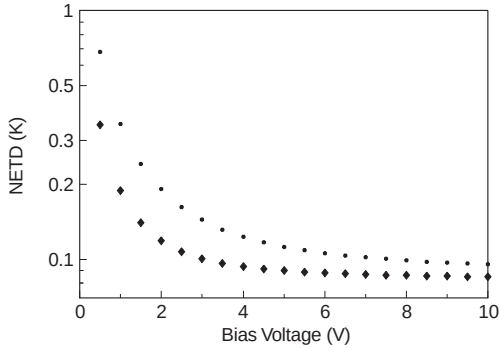


Figure 10.18: *Dependence of NETD on bias voltage for different array dimensions: ♦ 320× 240, • 640× 480.*

First, we note that any reduction of the thermal conductance must be accompanied by a reduction of the thermal mass, if the time constant is to be kept below 40 ms. This basically means that the bulky quarter-wavelength absorber must be replaced by the light, but less efficient, semitransparent NiCr layer. If this is done, and if a low stress material is used (namely poly SiGe deposited at reduced pressure), it is possible to realize a stable structure with thin poly SiGe layers. The gain in responsivity, obtained by the reduction of the thermal conductance, largely overcomes the loss due to the less efficient absorber.

As for the doping dose of the active area, a low dose corresponds to a large TCR, but also to a large noise; therefore, a compromise should be found. Furthermore, the noise depends also on the dimensions of the bolometer. Hence, the optimization of the doping dose and of the thermal conductance is coupled.

By following the preceding guidelines we found that the best performances can be obtained by using a 0.25 μm thick poly SiGe layer, prepared at reduced pressure with a resistivity of 2.5 Ωcm, coupled to a thin NiCr absorber. The NETD as a function of voltage, for two arrays composed respectively of 320 $\times$ 240 and 640 $\times$ 480 pixel, is reported in Fig. 10.18. Calculations are performed following the guidelines presented in section 3. It has been assumed that low noise external amplifiers are used, and that the pixel area is 50 μm x 50 μm, the fill factor is 80 % and the f/number of the IR optics is equal to one. It is possible to see that, for the smallest matrix, values of NETD as low as 85 mK can be obtained, comparable to state of the art values.

10.7 Conclusions

In this work, a brief survey of thermal imagers has been presented. It has been shown that microbolometers can provide a cheap detector technology for high performance uncooled FPAs. The different materials used for microbolometers have been mentioned, and the advantages of using poly SiGe have been highlighted. It has been shown that poly SiGe can be processed in a standard way without requiring ad hoc optimized processes. Furthermore, it is compatible with standard IC processes and, thus, it can be easily integrated with the driving electronics. This feature improves the fill factor of the pixels and reduces the cost of the FPA as it can be fabricated with standard foundry facilities and does not need a special setup.

Results achieved to date are very promising in terms of thermal insulation (close to 10^{-7} W/K), TCR (between (−1 % and −2 %) and IR emissivity (an average of 90 % over the wavelength range 8–12 μm). Responsivities in excess of 10^5 V/W have been demonstrated. The measured $1/f$ noise is somewhat large and is, at the moment, the most serious impediment in achieving outstanding results. Based on the physical properties of poly SiGe and on the performance of the already realized detectors, the NETD of a focal plane array comprising of 320×240 pixels, has been computed. A value of 85 mK has been found. This value is close to the best one reported for focal plane arrays based on vanadium oxide [66], and smaller than values obtained with metals (90 mK, 128×128 pixels [40]) and with amorphous semiconductors (100 mK, 256×64 pixels [68]).

We want to stress that these results are based on our actual knowledge of the properties of poly SiGe and there is still room for improvement. The most important point is to control the $1/f$ noise, without being obliged to use a high doping dose, which also implies a low TCR. We have seen that noise and strain depend on the deposition pressure, but, more generally, they depend on the deposition conditions. An optimization in this direction is now under development. To clarify this issue, we mention that an accurate choice of the deposition conditions of the material prepared at reduced pressure could bring its noise level to the one typical of atmospheric pressure material. In this way a decrease of the NETD by a factor of 2 will be readily obtained. We also note that $1/f$ noise in polycrystalline layers can be reduced by laser recrystallization, as observed in thin film transistors [69]. The use of this technique to reduce the $1/f$ noise in poly SiGe bolometers will also be explored in the near future. After optimizing the $1/f$ noise, the next step will be the realization of poly SiGe FPAs.

10.8 References

[1] Hudson, R. and Hudson, J., (1975). The military applications of remote sensing by infrared. *Proc. IEEE*, **63**:104–128.

[2] Ichikawa, M., (1989). Infrared spectra of penetration depth of into water and water refraction-index. *Proc. SPIE*, **1157**:318–328.

[3] Golay, M. J. E., (1947). A pneumatic infra-red detector. *Rev. Sci., Instr.*, **18**: 357–362.

[4] Scott Barr, E., (1962). The infrared pioneers-II. Macedonio Melloni. *Infrared physics*, **2**:67–73.

[5] Putley, E. H., (1964). The ultimate sensitivity of sub-mm detectors. *Infrared Physics*, **4**:1–8.

[6] Putley, E. H., (1977). Semiconductors and semi metals. In Willardson, R. K. and Beer, A. C. (eds.), *Infrared Detectors*, Vol. 2. New York: Academic Press.

[7] Abedini, Y. S., Barrett, O. R., Kim, J. S., Wen, D. D., and Yeung, S. S., (1996). 656×492-element platinum silicide infrared charge-coupled-device focal plane array. *Proc. SPIE*, **2020**:36–40.

[8] Wilson, T. E., Henricks, T. F., Halvis, J., Rosner, B. D., and Shiskowski, R. R., (1992). Versatile multimode $320 \times 240/256 \times 256$ hybrid InSb infrared focal plane array with selectable snapshot or rolling integration. *Proc. SPIE*, **1762**:401–406.

[9] Kanno, T., Saga, M., Kawahara, A., Oikawa, R., Ajisawa, A., Tomioka, Y., Oda, N., Yamagata, T., Murashima, S., Shima, T., and Yasuda, N., (1993). Development of MBE-grown HgCdTe 64×64 FPA for long-wavelength IR detection. *Proc. SPIE*, **2020**:41–48.

[10] Wenger, L. and Gaalema, S., (1992). Low power multiplexed lead salt arrays. *Proc. SPIE*, **1762**:407–417.

[11] Sedky, S., Fiorini, P., Caymax, M., Verbist, A., and Baert, C., (1998). IR bolometers made of polycrystalline silicon germanium. *Sensors and Actuators A*, **66 (1-3)**:193–199.

[12] Flanney, R. E. and Miller, J. E., (1992). Status of uncooled infrared imagers. *Proc. SPIE*, **1689**:379–395.

[13] Hanson, C., (1993). Uncooled thermal imaging at Texas Instruments. *Proc. SPIE*, **2020**:330–339.

[14] Horn, S. and Buser, R., (1993). Uncooled sensor technology. *Proc. SPIE*, **2020**:304–321.

[15] Owen, R., Belcher, J., Beratan, H., and Frank, S., (1994). Producability advances in hybrid uncooled infrared devices. *Proc. SPIE*, **2225**:79.

[16] Owen, R., Frank, S., and Daz, C., (1992). Producibility of uncooled IR FRA detectors. *Proc. SPIE*, **1683**:74.

[17] Watton, R., Denims, P. N. J., Gillhan, J. P., Manning, P. A., Perkins, M. C. J., and Todd, M. A., (1993). IR bolometer arrays, the route to uncooled, affordable thermal imaging. *Proc. SPIE*, **2020**:379–390.

[18] Wood, R. A., (1993). Uncooled thermal imaging with monolithic silicon focal plane. *Proc. SPIE*, **2020**:322–329.

[19] Liddiard, K. C., (1984). Thin-film resistance bolometer IR detectors. *Infrared Phys.*, **24**:57–64.

[20] Liddiard, K. C., (1986). Thin-film resistor bolometer IR detectors II. *Infrared Phys.*, **26**:43–49.

[21] Liddiard, K. C., (1993). Thin-film monolithic detector arrays for uncooled thermal imaging. *Proc. SPIE*, **1969**:206–216.

[22] Liddiard, K. C., Ringh, U., and Jansson, C., (1995). Staring focal plane arrays for advanced ambient temperature infrared sensor. *Proc. SPIE*, **2552**:564–572.

[23] Liddiard, K. C., Unewisse, M. H., and Reinhold, O., (1994). Design and fabrication of thin-film monolithic uncooled infrared detector arrays. *Proc. SPIE*, **2225**:62–71.

[24] Unewisse, M. H., Liddiard, K. C., and et al., B. I. C., (1995). Semiconductor film bolometer technology for uncooled IR sensor. *Proc. SPIE*, **2552**:77–87.

[25] Richards, P. L., (1994). Bolometers for infrared and millimeter waves. *J. Appl. Phys.*, **76(1)**:1–24.

[26] Kamins, T. L., (1988). *Polycrystalline Silicon for Integrated Circuit Applications*. Boston: Kluwer.

[27] Smith, R. A., Jones, F. E., and Chasmar, R. P., (1968). *The Detection and Measurement of Infra-Red Radiation,* 2nd edition. London: Oxford University Press.

[28] Laker, K. R. and Sansen, W. M. C., (1994). *Design of Analog Integrated Circuits and Systems*. New York: McGraw-Hill.

[29] Hooge, F. N., (1969). 1/f noise is no surface effect. *Physics Letter A*, **29**: 139.

[30] Gallo, M. A., Willits, D. S., Lubke, R. A., and Thiede, E. C., (1993). Low cost uncooled IR sensor for battlefield surveillance. *Proc. SPIE*, **2020**:351–362.

[31] Watton, R., Manning, P. A., Perkins, M., Gillham, J., and Todd, M., (1996). Uncooled IR imaging: Hybrid and integrated bolometer arrays. *Proc. SPIE*, **2744**:486–499.

[32] Herring, R. J. and Howard, P. E., (1996). Design and performance of the ULTRA 320×240 uncooled focal plane array and sensor. *Proc. SPIE*, **2746**: 2–12.

[33] Meyer, B., Cannata, R., Stout, A., Gim, A., Taylor, P., Woodbury, E., Deffner, J., and Ennerson, F., (1996). Amber's uncooled microbolometer LWIR camera. *Proc. SPIE*, **2746**:13–22.

[34] Marshall, C., Butler, N., Blackwell, R., Murphy, R., and Breen, I. T., (1996). Uncooled infrared sensor with digital focal plane array. *Proc. SPIE*, **2746**: 23–31.

[35] Marasco, P. L. and Dereniak, E. L., (1993). Uncooled infrared sensor performance. *Proc. SPIE*, **2020**:363–378.

[36] Lang, W., Steiner, P., Schaber, U., and Richter, A., (1994). A thin film bolometer using porous silicon technology. *Sensors and Actuators A*, **43**: 185–187.

[37] Shie, J. S., Chen, Y. M., and Chou, B. C. S., (1996). Characterization and modeling of metal film microbolometer. *Jour. Microelectromechanical Systems*, **5 (4)**:298–305.

[38] Shie, J. S. and Wenig, P. K., (1992). Design considerations of metal-film bolometer with micromachined floating membrane. *Sensors and Actuators A*, **33**:183–189.

[39] Tanaka, A., Matsumoto, S., Tsukamoto, N., Itoh, S., Endoh, T., Nakazato, A., Kumazawa, Y., Himikawa, M., Gotoh, H., Tomaka, T., and Teranishi, N., (1995). Silicon IC process compatible bolometer infrared focal plane array. In *The 8th International Conference on Solid State Sensors and Actuators and Eurosensors IX, Stockholm, Sweden, June 1995*, Vol. 2, pp. 632–635. IVA, Royal Swedish Academy of Engineering Sciences.

[40] Tanaka, A., Matsumoto, S., Tsukamoto, N., Itoh, S., Chiba, K., Endoh, T., Nakazato, A., Okayama, K., Kumazawa, Y., Hijikawa, M., Gotoh, H., Tanaka, T., and Teranishi, N., (1996). Infrared focal plane array incorporating silicon IC process compatible bolometer. *IEEE Trans. Electron Devices*, **43(11)**:1844–1850.

[41] Cole, B., Horning, R., Johnson, B., Nguyen, K., Kruse, P. W., and Foote, M. C., (1995). High performance infra red detector arrays using thin film microstructures. In *Proc. IEEE Int. Symp. on Applications of Ferroelectrics*, pp. 653–656.

[42] Umadevi, P., Negendra, C. L., and Thutupalli, G. K. M., (1993). Structural, electrical and infrared optical properties of vanadium pentoxide (V_2O_5) thick film thermisters. *Sensors and Actuators A*, **39**:59–69.

[43] Parker, T. W., Marshall, C. A., Kohin, M., and Murphy, R., (1997). Uncooled infrared sensors for surveillance and law enforcement applications. *Proc. SPIE*, **2935**:182–187.

[44] Butler, N., Blackwell, R., and et al., R. M., (1995). Low-cost uncooled microbolometer imaging system for dual use. *Proc. SPIE*, **2552**:583.

[45] Chudnovskii, F. A., (1975). Metal-semiconductor phase transition in vanadium oxides and its technical applications. *Sov. Phys. Tech. Phys.*, **20**:999.

[46] Jerominek, H., Picard, F., and Vicent, D., (1993). Vanadium oxide films for optimal switching and detection. *Opt. Eng.*, **32**:2092–2099.

[47] Kuznetsov, V. A. and Haneman, D., (1997). High temperature coefficient of resistance in vanadium oxide diodes. *Rev. Sci. Instrum.*, **68 (3)**:1518–1520.

[48] Umadevi, P., Negendra, C. L., and et al., G. K. M. T., (1991). A new thermistor material for thermistor bolometer: Material preparation and characterization. *Proc. SPIE*, **1485**:195.

[49] Zerov, V. Y., Kulikov, Y. V., Malyarov, V. G., Feokistov, N. A., and Kherbtov, I. A., (1997). Bolometric properties of silicon thin-film structures fabricated by plasmochemical vapor-phase deposition. *Tech. Phys. Lett.*, **23 (6)**:481–483.

[50] Enukova, T. A., Ivanova, N. L., Kulikov, Y. V., Malyarov, V. G., and Khrebtov, I. A., (1997). Amorphous silicon and germanium films for uncooled microbolometers. *Tech. Phys. Lett.*, **23(7)**:504–506.

[51] Ichihara, T., Watabe, Y., Honda, Y., and Aizawa, K., (1997). A high performance amorphous $Si_{1-x}C_x$:H thermister bolometer based on micromachined structure. In *1997 International Conference on Solid State Sensors and Actuators, Chicago*, pp. 1253–1256.

[52] NMRC, (1995). The development of integrated micro-bolometer arrays. In *Scientific Report'95*, p. 11. National Microelectronics Research Center.

[53] Paul, O., Korviet, J., and Boltes, H., (1994). Determination of the thermal conductivity of CMOS IC polysilicon. *Sensors and Actuators A*, **41-42**: 161–164.

[54] Maier-Schneider, D., Maibach, J., Obermeier, E., and Schneider, D., (1995). Variation in young's modulus and intrinsic stress of LPCVD-polysilicon due to high temperature annealing. *J. Micromech. Microeng.*, **5**:121–124.

[55] Vining, C. B., (1991). A model for the high temperature transport properties of heavily doped n-type silicon-germanium alloys. *J. Appl. Phys.*, **69**: 331–341.

[56] Steigmeier, E. F. and Abeles, B., (1964). Scattering of phonons by electrons in germanium-silicon alloys. *Phys. Rev.*, **136**:A1149.

[57] Slack, G. A. and Hussain, M. A., (1991). The maximum possible conversion efficiency of silicon germanium thermoelectric generators. *J. Appl. Phys.*, **70**:2694–2718.

[58] Dismukes, J., Ekstrom, L., Steigmeier, E., Kudam, I., and Beers, D., (1964). Thermal and electrical properties of heavily doped Ge-Si alloys up to 1300 °C. *J. Appl. Phys.*, **35**:2899.

[59] Fiorini, P., Sedky, S., Caymax, M., and Baert, K., (1997). Preparation and residual stress characterization of poly-silicon germanium films prepared by atmospheric pressure chemical vapor deposition. *Proc. Mat. Res. Soc. Symp.*, **472**:227–231.

[60] Core, T. A., Tsang, W. K., and Sherman, S. J., (1993). Fabrication technology for an integrated surface micromachined sensor. *Solid State Technology*, **36**:39–48.

[61] Tas, N., Sonnenberg, T., Jansen, H., Legtenberg, R., and Spoek, M. E., (1996). Stiction in surface micromachining. *J. Micromech. Microeng.*, **6**:385–397.

[62] Strimer, P., Gerbaux, X., Hadni, A., and Souel, T., (1981). Black coatings for infrared and visible, with high electrical resistivity. *Infra Red Physics*, **21**:37–39.

[63] Betts, D. B., Clarke, F. J. J., Cox, L. J., and Larkin, J. A., (1985). Infrared reflection properties of five types of black coating for radiometric detectors. *J. Physics*, **18**:689–696.

[64] Veremei, V. V. and Pankrotov, N. A., (1974). Interference phenomena in semiconductor bolometers. *Sov. J. Opt. Technol.*, **41**:199.

[65] Parsons, A. D. and Pedder, D. J., (1988). Thin-film infrared absorber structures for advanced thermal detectors. *J. Vac. Sci. Technol.*, **A6 (3)**:1686–1689.

[66] Stout, A. and Rittenberg, E., (1997). High performance hand-held thermal imager for law enforcement. *Proc. SPIE*, **2935**:154–157.

[67] Pankratov, N. A. and Malyarov, N. G., (1985). Detector layout of a submillimeter photometer. *Zh. Prikl. Specktrosk.*, **42**:1028.

[68] Tissot, J. L., (1998). What is an uncooled infrared microbolometer? (French). *CLEFS CEA*, **37**:28–33.

[69] Carluccio, R., Corradetti, A., Fortunatto, G., Reita, C., Legagneux, P., Plais, F., and Pribat, D., (1997). Noise performances in polycrystalline silicon thin-film transistors fabricated by excimer laser crystallization. *Appl. Phys. Lett.*, **71**:578–580.

11 Hyperspectral and Color Imaging

Bernd Jähne

Interdisziplinäres Zentrum für Wissenschaftliches Rechnen (IWR)
Universität Heidelberg, Germany

11.1 Spectral signatures

Single measured image irradiance is generally a poor indicator of object properties because it is a product of the object reflectivity or eventually other optical properties (e.g., see Chapters 3 and 5) and the irradiance by external illumination sources. Absolute measurements thus require careful calibration. Except for these principal difficulties one scalar feature is often not sufficient to identify a certain object and to measure its properties.

Spectroscopic imaging is, in principle, a very powerful tool to identify objects and their properties because almost all optic material constants such as

- *reflectivity*
- *index of refraction*
- *absorption coefficient*
- *scattering coefficient*

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

Table 11.1: *Examples of some strategies for spectral sampling*

Sampling Method	Description and Application
Line sampling	Channels with narrow spectral range (line); suitable for absorption, emission, and luminescence imaging for specific chemical species and/or specific processes; orthogonal base for color space.
Band sampling	Channels with wide spectral range (band) of uniform responsivity, adjacent to each other; suitable for measurements of spectral radiance with rather coarse resolution; orthogonal base for color space.
Parameter-based sampling	Sampling optimized for a certain model parameters of the spectral distribution. The parameters of the spectral distribution are estimated; generally nonorthogonal base for color space.

- *optical activity*
- *luminescence*

depend on the wavelength of the radiation (Chapter 3).

The trouble with spectroscopic imaging is that it adds another coordinate to imaging and the required amount of data is multiplied correspondingly. Therefore, it is important to sample the spectrum with a minimum number of samples that is sufficient to perform the required task. We introduce here several sampling strategies and discuss, from this point of view, human color vision as one realization of spectral sampling in Section 11.3.

11.2 Spectral sampling methods

Table 11.1 illustrates three different types of sampling that will be discussed in the following sections.

11.2.1 Line sampling

With this technique, each channel picks only a narrow spectral range. This technique is useful if processes are to be imaged that are related to the emission or the absorption at specific spectral lines. The technique is very selective. One channel "sees" only a specific wavelength and is insensitive—at least to the degree that such a narrow bandpass filtering can be realized technically—to all other wavelengths. Thus, a

very specific effect or a specific chemical species can be imaged with this technique. This technique is, of course, not appropriate to make an estimate of the total radiance from objects because it misses most wavelengths.

11.2.2 Band sampling

This is the appropriate technique if the total radiance in a certain wavelength range has to be imaged and some wavelength resolution is still required. Ideally, the individual bands have even responsivity and are adjacent to each other. Thus, band sampling gives the optimum resolution with a few channels but does not allow any distinction of the wavelengths *within* one band. Thus, we can measure the spectral radiance with a resolution given by the width of the spectral bands.

11.2.3 Parameter-based spectral sampling

In almost all applications, the spectral radiance is not of interest by itself but the object features that characterize the radiation emitted by the object. Often there are only a few parameters. Extracting these few, say P, parameters from a complete scan of the spectrum with Q samples ($Q \gg P$) is certainly a waste of memory. Remember that at each point of the image Q samples must be taken just to extract P parameters. It is obvious that—at least in principle—only P samples are required to determine P parameters.

Before we treat this problem formally, it will be illustrated with two simple examples.

Example 11.1: Measurement of total radiative flux and mean wavelength

This example demonstrates that it is possible to determine the total radiative flux Φ ("intensity")

$$\Phi = \int_{\lambda_1}^{\lambda_2} \Phi(\lambda)\, \mathrm{d}\lambda \tag{11.1}$$

and the mean wavelength $\langle\lambda\rangle$ ("color")

$$\langle\lambda\rangle = \int_{\lambda_1}^{\lambda_2} \lambda\Phi(\lambda)\, \mathrm{d}\lambda \Big/ \int_{\lambda_1}^{\lambda_2} \Phi(\lambda)\, \mathrm{d}\lambda \tag{11.2}$$

with just two sensors having an adequate spectral sensitivity.

In order to measure the total radiative flux according to Eq. (11.1), it is required to use a sensor with a wavelength-independent responsivity

$$R_1(\lambda) = R_0 \tag{11.3}$$

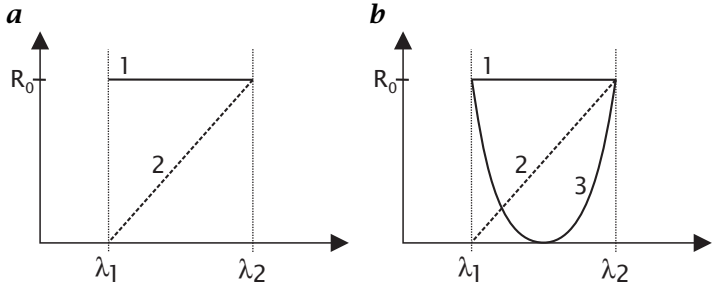


Figure 11.1: *Examples of model-based spectral sampling* ***a*** *to determine the total radiance and the mean wavelength; and* ***b*** *same as* ***a*** *plus the variance of the spectral width.*

R is the *responsivity* of the sensor given as $R(\lambda) = s(\lambda)/\Phi(\lambda)$ (units A/W). The sensor signal s is usually given in units for the electric current. From the multiplication of the spectral flux by the wavelength in Eq. (11.2), it is evident that we need a second sensor that has a sensitivity that varies linearly with the wavenumber (Fig. 11.1a)

$$R_2(\lambda) = \frac{\lambda - \lambda_1}{\lambda_2 - \lambda_1} R_0 = \left(\frac{1}{2} + \tilde{\lambda}\right) R_0 \tag{11.4}$$

where $\tilde{\lambda}$ the normalized wavelength

$$\tilde{\lambda} = \left(\lambda - \frac{\lambda_1 + \lambda_2}{2}\right) / (\lambda_2 - \lambda_1) \tag{11.5}$$

$\tilde{\lambda}$ is zero in the middle of the interval and $\pm 1/2$ at the edges of the interval. Note that the offset $1/2$ is required in Eq. (11.4) since only positive signals can be measured. The signal given by a sensor is

$$s = \int_{\lambda_1}^{\lambda_2} R(\lambda)\Phi(\lambda)\,\mathrm{d}\lambda \tag{11.6}$$

Using Eqs. (11.4) and (11.5), we can infer that the mean wavelength as defined by Eq. (11.2) is directly related to the ratio of the two sensor signals:

$$\frac{s_2}{s_1} = \left\langle \tilde{\lambda} \right\rangle + 1/2 \quad \text{or} \quad \langle \lambda \rangle = \lambda_1 + \frac{s_2}{s_1}(\lambda_2 - \lambda_1) \tag{11.7}$$

while the total radiant flux is given by

$$\Phi = \frac{s_1}{s_2} / R_0 \tag{11.8}$$

It is interesting to observe that only the determination of the total radiant flux requires an *absolute calibration* of the sensor. It is not needed for the determination of the mean wavelength because it is given as the ratio of two sensor signals (*ratio imaging*).

It is important to note that both line and band sampling are not suitable to determine the mean wavelength and total radiant flux. Sampling just at two lines misses all wavelengths except the two selected lines. Thus the total radiant flux is incorrect.

With band sampling it is possible to get the total radiant flux right provided that the two selected bands are adjacent to each other and cover the whole wavelength range of interest. The mean wavelength, however, comes out incorrectly in the general case. This is directly related to the fact that within the selected band all wavelengths are equally weighted. The wavelength of a monochromatic radiative flux, for example, cannot be determined with better resolution than the bandwidth of the individual channels. What is needed according to Eq. (11.4) is linearly changing responsivity over the wavelength interval for the second sensor.

The example also illustrates that measurements of this type are always a many-to-one mapping. The two sensors receive the same signal for all types of spectral distributions that have the same total radiant flux and mean wavelength as defined by Eqs. (11.1) and (11.2).

Example 11.2: Measurement of total radiative flux, mean, and variance of the wavelength

The two-channel system discussed in Example 11.1 cannot measure the width of a spectral distribution at all. This deficit can be overcome with a third channel that has a sensitivity that increases with the square of the distance from the mean wavelength (Fig. 11.1b).

The responsivity of the third sensor is given by

$$R_3(\lambda) = 4\tilde{\lambda}^2 R_0 \tag{11.9}$$

Consequently, the mean squared wavelength is given by

$$\left\langle \tilde{\lambda}^2 \right\rangle = \frac{1}{4}\frac{s_3}{s_1} \tag{11.10}$$

The variance $\sigma_{\tilde{\lambda}}^2 = \left\langle \left(\tilde{\lambda} - \left\langle \tilde{\lambda} \right\rangle\right)^2 \right\rangle = \left\langle \tilde{\lambda}^2 \right\rangle - \left\langle \tilde{\lambda} \right\rangle^2$ is then given by

$$\sigma_{\tilde{\lambda}}^2 = \frac{1}{4}\frac{s_3}{s_1} - \left(\frac{s_2}{s_1} - \frac{1}{2}\right)^2 \tag{11.11}$$

For a monochromatic distribution at the wavelength λ_0 the variance is zero. Then

$$\sigma_{\tilde{\lambda}}^2 = \left\langle \lambda_0^2 \right\rangle - \langle \lambda_0 \rangle^2 = \lambda_0^2 - \lambda_0^2 = 0 \tag{11.12}$$

The estimates given by Eqs. (11.10) and (11.11) are only valid as long as the spectral distribution is confined to the interval $[\lambda_1, \lambda_2]$ to which the sensors respond.

After these two introductory examples, we formulate linear parameter-based sampling in a general way as a *linear discrete inverse problem* [1]. As in Examples 11.1 and 11.2 we assume that P parameters $\boldsymbol{p}$ of interest are a linear combination of the spectral flux density and we want to measure them from other linear combinations of the spectral flux density by the use of Q sensor signals $\boldsymbol{q}$ with various spectral sensitivities. The general question is whether this is possible at all in general and if yes under which conditions.

In order to derive the relation between the parameters $\boldsymbol{p}$ and the sensor signals $\boldsymbol{q}$, we assume a hypothetical band-sampled spectral density $\boldsymbol{s}$ with S samples. The sampling must be dense enough so that the sampling theorem (see Volume 2, Section 2.4.2) is met. Then the linear relations between the band-sampled spectral density $\boldsymbol{s}$ and the parameter vector $\boldsymbol{p}$ and the signal vector $\boldsymbol{q}$ can be written as:

$$\boldsymbol{p} = \boldsymbol{P}\boldsymbol{s} \quad \text{and} \quad \boldsymbol{q} = \boldsymbol{Q}\boldsymbol{s} \tag{11.13}$$

where $\boldsymbol{P}$ and $\boldsymbol{Q}$ are P × S and Q × S matrices, respectively, with $P \leq Q \leq S$. Given the linear nature of the ansatz, the direct relation between $\boldsymbol{p}$ and $\boldsymbol{q}$ must also be linear provided that a solution exists at all:

$$\boldsymbol{p} = \boldsymbol{M}\boldsymbol{q} \tag{11.14}$$

Replacing $\boldsymbol{q}$ in Eq. (11.14) by $\boldsymbol{q} = \boldsymbol{Q}\boldsymbol{s}$ and using $\boldsymbol{p} = \boldsymbol{P}\boldsymbol{s}$, a direct relation between the three matrices $\boldsymbol{M}, \boldsymbol{P}$, and $\boldsymbol{Q}$ is obtained:

$$\boldsymbol{P} = \boldsymbol{M}\boldsymbol{Q} \tag{11.15}$$

Using standard least-squares techniques (see, e. g., Menke [1]), the P × Q matrix $\boldsymbol{M}$ is given by

$$\boldsymbol{M} = \boldsymbol{P}\boldsymbol{P}^T\boldsymbol{P}\boldsymbol{Q}^T(\boldsymbol{Q}\boldsymbol{P}^T\boldsymbol{P}\boldsymbol{Q}^T)^{-1} \tag{11.16}$$

provided that the inverse of the Q × Q matrix $\boldsymbol{Q}\boldsymbol{P}^T\boldsymbol{P}\boldsymbol{Q}^T$ exists. Thus this equation tells us both the condition for the existence of a solution and how to compute it by matrix-matrix multiplications. The solubility does not depend at all on the actual measurements $\boldsymbol{q}$ but only on the spectral responsitivity of the Q sensors $\boldsymbol{Q}$ and the spectral shape of the parameters $\boldsymbol{p}$ to be estimated that are contained in the matrix $\boldsymbol{P}$.

With this general formulation we have a powerful general concept. We can use it to handle any multispectral or multichannel image processing task where we measure Q channels and want to retrieve P parameters that are linear combinations of the measured signals. The two simple Examples 11.1 and 11.2 discussed at the beginning of this section, human color vision (see Section 11.3), and *differential optical absorption spectroscopy* (DOAS) discussed in Volume 3, Chapter 37 are just four examples of this general type of problems.

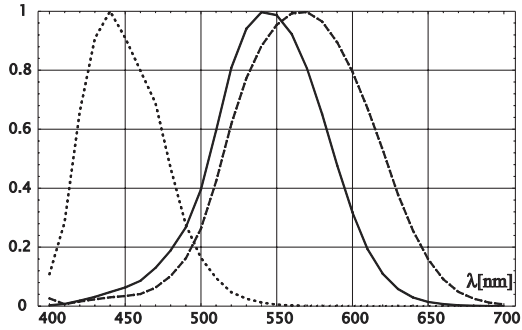


Figure 11.2: *Estimates of the relative cone sensitivities of the human eye after DeMarco et al. [2].*

11.3 Human color vision

Human color vision can be regarded in terms of the spectral sampling techniques summarized in Table 11.1 as a parameter-based sampling. It does not measure the spectral radiant flux directly but rather properties of the spectral distribution such as the total radiant flux (*intensity*), the mean wavelength (*color*), and the width of the spectral distribution (*saturation* of the color). If the width of the spectral distribution is narrow we have a pure color with high saturation. If the spectral distribution is wide, the color has a low saturation. If the spectral distribution is flat, we sense no color. With the respect to this discussion, it appears that the three-sensor system discussed in Example 11.2 appears to be an ideal intensity-color-saturation sensor. It is ideal in the sense that it has a linear response and the wavelength (color) and width (saturation) resolution are independent of the wavelength. Thus it is interesting to compare this three-sensor system with the color-sensing system of the human eye.

For color sensing, the human eye has also three types of photopigments in the photoreceptors known as cones with different spectral sensitivities (Fig. 11.2). The sensitivities cover different bands with maximal sensitivities at 445 nm, 535 nm, and 575 nm, respectively (band sampling), but overlap each other significantly (parameter-based sampling). In contrast to our model examples, the three sensor channels are unequally spaced and cannot simply be linearly related. Indeed, the color sensitivity of the human eye is uneven and all the nonlinearities involved make the science of color vision rather difficult. Here, only some basic facts are given—in as much as they are useful to handle color imagery.

11.3.1 Three-dimensional color space

Having three color sensors, it is obvious that color signals cover a 3-D space. Each point in this space represents one color. From the discussion on spectral sampling in Section 11.2, it is clear that many spectral distributions called *metameric color stimuli* or short *metameres* map onto one point in this space. Generally, we can write the signal s_i received by a sensor with a spectral responsivity $R_i(\lambda)$ as

$$s_i = \int R_i(\lambda)\Phi(\lambda)d\lambda \tag{11.17}$$

With three primary color sensors, a triple of values is received, often called *tristimulus* and represented by the 3-D vector $\boldsymbol{s} = [s_1, s_2, s_3]^T$.

11.3.2 Primary colors

One of the most important questions in *colorimetry* is a system of how to represent colors as linear combinations of some basic or *primary colors*. A set of three spectral distributions $\Phi_j(\lambda)$ represents a set of primary colors and results in an array of responses that can be described by the matrix P with

$$P_{ij} = \int R_i(\lambda)\Phi_j(\lambda)d\lambda \tag{11.18}$$

Each vector $\boldsymbol{p}_j = \left[p_{1j}, p_{2j}, p_{3j}\right]^T$ represents the tristimulus of the primary colors in the 3-D color space. Then, it is obvious that any color can be represented by the primary colors that are a linear combination of the base vectors $\boldsymbol{p}_j$ in the following form:

$$\boldsymbol{s} = R\boldsymbol{p}_1 + G\boldsymbol{p}_2 + B\boldsymbol{p}_3 \qquad \text{with} \qquad 0 \le R, G, B \le 1 \tag{11.19}$$

where the coefficients are denoted by R, G, and B, indicating the three primary colors red, green, and blue. Note that these coefficients must be positive and smaller than one. Because of this condition, all colors can be presented as a linear combination of a set of primary colors only if the three base vectors are orthogonal to each other. This cannot be the case as soon as more than one of the color sensors responds to one primary color. Given the significant overlap in the spectral response of the three types of cones (Fig. 11.2), it is obvious that none of the color systems based on any type of real primary colors will be orthogonal. The colors that can be represented lie within the parallelepiped formed by the three base vectors of the primary colors. The more the primary colors are correlated with each other (i. e., the smaller the angle between two of them is), the smaller is the color space that can be represented

Table 11.2: *Most often used primary color systems. The second column gives also the conversion matrix of the corresponding color system to the XYZ color system (values taken from Wendland [3, Section 5.7.4] and Pratt [5, Table 3.5-1]).*

Name	Description
Monochromatic primaries R_c, G_c, B_c	Adapted by C.I.E. in 1931 λ_R = 700 nm, λ_G = 546.1 nm, λ_B = 435.8 nm $\begin{bmatrix} 0.490 & 0.310 & 0.200 \\ 0.177 & 0.812 & 0.011 \\ 0.000 & 0.010 & 0.990 \end{bmatrix}$
NTSC Primary Receiver Standard R_N, G_N, B_N	FCC Standard, 1954, to match phosphors of RGB color monitors $\begin{bmatrix} 0.6070 & 0.1734 & 0.2006 \\ 0.2990 & 0.5864 & 0.1146 \\ 0.0000 & 0.0661 & 1.1175 \end{bmatrix}$
S.M.P.T.E. Primary Receiver Standard R_S, G_S, B_S	Better adapted to modern screen phosphors $\begin{bmatrix} 0.393 & 0.365 & 0.192 \\ 0.212 & 0.701 & 0.087 \\ 0.019 & 0.112 & 0.985 \end{bmatrix}$
EBU Primary Receiver Standard R_e, G_e, B_e	Adopted by EBU 1974 $\begin{bmatrix} 0.4303 & 0.3416 & 0.1780 \\ 0.2219 & 0.7068 & 0.0713 \\ 0.0202 & 0.1296 & 0.9387 \end{bmatrix}$

by them. Mathematically, colors that cannot be represented by a set of primary colors have at least one negative coefficient in Eq. (11.19). The most often used primary color systems are summarized in Table 11.2.

11.3.3 Chromaticity

One component in the 3-D color space is intensity. If a color vector is multiplied by a scalar, only its intensity is changed but not its color. Thus, all colors could be normalized by the intensity. This operation reduces the 3-D color space to a 2-D color plane or *chromaticity diagram*:

$$r = \frac{R}{R+G+B}, \quad g = \frac{G}{R+G+B}, \quad b = \frac{B}{R+G+B} \tag{11.20}$$

with

$$r + g + b = 1 \tag{11.21}$$

It is sufficient to use only the two components r and g. The third component is then given by $b = 1 - r - g$, according to Eq. (11.21).

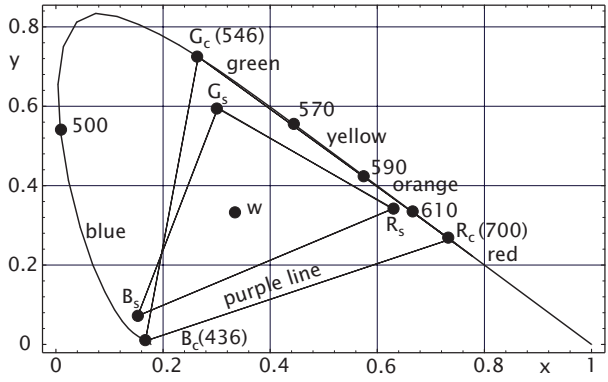


Figure 11.3: *Chromaticity diagram shown in the xy color space. The u-shaped curve of monochromatic colors with wavelengths in nm as indicated and the purple line includes all possible colors. Shown are also range of colors (triangles) that can be represented with monochromatic primaries R_c, G_c, B_c and the SMPTE primary receiver standard R_S, G_S, B_S.*

Thus, all colors that can be represented by the three primary colors R, G, and B are confined within a triangle. As already mentioned, some colors cannot be represented by the primary colors. The boundary of all possible colors is given by all visible monochromatic colors from deep red to blue. The line of monochromatic colors form a u-shaped curve (Fig. 11.3). Thus, most monochromatic colors cannot be represented by the monochromatic primaries. As all colors that lie on a straight line between two colors can be generated as a mixture of these colors, the space of all possible colors covers the area filled by the u-shaped spectral curve and the straight mixing line between its two end points for blue and red color (*purple line*).

In order to avoid negative color coordinate values, often a new coordinate system is chosen with virtual primary colors, that is, primary colors that cannot be realized by any physical colors. This color system is known as the *XYZ color system* and constructed in such a way that it includes just the curve of monochromatic colors with only positive coefficients (Fig. 11.3).

11.3.4 Hue and saturation

The color systems discussed so far do not directly relate to the human color sensing. From the rg or xy values, we cannot directly infer colors such as green, blue, etc. In addition to *luminance* (*intensity*), a description of colors would also include the type of color such as green or blue (*hue*) and the purity of the color (*saturation*). From a pure color, we can obtain any degree of saturation by mixing it with white.

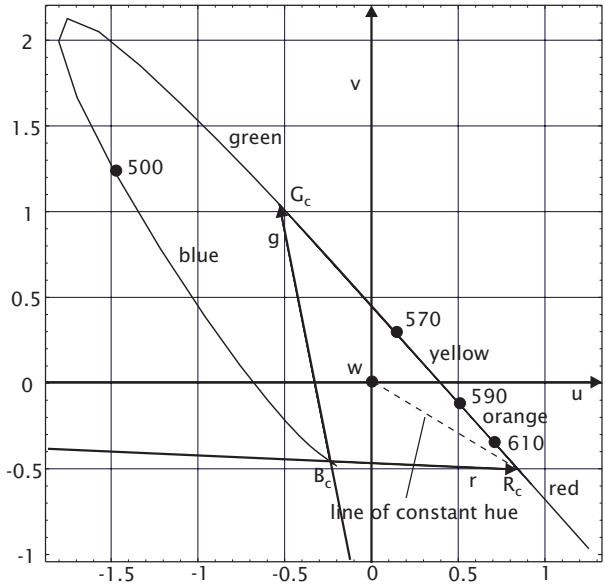


Figure 11.4: *Chromaticity diagram shown in the uv color difference system centered at the white point w. The color saturation is proportional to the distance from the center and the color hue is given by the angle to the x axis. Shown are also the axes of the rg color system marked with r and b.*

Hue and saturation can be extracted from chromaticity diagrams by simple coordinate transformations. The essential point is the *white point* in the middle of the chromaticity diagram (Fig. 11.4). If we draw a line from this point to a pure (monochromatic) color, it constitutes a mixing line for a pure color with white and is thus a line of constant hue. From the white point to the pure color, the saturation increases linearly. The *white point* is given in the rg chromaticity diagram by $w = (1/3, 1/3)$. A color system that has its center at the white point is called a *color difference system*. From a color difference system, we can infer a hue-saturation color system by simply using polar coordinate systems. Then, the radius coordinate is proportional to the saturation and the hue to the angle coordinate (Fig. 11.4).

Color science is, in the abstract, relatively simple. However, real difficulties arise from what is required to adapt the color system in an optimum way to display and print devices, for transmission by television signals, or to correct for the uneven color resolution of the human visual system that is apparent in the chromaticity diagrams of simple color spaces (Figs. 11.3 and 11.4). The result to date is a confusing manifold of different color systems. For a detailed treatment of color vision, the reader is referred to the monography written by the Commit-

tee on Colorimetry of the Optical Society of America [4]. An excellent treatment of color with respect to digital image processing is given by Pratt [5] and with respect to video engineering by Inglis [6].

11.3.5 Intensity-hue-saturation color coordinate system

Here, we discuss only one further color coordinate system that is optimally suited to present vectorial image information as colors on monitors. With a gray scale image, only one parameter can be represented. In color, it is, however, possible to represent three parameters simultaneously, for instance as intensity, hue, and saturation (IHS). This representation is known as the IHS *color coordinate system*. The transformation is given by

$$\begin{aligned} \begin{bmatrix} I \\ U \\ V \end{bmatrix} &= \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \\ H &= \arctan\left(\frac{V}{U}\right) \\ S &= (U^2 + V^2)^{1/2} \end{aligned} \tag{11.22}$$

This transformation essentially means that the zero point in the chromaticity diagram has been shifted to the white point. The pairs $[U,V]^T$ and $[S,H]^T$ are the Cartesian and polar coordinates in this new coordinate system, respectively.

11.4 References

[1] Menke, W., (1984). *Geophysical data analysis: discrete inverse theory*. Orlando: Academic Press.

[2] DeMarco, P., Pokorny, J., and Smith, V. C., (1992). Full-spectrum cone sensitivity functions for X-chromosome-linked anomalous trichromats. *J. Optical Society*, **A9**:1465–1476.

[3] Wendland, B., (1988). *Fernsehtechnik I: Grundlagen*. Heidelberg: Hüthig.

[4] Committee on Colorimetry, Optical Society of America, (1953). *The Science of Color*. Washington, D. C.: Optical Society of America.

[5] Pratt, W., (1991). *Digital image processing*. New York: Wiley.

[6] Inglis, A. F., (1993). *Video engineering*. New York: McGraw-Hill.

Part III

Two-Dimensional Imaging

12 Dynamic Fluorescence Imaging

Dietmar Uttenweiler and Rainer H. A. Fink

II. Physiologisches Institut, Universität Heidelberg, Germany

12.1 Introduction

The use of dynamic fluorescence imaging techniques has grown in many fields of scientific applications. Dynamic fluorescence imaging comprises the acquisition, the digital image processing and the mathematical analysis of sequences of images obtained from the spatially resolved emission spectra of fluorescent indicators. Especially, the possibility of monitoring processes with high spatial and temporal resolution has led to the enormous spread of this technique. Examples can be found

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

from such diverse fields as environmental physics (Volume 3, Chapter 33) to the broad range of life sciences including molecular biology, DNA-sequencing, neurobiology and biophysical studies of functional mechanisms in living cells. As living cells are highly compartimentalized on a nano- to micrometer scale and respond to environmental changes on a millisecond time scale, many of their properties can only be studied with a comprehensive approach to dynamic fluorescence imaging. With biological cellular preparations as examples, the present chapter discusses the most important properties of selected fluorescent indicators, microscopic techniques and essential steps for the required mathematical image analysis.

Particular emphasis is given to the use of fluorescence imaging techniques in the determination of intracellular ion concentrations in living cells under *in vivo* and *in vitro* conditions. For this application fluorescence techniques are the most popular tools. In addition to the high spatial and temporal resolution fluorescence imaging offers, sensitivity and selectivity for specific ions are among the largest benefits. Furthermore, the possibility of spectroscopic analysis allows direct information to be obtained about the molecules and the interaction with their environment. First, we want to give a brief account of the nature of fluorescence and its molecular origin.

12.2 Fluorescence

Fluorescence has been used by physicists and biochemists since the 1920s. Many biochemical intermediates are naturally fluorescent, for example, the enzyme cofactor NADH involves a drop in the fluorescence emission when oxidized to NAD^+, which can be used as a sensitive indicator for cellular metabolism [1].

Today, there is a rapid increase in fluorescence techniques in cellular biology, especially in monitoring intracellular signaling pathways. For example, the development of the ratiometric dyes by Prof. Roger Y. Tsien's group in the mid-1980s [2] has led to enormous progress in the accurate determination of intracellular calcium concentrations. Recently, the so-called *green fluorescent protein* (GFP), produced in chimeric target constructs with an attached fluorophore, extended as a very specific marker for genetic expression even further the use of fluorescent techniques in cell biology and molecular physiology [3].

12.2.1 Physical properties of fluorescence

Fluorescence is the result of a quantum mechanically "allowed" transition of electrons in certain molecules typically called *fluorophores* or *fluorescent* dyes from an excited state to the ground state. The energy

for the excitation of the dye is mostly supplied as photons by an excitation light source as, for example, high-pressure arc lamps or lasers. The typical fluorescence lifetimes, that is, the average time the fluorophore stays in the excited state, range from 10^{-9} s to 10^{-7} s . Following light absorption, several processes occur (see [4]). A fluorophore is usually excited to some higher vibrational level of either S_1 ,S_2, or S_3, as shown in Fig. 12.1, where S_i denotes the different electronic singlet states of the fluorophore. Mostly relaxation to the lowest vibrational level of S_1 occurs in the order of 10^{-12} s. This so-called *internal conversion* is basically a transfer of energy to the surrounding medium as heat. From this state, either fluorescence emission or radiationless decay to one of the vibrational states of S_0 occurs. The emission of fluorophores generally occurs at wavelengths longer than those of absorption. This "*Stokes shift*" is a result of several processes, including internal conversion of excited states to the lowest vibrational level of S_1 or *solvent relaxation* effects as shown in Fig. 12.1. The sensitivity of fluorescence techniques is based on the fact that the emitted fluorescence photons can be detected against a low background, separated from the excitation photons by the Stokes shift.

The energy for the excitation of the fluorophore can either originate from the absorption of a single photon with energy $h\nu_a$, where h is Planck's constant and ν is the frequency of the light, or from the absorption of either two photons with energy $h\nu_a/2$ or from multiple photon absorption. The absorption of two or more photons is used in nonlinear scanning microscopy and will be described in Section 12.4.4. The absorption probability for a single photon is linearly proportional to the initial excitation light intensity, whereas there is a quadratic dependence for the two-photon absorption process.

From the *law of Lambert-Beer* that states that the extinction for one photon absorption is proportional to the concentration of the absorbing species and to the optical path length, it can be derived that under the assumption of a dilute solution the fluorescence intensity I_f is proportional to the concentration c of the absorbing substance

$$I_f \propto \epsilon(\lambda) c x I_0 q \tag{12.1}$$

where $\epsilon(\lambda)$ is the wavelength-dependent extinction coefficient, x is the thickness of the sample volume, and I_0 is the excitation light intensity. The *quantum yield* q is defined as the ratio of fluorescence photons to the number of photons absorbed. Neglecting intersystem crossings in Fig. 12.1, the quantum yield can be written approximately as

$$q = \frac{\sigma}{\sigma + k} \tag{12.2}$$

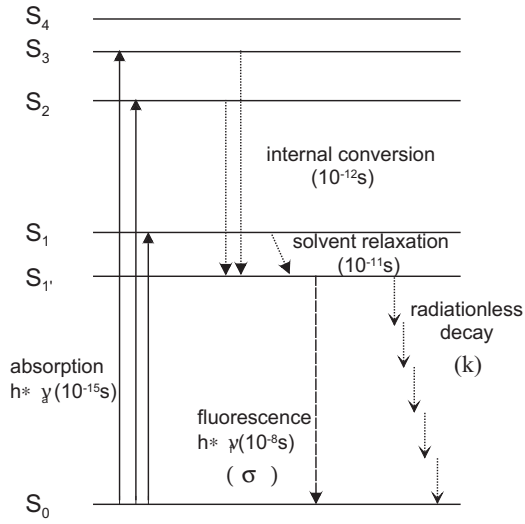


Figure 12.1: *Typical energy levels and time scales of transitions for a fluorophore. The ground, first and second electronic states are depicted by S_0, S_1 and S_2 further divided into various vibrational and rotational states that are omitted for simplicity; $S_{1'}$ is the lowered S_1 state due to solvent relaxation. The rate of fluorescence emission is denoted by σ and the rate of all possible radiationless decay processes by k. Possible intersystem crossings from the S_1 singlet state to the first triplet state T_1, which result in phosphorescence emission, are neglected. Adopted from [4].*

where σ denotes the rate of fluorescence emission and k denotes the rate of all possible radiationless decay processes.

If the fluorescence spectrum is different for the free and the bound form of the fluorescent dye the amount of ions complexed by the dye can be determined. The free ion-concentration can be subsequently obtained with the knowledge of the dissociation constant K_d of the dye-ion complexation reaction.

It is important to note that the fluorescence intensities are proportional to the concentration over only a limited range of concentrations of the fluorescent specimen; for larger concentrations the relation becomes nonlinear. Additionally, there are several attenuation effects, which largely decrease the observed fluorescence. When the fluorescence emission passes through the solution, it can be reabsorbed leading to a decrease in fluorescence intensity. This loss is called the *inner filter effect* and the attenuation of fluorescence emission increases with increasing concentrations and increasing thickness of the sample volume.

The *sensitivity* of a given fluorescent indicator not only depends on the quantum yield defined by Eq. (12.2), but also on the absorption characteristics of the fluorophore. The sensitivity S can be defined as

$$S = \epsilon(\lambda) I_0 q \tag{12.3}$$

and is an important parameter, for example, in choosing the optimum excitation wavelength, particularly when the main excitation band cannot be used due to technical limitations of the excitation light source.

Fluorescence quenching generally refers to bimolecular processes, which reduce fluorescence emission by inducing the relaxation of excited fluorophore molecules back into the ground state without the emission of a fluorescence photon. This includes self- or concentration-quenching, where one fluorophore is quenched by another, and collisional quenching, which is due to transient excited state interactions. The quenching effects include specific interactions with the solvent, which lead to a decrease in fluorescence intensity, or interactions with other substances present in the solution. The lifetime τ_0 of an excited state of a fluorophore without quenching is given by

$$\tau_0 = \frac{1}{\sigma + k} \tag{12.4}$$

and in the presence of a quenching substance with concentration c, the lifetime τ is given by

$$\frac{\tau}{\tau_0} = 1 + Kc \tag{12.5}$$

where K is the so-called quenching constant.

12.2.2 The oxygen quenching method

The decrease in fluorescence due to quenching can also be used for the quantitative determination of concentrations of substances, which act as quenchers. A major application is the measurement of oxygen concentrations in aqueous solutions. Oxygen quenches almost all known fluorophores via collisional quenching. The decrease in fluorescence intensity is described by the *Stern-Volmer equation* [4]:

$$I_f(c) = \frac{I_{f_0}}{1 + Kc} \tag{12.6}$$

where I_{f_0} is the fluorescence intensity without quenching, K is the quenching constant, and c is the concentration of the quenching substance.

This technique can be used in a variety of different scientific applications. In the field of life sciences it can be used to measure the oxygen

concentration in living cells and tissue [5]. The method is also successfully applied in environmental physics, where concentration fields of dissolved oxygen are measured in the boundary layer at the water surface to study the mechanisms of air-water gas transfer [6].

12.3 Fluorescent indicators

The optical methods to quantify intracellular ion concentrations can be discussed very well using the example of Ca^{2+}-sensitive techniques, as the investigation of intracellular Ca^{2+}-levels has been the major application. The three most important techniques to quantify intracellular Ca^{2+}-levels are measurements with photoproteins, absorption measurements and fluorescence measurements (see Thomas [7]). *Photoproteins* are obtained from luminescent organs of coelenterates and emit light when reacting with Ca^{2+}. This method was first used by Ridgway and Ashley [8], who could measure intracellular Ca^{2+}-levels in muscle fibers.

The development of Ca^{2+}-sensitive dyes (azo-dyes, namely, arsenazo III and antipyrylazo III) in the mid-1970s has allowed the start of absorption measurements and has been particularly useful in studying the fast Ca^{2+}-transients in skeletal muscle fibers, although the absorption changes are difficult to detect and to interpret, especially in larger cellular preparations.

The introduction of fluorescent indicators has greatly facilitated the measurement of intracellular ion concentrations and, combined with the various microscopic techniques available today, they provide the most powerful tools.

12.3.1 Calcium indicators

Many useful calcium indicators have been developed in Tsien's lab and are derivatives of the Ca^{2+} buffer BABTA, including Quin-2, Fura-2, Indo-1, Fluo-3 and Rhod-2 (see Haugland and Minta [9]). They are chemically designed fluorescent dyes having an additional Ca^{2+}-complexing moiety like EDTA or EGTA. For example, Fluo-3 is obtained by using fluorescein and EDTA.

Calcium indicators need to have high affinities, as typical Ca^{2+}-concentrations in cells are very low (from 10-100 nM at rest to 1-10 μM during activation). Additionally, the selectivity against Mg^{2+}-ions has to be very high as these ions competitively bind and are present in much higher concentrations (about 1 mM).

Since the development of the *ratiometric dyes*, for example, Fura-2 and Indo-1, it is possible, even using conventional fluorescence microscopy, to accurately determine intracellular calcium ion concentra-

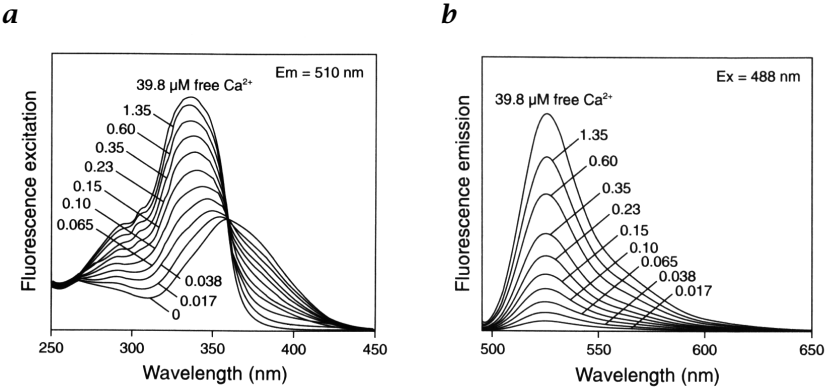


Figure 12.2: ***a*** *Excitation spectrum of Fura-2. For an excitation wavelength of 340 nm, the fluorescence signal increases with increasing Ca^{2+}-concentration and the opposite effect can be seen for an excitation wavelength of 380 nm. The fluorescence signal is independent of the calcium concentration at the isosbestic point at $\lambda = 360$ nm;* ***b*** *emission spectrum of Fluo-3. Figure courtesy of Molecular Probes Inc., Eugene, OR [10].*

tions. In addition to higher fluorescence intensities and better calcium selectivity, these dyes exhibit a strong shift in their excitation or emission wavelength upon binding of calcium. Fura-2 is designed to shift its wavelength of excitation to shorter wavelengths with the binding of Ca^{2+}-ions. As seen in Fig. 12.2, the excitation maximum for the free dye is at a wavelength of about 370 nm and shifts to 340 nm with the binding of Ca^{2+}-ions, a much larger shift than in the emission spectrum. This allows *dual-excitation ratio measurements* by sampling the fluorescence intensities at two appropriate wavelengths λ_1 and λ_2 (mostly 340 nm/380 nm). Forming the ratio R of the fluorescence emissions I_{f_1} and I_{f_2}, the calcium concentration can be calculated according to the equation originally derived by Grynkiewicz et al. [2]:

$$[Ca^{2+}] = K_d \beta \frac{R - R_{\min}}{R_{\max} - R} \tag{12.7}$$

where K_d is the dissociation constant of the Fura-calcium complex; $R_{\min}$ and $R_{\max}$ are the ratios of the fluorescence emission in the virtual absence or with a saturating amount of calcium; and β corresponds to the ratio of fluorescence emission of the free dye to the calcium bound dye measured at the second wavelength. Thus, calcium concentrations can be calibrated independently of the dye concentration, specimen thickness and illumination intensity. *Fura-2* is a chemical derivative of the calcium buffer BAPTA and the absorption maximum is in the near UV with an extinction coefficient in the range of $2\text{-}3 \times 10^4\,\mathrm{M^{-1}cm^{-1}}$. The

emission maximum is at a wavelength of 512 nm for the free dye and shows a shift to a wavelength of 505 nm for the Ca^{2+}-complexed dye. The apparent K_d of the Ca^{2+}-Fura-2 complex, is in the range of 135-300 nM, strongly depending on the ionic strength of the solution and several other factors, such as, for example, viscosity [11].

The nonratiometric Ca^{2+}-indicator *Fluo-3* (see Fig. 12.2) offers some advantages over UV-excitable indicators, such as it is excited in the visible part of the spectrum. It can be excited with the 488 nm line of an argon-ion laser and, therefore, it is frequently used in laser scanning microscopy. Due to the longer excitation wavelength, there is reduced photodamage, light scatter and reduced cell autofluorescence. Unlike Fura-2, Fluo-3 is essentially nonfluorescent unless bound to Ca^{2+} and exhibits a more than 100-fold increase in fluorescence upon complexation. The K_d of ~ 400 nM allows the detection of Ca^{2+}-concentrations of more than 10 μM without saturation.

A new generation of calcium indicators, known as "*cameleons*," has been recently developed by Tsien's group based on the *green fluorescent protein* (GFP) [3]. These indicators combine the brightness of fluorescent indicators with the target-ability of biosynthetic indicators and are generated *in situ* by gene transfer into the cells. This can be used to target the indicator to specific intracellular locations with more or less molecular resolution, allowing the monitoring of Ca^{2+}-signals that are extremely localized with very sharp gradients to their surrounding.

12.3.2 Other ions and membrane potential

The very fundamental process of electrical excitation in nerve and muscle cells is governed by diffusion potentials due to intracellular and extracellular differences for Na^+-, K^+- Cl^-- and Ca^{2+}-ion concentrations and due to the regulation of membrane permeabilities for these ions. These mechanisms also play very important functional roles in nonexcitable cells in the human body and almost all plant and animal cells. Therefore, a number of fluorescent probes have been developed in addition to the important Ca^{2+}-indicators, which are sensitive to the forementioned ions and also for the other regulating ions, namely, Mg^{2+} and H^+ (for review, see Mason [12]). All these optical probes can be categorized based on their ion dependent excitation and emission spectra as single wavelength, dual (wavelength-) excitation and dual (wavelength-) emission dyes. There is still a largely untapped potential in particular for the combination of those dyes to provide a most powerful tool usable with dynamic imaging to monitor fast biologically or clinically important concentration changes for several ions simultaneously to gain insight into complex regulatory processes. Often, dynamic fluorescence imaging techniques are combined with high-resolution electrophysiological experiments measuring transmembrane currents through

specific ion channel proteins. The electrophysiological techniques use either intracellular microelectrodes (tip diameter<0.5 μm) or external patch electrodes of similar size. It is very difficult to use these electrodes for intracellular organelles or for spatially resolved measurements. Therefore, potentiometric fluorescent probes were developed to record in a noninvasive way the membrane potential changes with high spatial resolution (see Loew [13], Wu and Cohen [14]). These dyes can be divided into fast response probes, which can be used to measure membrane potential changes on the millisecond time scale, as, for example, in cardiac cells and neurons in the central nervous system. The slow-response dyes offer in general a much larger response than the fast dyes, but with a slower kinetic and they are particularly useful for nonexcitable cells and cellular organelles.

12.3.3 Dye kinetic and buffering

It should be noted that fluorescent indicators all have an intrinsic kinetic delay to changes in their environment. The quantitative analysis of fluorescence changes, therefore, generally has to consider the kinetic on- and off-rate constants $s(k_{\text{on}}, k_{\text{off}})$ of the fluorophore-ion interaction, which are related to the dissociation constant K_D by the following relation:

$$K_D = k_{\text{off}}/k_{\text{on}} \tag{12.8}$$

A second aspect of ion binding to a fluorophore is that thereby each fluorescent indicator acts as an *ion buffer* itself. In many cases this can result in pronounced alterations of the complex ion distribution in cellular systems. Therefore, the buffering effects should be generally taken into account in the quantitative analysis of fluorescence signals.

12.3.4 Photobleaching and photodamage

The irreversible destruction of fluorophores (*photobleaching*) is mostly influenced by the excitation illumination intensity, but also by other experimental and surrounding environmental conditions of the dye (e.g., impeded diffusional exchange and *compartmentalization* of the dye, pH, formation of radicals and oxidation and radical formation, etc.). Therefore, high excitation intensities should be avoided, mainly by increasing the detection sensitivity, or by detecting the fluorescence emission over the broadest possible wavelength band.

Photodamage is the result of interactions of the excitation photons with the specimen. There are many different processes, which can result in the damage of cells, proteins or DNA (see Niemz [15] for a general discussion of laser tissue interaction). In general, it can be stated that

photodamage is stronger the higher the illumination intensities and the shorter the excitation wavelengths.

12.3.5 Dye loading of cells

Fluorescence dyes also offer the advantage that cells can relatively easily be loaded with the indicator of choice. Normally, fluorescent probes are polar and, therefore, unable to cross the cell membrane, which is a lipid bilayer. Basically, there are two ways to insert the dye into a cell. In the first method, the dye is directly injected into the cell with a micropipette. In the second method, the dye is chemically transformed to the lipophilic acetoxymethyl (AM)-ester (e. g., Fura-2-AM, Fluo-3-AM). These dyes have their polar carboxy groups esterized and, therefore, they are able to cross the cell membrane by diffusion. In this form the dye can not bind Ca^{2+}-ions and is not fluorescent. Inside the cell, the dye-ester is hydrolyzed to its free polar form by cytosolic esterases or compartmentalized enzymes, and the free nonlipophilic dye is trapped in the interior of the cell.

12.4 Microscopic techniques

In the past decade, enormous progress has been made in the development of very sensitive fluorescence imaging techniques. At present, it is possible to choose among various methods for recording intracellular ion concentrations, which allows a great flexibility in selecting an appropriate technique for a particular application.

The question of a best method for intracellular ion imaging can not be answered in general, as all techniques have their strengths and limitations. Among them are spatial and temporal resolution, photodamage caused in the specimen, and important enough financial aspects and ease of use. In the following we will present examples of fluorescence imaging techniques applied to study relevant biophysical and physiological questions.

12.4.1 Conventional fluorescence microscopy

Quantitative *fluorescence microscopy* has made enormous progress including the development of the ratiometric fluorescent probes. The ratiometric fluorescence imaging method is now very well established and a large amount of literature is available, which discusses the method and its potential pitfalls (e. g., Silver et al. [16]; a collection of papers can be found in Volume 11 of *Cell Calcium*, 1990).

As already described in Section 12.3.1 this method allows the quantitative recording of spatial and temporal ion concentration changes with

commercially available standard equipment. For ratiometric ion concentration determination a typical setup consists of an epi-fluorescence microscope equipped either with a dual excitation (monochromators, or interference filter-based devices, for example, Uttenweiler et al. [17]) or a dual emission device.

The detectors can either be photomultipliers or the various types of sensitive charge-coupled-device (CCD) cameras. The temporal resolution is limited by the time necessary for excitation wavelength changes and by CCD readout, which is typically done with video frequency.

Recently, 12 to 14 bit digital CCD cameras with custom timing and readout have become available, which generally allow much faster frame rates without the need of image intensifiers. Therefore, it becomes more and more feasible that conventional fluorescence imaging measurements can be carried out with a very high temporal resolution. This fast recording method was used in Fig. 12.3 to record Ca^{2+}-waves in spontaneously activated rat cardiac myocytes.

In general, conventional microscopic fluorescence signals not only contain information from the focal plane, but also from the entire cone of the excitation light. Therefore, spatial gradients are mostly underestimated [18, 19]. In the next section, we will describe the methodological approach of image deconvolution to reduce the effect of out-of-focus information in fluorescence images.

12.4.2 Image deconvolution

With the help of sophisticated algorithms used in digital image analysis, it is now possible to achieve a much better depth resolution in conventional microscopy. Several deblurring techniques remove out-of-focus information either by *"nearest-neighbors"* or *"no-neighbors" algorithms* [20, 21]. Nearest-neighbors algorithms assume that the majority of out-of-focus information comes from adjacent sections of the specimen. This information is estimated by sampling images from adjacent sections and blurring them with the out-of-focus point spread function. By subtraction of this contribution from the original image, the remaining signal predominantly reflects in-focus information. If one assumes that all the light in an observed image comes from the in-focus image and the two adjacent image planes, the observed image o_j can be written as [21]:

$$o_j = i_j * s_0 + i_{j+1} * s_1 + i_{j-1} * s_{-1} \tag{12.9}$$

where i_j is the in-focus image; i_{j+1} and i_{j-1} are the images in the neighboring planes; s_0 is the in-focus point spread function (PSF); s_1 and s_{-1} are the out-of-focus point spread functions; and $*$ denotes the convolution operation. By taking the Fourier transform the equation simplifies

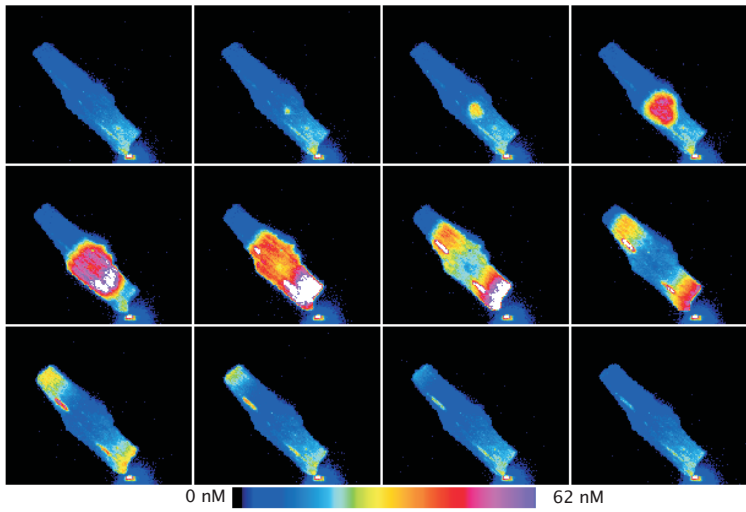


Figure 12.3: *Example of a fast spatially resolved Ca^{2+}-image sequence. Cardiac myocytes were labeled with the Ca^{2+}-sensitive indicator Fluo-3 (2 µM) and spontaneous Ca^{2+}-waves propagating inside the myocyte can be seen. The sequence was recorded with a MERLIN system and an Astrocam frame transfer camera (Life Science Resources Inc., Cambridge, UK) with an EEV37 CCD chip read out at 5.5 MHz to capture the images. [Figure courtesy of Dr. B. Somasundaram, Life Science Resources Inc. and Dr. N. Freestone of Babraham Institute, Babraham, Cambridge UK]; (see also Plate 3).*

to

$$O_j = I_j \cdot S_0 + I_{j+1} \cdot S_1 + I_{j-1} \cdot S_{-1} \tag{12.10}$$

where S_1, S_{-1} and S_0 are the Fourier transforms of the respective point spread functions, the optical transfer functions (OTF).

If one assumes that the true images in the adjacent planes can be replaced by the observed images and that $S_1 \approx S_{-1}$, that is, $I_{j+1}S_1 \approx O_{j+1}S_1$ and $I_{j-1}S_{-1} \approx O_{j-1}S_1$, Eq. (12.10) can be rewritten as

$$O_j = I_j \cdot S_0 + O_{j+1} \cdot S_1 + O_{j-1} \cdot S_1 \tag{12.11}$$

and subsequently

$$I_j = (O_j - c \cdot (O_{j+1} + O_{j-1}) \cdot S_1) \cdot S_0^{-1} \tag{12.12}$$

where c is an empirical constant and S_0^{-1} is the inverted in-focus OTF. In the range where the OTF is close to zero, the inverted OTF is not suited for the inverse filtering. The use of S_0^{-1} as a filter would result

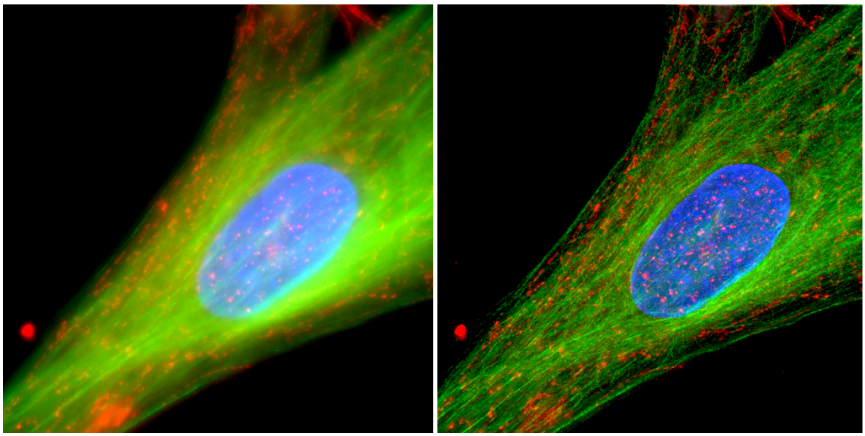


Figure 12.4: *Example of a nearest-neighbors deblurring algorithm. The image of a human skin fibroblast consists of three separate 12 bit gray-scale images, each recorded with a different fluorescent dye. The cell has been processed for double immunofluorescence and counterstained with Hoechst 33258 for DNA. Microtubules in the cell are localized with a monoclonal IgG antibody to beta-tubulin followed by a secondary antibody tagged with FITC. Mitochondria are localized with a monoclonal IgM antibody to a novel protein, followed by a secondary antibody tagged with Texas Red. For FITC, excitation filter = 485 nm, with a barrier filter at 530 nm. For Texas Red, the excitation filter was 560 nm, with a 635 nm barrier filter. Traditional UV filters were used for the Hoechst dye. Images were captured at six depths at 1 µm steps and each color channel was deblurred separately resulting in the deconvolved image at the right side. [Figure courtesy of Dr. R. Zinkowski, Molecular Geriatrics Corp., Vernon Hills, IL and Dr. Chris MacLean, VayTek Inc., Fairfield, IA]; (see also Plate 4).*

in a domination of noise at high spatial frequencies. Therefore it is replaced by a Wiener inverse filter [21].

Figure 12.4 is an example of a nearest-neighbors deblurring algorithm applied to a multiple stained human skin fibroblast. The color image is the overlay of three separate fluorescence images, each recorded with a different fluorophore. A nearest-neighbors deblurring algorithm was applied to each separate spectral channel, with 1 µm stepsize for six image sections. The resulting image is an impressive example of the power of this technique even with multispectral image data, as it reveals many more details than the unprocessed image.

The no-neighbors deblurring schemes make the additional assumption that the blurred image $O_j S_1$ can be used instead of the blurred neighbors $O_{j+1} S_1$ and $O_{j-1} S_1$. Therefore, Eq. (12.12) can be rewritten as

$$I_j = (O_j - 2c \cdot S_1 \cdot O_j) \cdot S_0^{-1} \tag{12.13}$$

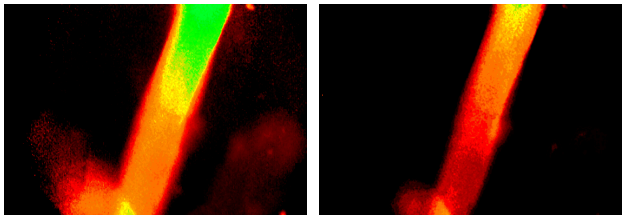


Figure 12.5: *Example of a no-neighbors deblurring algorithm. A muscle fiber from Xenopus laevis M. lumbricalis with a diameter of 100 µm was stained with the ratiometric Na^+-indicator SBFI-AM (340 nm/380 nm ratio). The deblurred image on the right contains less out-of-focus information than the original ratio image on the left; (see also Plate 5).*

where the inverse filtering with S_0^{-1} is again replaced by a Wiener inverse filter; S_0 and S_1 are modeled by theoretical transfer functions. These functions are calculated from characteristic parameters of the optical setup including the wavelength of the emitted light, pixel size of the detection unit, the aperture of the objective lens and the index of refraction. The negative values introduced by the filtering are set to zero by a threshold operation [21].

Because no-neighbors schemes model the information from adjacent sections from the image itself, they lack the need to acquire images in different sections. Therefore, these algorithms are well suited for high temporal resolution studies and an example of such an algorithm is given in Fig. 12.5. Although no sectioning is necessary, out-of-focus information is effectively reduced in the deblurred image.

However, in general, it should be noted that deblurring techniques used with conventional imaging still are not able to resolve structures parallel to the optical axis, as the optical transfer function is zero for structures in the axial direction.

12.4.3 Confocal microscopy

For detecting highly localized gradients in ion concentrations, more expensive equipment is needed as the observed image is contaminated with out-of-focus information. One solution is to work with a scanning confocal microscope, where out-of-focus information is reduced by using a detection pinhole. This approach is the most commonly used, when depth resolution is a crucial point. In confocal microscopes the fluorescently stained specimen is illuminated with a diffraction limited spot created by a laser with appropriate wavelength and the discrimination of out-of-focus information is achieved by the detection pinhole. With the confocal setup the intensity discrimination in z-direction can be said to be roughly proportional to $1/z^4$ [22]. An image is obtained

by scanning the image plane in xy-direction with the focused laser spot via galvanometer-driven mirrors and by detecting the fluorescence emission with a photomultiplier or other sensitive detector through the detection pinhole. A 3-D image can be reconstructed by sampling images in different z-planes (see also Chapter 21). Confocal images yield greatly improved axial resolution and also improved lateral resolution.

Modern point-scanning confocal microscopes allow frame rates up to 30 frames per second with an image size of 128×128 pixel. Therefore, it is possible to follow fast intracellular ion changes in space and time, as, for example, demonstrated in the detection of spherical Ca^{2+}-waves in rat cardiac myocytes using the Ca^{2+}-indicator Fluo-3 [23].

Potential problems of the technique especially arise with dynamic low light level applications in highly scattering specimen. For low light level fluorescence the amount of light detected may be too weak to yield a reasonable signal-to-noise ratio. Opening the detection pinhole, which results in a higher signal will lead to the loss of the advantages of increased axial and lateral resolution. When increasing the illumination power, severe *photodamage*, *photobleaching* and fluorophore saturation have to be avoided.

A significantly higher temporal resolution can be obtained using the *linescan mode* of confocal microscopes. Although only information of a 1-D line is available it is possible to reconstruct the full 3-D spatio-temporal structure of ion concentration changes.

An impressive example of the power of this technique is the detection of *Ca^{2+}-sparks*, the elementary events of Ca^{2+}-release in skeletal and heart muscle [24, 25, 26]. Ca^{2+}-sparks are the elementary events of Ca^{2+}-release through ion channels of the ryanodine family from the sarcoplasmic reticulum of muscle fibers. Figure 12.6 shows a very elegant experiment to analyze small event Ca^{2+}-release in skeletal muscle (for details see Shirokova and Rios [27] and Tsugorka et al. [26]), which are even smaller than elementary events in heart muscle. The spatial resolution of these measurements reach 300 nm, close to the theoretical resolution of the confocal microscope and the temporal resolution is around 2 ms. The high spatial and temporal resolution allows the reconstruction of the profile of these small elementary Ca^{2+}-release events, as shown in Fig. 12.6.

12.4.4 Two-photon microscopy

Multiphoton and especially *two-photon laser scanning microscopy* is a newly developed technique [28, 29]. Instead of absorbing one photon and emitting fluorescence that is Stokes-shifted to a longer wavelength, two-photon excitation involves the absorption of two photons simultaneously, with mostly the same wavelength and, thus, the same energy. Therefore, the fluorescence wavelength is shorter than the excitation

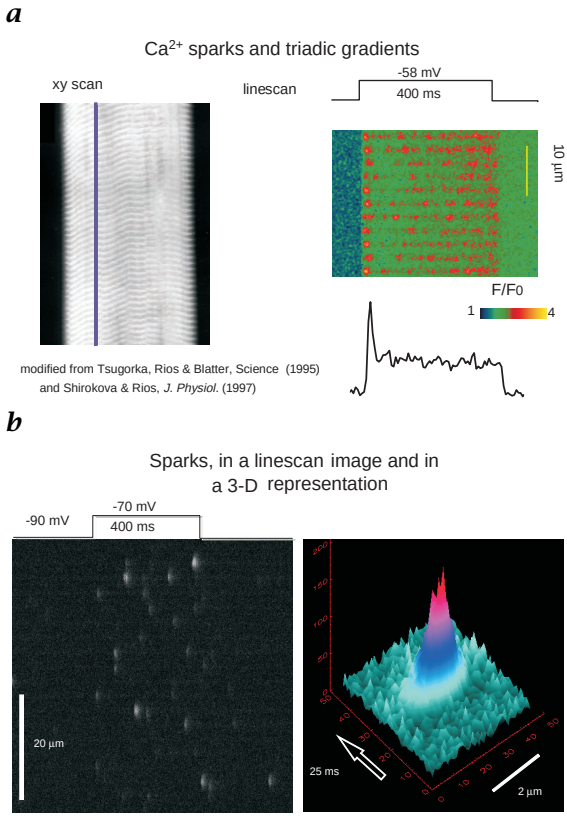


Figure 12.6: *Ca^{2+}-sparks measured in skeletal muscle fibers from Rana pipiens with the fluorescent indicator Fluo-3 under voltage clamp conditions:* ***a*** *line-scan image of Fluo-3 fluorescence upon 400 ms depolarization, F/F_0 is the normalized fluorescence;* ***b*** *3-D representation of a spark as reconstructed from the linescan image data. [Figure courtesy of Prof. E. Rios, Rush University, Chicago, IL, USA]; (see also Plate 6).*

wavelength, opening UV-excited dyes to excitation in the visible part of the spectrum.

The axial resolution is determined by the quadratic dependence of the two-photon absorption rate and fluorescence intensity on local excitation laser power. Using a pulsed near-infrared fs-laser the two-photon absorption probability becomes appreciable for the excitation of the fluorescent dye and the fluorescence is highly localized in the vicinity of the focal point. Therefore, the depth resolution of this technique is achieved without the confocal aperture, thus enabling efficient fluorescence photon collection. In confocal scanning microscopy the fluorescent dye is excited in the entire cone of the laser beam. Scat-

tered fluorescence photons seem to originate from out-of-focus planes and therefore they are rejected by the detection pinhole. Since in two-photon microscopy the excitation is limited to the focal region, all emitted fluorescence photons can be detected, resulting in a much better fluorescence photon collection efficiency, further improved by the lack of the necessity for a descanning optics. The localization of the excitation to a very small volume, in contrast to the excitation in the entire light cone in confocal microscopy, dramatically reduces the effects of *photodamage* of the tissue and *photobleaching* of the dye, although basic studies of potential thermodynamic damages still have to be carried out.

Another advantage of the two-photon excitation method is that the infrared photons have a deeper penetration depth into biological specimen and, therefore, deep tissue and thick cellular preparations can be studied.

As two-photon microscopy is a very new technique, two-photon excitation setups are still very expensive and rarely commercially available. Additionally, there is still incomplete data on the two-photon absorption and fluorescence properties of commonly used fluorophores.

Figure 12.7 shows an example of the imaging strength of two-photon laser scanning microscopy in highly scattering tissue. A neocortical layerV pyramidal cell in a brain slice was imaged using two-photon excitation with 90 to 110 fs pulses at 76 MHz from a Ti:Sa-Laser operated at a wavelength centered at 840 nm coupled into an upright microscope (BX50Wi, Olympus) equipped with a 60×-objective with high infrared transmission. The calcium transients in a basal dendrite were measured under current clamp conditions and the action potential was elicited by current injection of 500 to 1000 pA for 10 ms into the soma. Again, the use of the linescan mode allows the recording of the very fast calcium response to the physiological stimulation by an action potential under close to *in vivo* conditions.

12.4.5 Miscellaneous techniques

Several other microscopic techniques exist that yield improved spatial and temporal resolution. Two examples shall briefly be mentioned.

Total internal reflection microscopy. The axial resolution in fluorescence microscopy can be dramatically increased with the use of the *evanescent field* for the excitation of the fluorophore [31]. This technique is especially valuable when single molecules have to be studied and when surface-associated processes have to be visualized. Normally, even with confocal or nonlinear fluorescence microscopy, the fluorescence signal would consist of many layers of molecules, resulting in a blurred image, where fluorescent changes of a single molecule

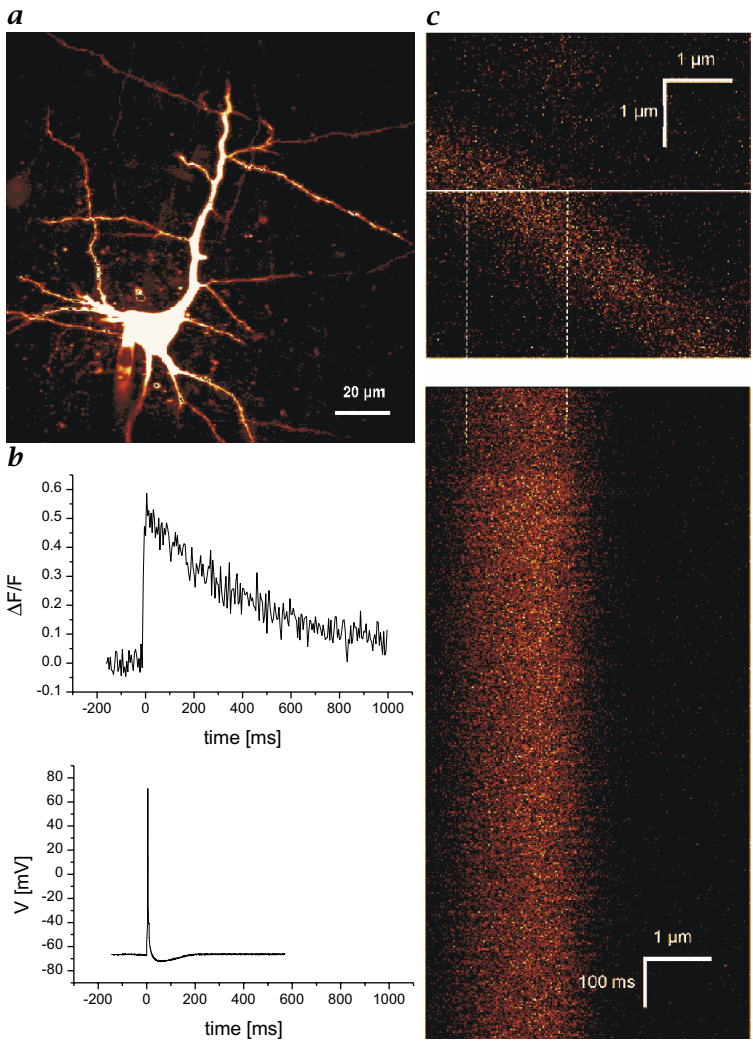


Figure 12.7: *Example of two-photon microscopy in brain slices.* ***a*** *A neocortical layerV pyramidal cell in a rat brain slice was filled via somatic whole-cell patch pipettes with the calcium indicator Calcium Green-1 (100 µM) or Oregon Green 488 BAPTA-1 (100 µM);* ***b*** *upper trace: calcium fluorescence transient evoked by a single backpropagating dendritic action potential; lower trace: Electrophysiological recording of the AP with somatic whole cell recording in current-clamp mode;* ***c*** *Linescan through a basal dendrite: fluorescence was recorded in linescan-mode. Upper picture: The line in the xy-image shows the position of the linescan. Lower picture: The linescan had a length of 1160 ms. All points in one line between broken lines were averaged. (Figure courtesy of Helmut Köster, Max-Planck-Institut für Medizinische Forschung, Heidelberg; see Köster and Sakmann [30]); (see also Plate 7).*

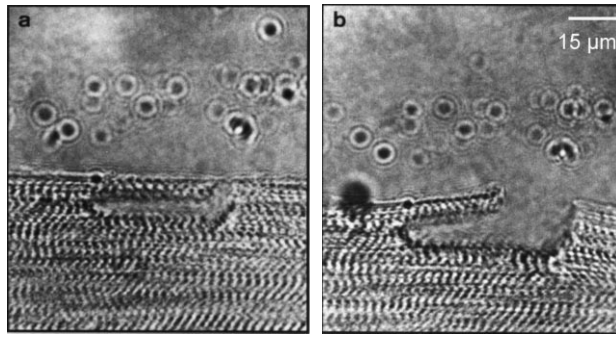


Figure 12.8: *Example of a functionally intact UV-laser microdissected myofibrillar preparation from Xenopus laevis muscle (panel **a**). The small myofibrillar bundle retains the ability to contract as demonstrated by the release of a caged Ca^{2+}-compound (nitr-7) in the vicinity of the bundle by a UV-laser pulse (panel **b**). Taken from Veigel et al. [33].*

or processes at biological surface are overwhelmed by the background fluorescence. By means of total internal reflection microscopy it was even possible to visualize ATP-turnover reactions at the single myosin molecule level, which is important for answering the question of how ATP hydrolysis is coupled to mechanical work at the level of the single molecule [32].

UV-laser microdissection. In addition to the development of the new microscopic techniques with improved spatial and temporal resolution described in the foregoing, the microdissection of large preparations can also result in improved optical properties. A UV-laser (N_2-laser, 337 nm, with 1 μJ energy per pulse and 20 Hz repetition rate) coupled into the fluorescence entry of an inverted microscope and focused onto the specimen via a high NA-objective can be used to dissect functionally intact areas of cellular preparations with a precision unmatched by other techniques. The effective cutting diameter can be shown to be as small as 0.5 μm. The selective preparation of functionally intact subunits of cellular probes by UV-laser microdissection offers the possibility of recording fluorescence data from samples with minimal thickness, thus avoiding many artifacts and potential problems that arise from the thickness of the preparation. In muscle fibers, for example, laser microdissected myofibrillar preparations with intact sarcoplasmic reticulum have been isolated [33]. Due to their small diameter (around 2 μm) they are ideally suited for quantitative microscopic measurements and additionally offer the advantage that diffusional delays are dramatically reduced.

12.5 Analysis of fluorescence images

The analysis of high temporal and spatial resolution fluorescence images acquired with the various methods described in the foregoing requires sophisticated techniques in order to derive the unbiased biophysical and physiological parameters from the spatiotemporal changes in fluorescence.

First, the analysis of the fluorescence images should, whenever possible, involve the correction for experimental distortions or errors inherent to a certain acquisition technique. For a fluorescence microscope setup this includes, for example, the correction of optical inaccuracies. When the *point spread function* of the microscope is determined experimentally, the distorting effects can at least partly be reversed [34]. Second, a detailed analysis of fluorescence images also has to include the corrections for the interaction of the *fluorescent indicator* with its environment. Only in few cases, a fluorescence transient is the direct reflection of the underlying chemical, biophysical or physiological process, which is to be described. Mostly, the interaction of the fluorescent indicator with the various constituents of the experimental system has to be characterized in detail. For intracellular ion concentration determinations this involves, for example, studies of the buffering effects of the fluorophore. Because the indicator itself acts as an *ion buffer*, significant alterations in the intracellular concentration distributions very often result. Also the kinetic properties of the indicators have to be considered, as fluorophores have more or less pronounced delays in their binding to ions due to the limited kinetic on- and off-rate constants for ion binding. Furthermore, the spatiotemporal distribution of ions can only be correctly described by comprehensive spatially resolved mathematical models of ion distributions, which account both for the limited kinetic of the dye and the spatially inhomogeneous and compartmented structure of all cellular preparations. An example of such a detailed mathematical *model-based analysis of fluorescence images* is given in Volume 3, Chapter 34 of this handbook.

Once corrected for these biases, the broad spectrum of techniques described in this handbook for an automated image analysis can be applied. Fluorescence images in general pose high demands on algorithms used for their analysis due to their high level of noise, all the more considering the analysis of very fast processes with high temporal resolution and, when increasing the spatial resolution, down to the molecular scale. An example of *molecular visualization* with the fluorescence imaging technique is shown in Fig. 12.9, where the movement of actin filaments (diameter 5 nm) labeled with rhodamine-phalloidin over a myosin-decorated surface is shown. The images were captured with video rate; due to the high temporal resolution, these images exhibit significant noise levels that pose very high demands on algorithms

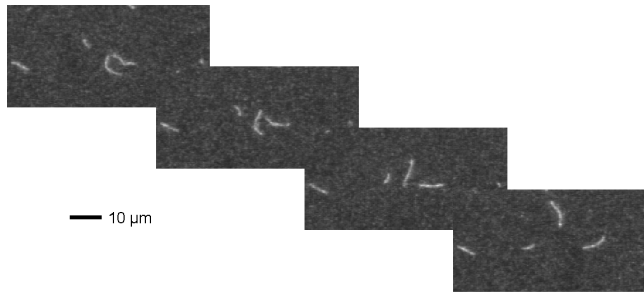


Figure 12.9: *Example of molecular visualization in an in vitro motility assay. The movement of rhodamine-phalloidin labeled actin filaments over a myosin-decorated surface is visualized and yields information about the basic interaction of the motor proteins actin and myosin. For display each image is the average of three raw images. From Uttenweiler et al. [35].*

detecting the velocity of actin filament movement (see Uttenweiler et al. [35]).

It should be noted that many image analysis techniques discussed in this book have been successfully applied to noisy fluorescence images to yield an accurate description of the chemical, biophysical or physiological processes of interest for the various fluorescence imaging applications.

12.6 Summary

This chapter presents and discusses new methods and applications of dynamic fluorescence imaging. Also, we aimed to draw the readers attention in particular to aspects of the acquisition and analysis of high spatially and temporally resolved fluorescence images. The reader is also referred to the great amount of recent literature, which covers the majority of fluorescence imaging aspects (the citations in this paper shall be seen as some starting points).

The great amount of fluorescence imaging techniques has significantly improved the choice of a suitable technique for chemical, biophysical and physiological investigations. Although the temporal and spatial resolution has dramatically increased, the unbiased information about the underlying processes can only be gained with powerful mathematical models, which account for experimental inaccuracies, fluorescent indicator properties and the complex nature of molecular and cellular processes.

As the field of dynamic fluorescence imaging is steadily growing, many more improvements, new techniques, and sophisticated methods for the analysis will certainly be available in the future.

Acknowledgments

The authors would like to thank the following persons for their contribution of figures: Dr. R.P. Haugland and Dr. I. Johnson, Molecular Probes Inc., Eugene, OR, USA; Dr. B. Somasundaram and Dr. W.T. Mason, Life Science Resources, Cambridge, UK; Dr. N. Freestone, Babraham Institute, Babraham, Cambridge, UK; H. Ostermann, Chromaphore GMBH, Duisburg, Germany; Dr. R. Zinkowski, Molecular Geriatrics Corp., Vernon Hills, IL, USA; Dr. C. MacLean, VayTek Inc., Fairfield, IA, USA; Prof. E. Rios, Rush University, Chicago, USA; Dipl.-Phys. H. Köster, Max-Planck Institut für Medizinische Forschung, Heidelberg, Germany. The authors would additionally thank Dr. M.J. Salzer, Institute of Biochemistry, Heidelberg, Germany, for careful reading of the manuscript and helpful comments.

12.7 References

[1] Master, B. and Chance, B., (1993). Redox confocal imaging: intrinsic fluorescent probes of cellular metabolism. In *Fluorescent and Luminescent Probes for Biological Activity*, W. Mason, ed., pp. 44–57. London: Academic Press.

[2] Grynkiewicz, G., Poenie, M., and Tsien, R., (1985). A new generation of Ca^{2+} indicators with greatly improved fluorescence properties. *The Jour. Biological Chemistry*, **260**:3440–3450.

[3] Miyawaki, A., Liopis, J., Heim, R., McCaffery, J., Adams, J., Ikural, M., and Tsien, R., (1997). Fluorescent indicators for $Ca2^{+}$ based on green fluorescent proteins and calmodulin. *Nature*, **388**:882–887.

[4] Lakowicz, J. R., (1983). *Principles of Fluorescence Spectroscopy*. New York: Plenum Press.

[5] Barnikol, W., Burkhard, O., Trubel, H., Petzke, F., Weiler, W., and Gaertner, T., (1996). An innovative procedure of oxygen detection in medicine, biology, environmental research and biotechnology based on luminescence quenching. *Biomed. Tech. Berl.*, **41(6)**:170–177.

[6] Münsterer, T., Mayer, H. J., and Jähne, B., (1995). Dual-tracer measurements of concentration profiles in the aqueous mass boundary layer. In *Air-Water Gas Transfer, Selected Papers, 3rd Intern. Symp. on Air-Water Gas Transfer*, B. Jähne and E. Monahan, eds., pp. 637–648. Hanau: Aeon.

[7] Thomas, M. V., (1982). *Techniques in Calcium Research*. London: Academic Press.

[8] Ridgway, E. B. and Ashley, C. C., (1967). Calcium transients in single muscle fibers. *Biochem. Biophys. Res. Commun.*, **29(2)**:229–234.

[9] Haugland, R. and Minta, A., (1990). Design and application of indicator dyes. In *Noninvasive Techniques in Cell Biology*, J. Foskett and S. Grinstein, eds., pp. 1–20. New York: Wiley-Liss.

[10] Haugland, R., (1996). *Handbook of Fluorescent Probes and Research Chemicals*. Eugene, OR: Molecular Probes Inc.

[11] Uto, A., Arai, H., and Ogawa, Y., (1991). Reassessment of Fura-2 and the ratio method for determination of intracellular Ca^{2+} concentrations. *Cell Calcium*, **12**:29–37.

[12] Mason, W., (1993). *Fluorescent and Luminescent Probes for Biological Activity*. London: Academic Press.

[13] Loew, L., (1993). Potentiometric membrane dyes. In *Fluorescent and Luminescent Probes for Biological Activity*, W. Mason, ed., pp. 150–160. London: Academic Press.

[14] Wu, J.-Y. and Cohen, L., (1993). Fast multisite optical measurement of membrane potential. In *Fluorescent and Luminescent Probes for Biological Activity*, M. Mason, ed., pp. 389–404. London: Academic Press.

[15] Niemz, M., (1996). *Laser tissue interactions*. Heidelberg: Springer Verlag.

[16] Silver, R. A., Whitaker, M., and Bolsover, S. R., (1992). Intracellular ion imaging using fluorescent dyes: artifacts and limits to resolution. *Pfluegers Arch.*, **420**:595–602.

[17] Uttenweiler, D., Wojciechowski, R., Makabe, M., Veigel, C., and Fink, R. H. A., (1995). Combined analysis of intracellular calcium with dual excitation fluorescence photometry and imaging. *Optical Engineering*, **34(10)**: 2864–2871.

[18] Duty, S. and Allen, D., (1994). The distribution of intracellular calcium concentration in isolated single fibres of mouse skeletal muscle during fatiguing stimulation. *Pflügers Arch.*, **427**:102–109.

[19] Uttenweiler, D., Weber, C., and Fink, R. H. A., (1998). Mathematical modeling and fluorescence imaging to study the Ca^{2+}-turnover in skinned muscle fibers. *Biophys. J.*, **74(4)**:1640–1653.

[20] Agard, D., (1984). Optical sectioning microscopy: cellular architecture in three dimensions. *Ann. Rev. Biophys. Bioeng.*, **13**:191–219.

[21] Monck, J., Oberhauser, A., Keating, T., and Fernandez, J., (1992). Thin-section ratiometric $Ca2^{+}$ images obtained by optical sectioning of Fura-2 loaded mast cells. *The Jour. Cell Biology*, **116 (3)**:745–759.

[22] Wilson, T., (1990). *Confocal Microscopy*. London: Academic Press.

[23] Wussling, M. H. P. and Salz, H., (1996). Nonlinear propagation of spherical calcium waves in rat cardiac myocytes. *Biophys. J.*, **70**:1144–1153.

[24] Klein, M., Cheng, H., Santana, L., Y.-H-Jiang, Lederer, W., and Schneider, M., (1996). Two mechanisms of quantized calcium release in skeletal muscle. *Nature*, **379**:455–458.

[25] Lipp, P. and Niggli, E., (1996). Submicroscopic calcium signals as fundamental events of excitation contraction coupling in guinea pig cardiac myocytes. *J. Physiol.*, **492**:31–38.

[26] Tsugorka, A., Rios, E., and Blatter, L., (1995). Imaging elementary events of calcium release in skeletal muscle cells. *Science*, **269**:1723–1726.

[27] Shirokova, N. and Rios, E., (1997). Small event Ca^{2+} release: a probable precursor of Ca^{2+}-sparks in frog skeletal muscle. *J. Physiol.*, **502(1)**:3–11.

[28] Denk, W., Piston, D., and Webb, W., (1995). Two-photon molecular excitation in laser scanning microscopy. In *The handbook of confocal microscopy*, J. Pawley, ed., pp. 445–458. New York: Plenum Press.

[29] Denk, W., Strickler, J., and Webb, W., (1990). Two-photon laser scanning fluorescence microscopy. *Science*, **248**:73–76.

[30] Köster, H. J. and Sakmann, B., (1998). Calcium dynamics in single spines during pre- and postsynaptic activity depend on relative timing of back-propagating action potentials and subthreshold excitatory postsynaptic potentials. *P.N.A.S.*, **95(16)**:9596–9601.

[31] Axelrod, D., (1990). Total internal refelection fluorescence at biological surfaces. In *Noninvasive Techniques in Cell Biology*, J. Foskett and S. Grinstein, eds., pp. 93–127. New York: Wiley-Liss.

[32] Funatsu, T., Harada, Y., Tokunaga, M., Saito, K., and Yanagida, T., (1996). Imaging of single fluorescent molecules and individual ATP turnovers by single myosin molecules in aqueous solution. *Nature*, **374**:555–559.

[33] Veigel, C., Wiegand-Steubing, R., Harim, A., Weber, C., Greulich, K. O., and Fink, R. H. A., (1994). New cell biological applications of the laser microbeam technique: the microdissection and skinning of muscle fibres and the perforation and fusion of sarcolemma vesicles. *European Jour. Cell Biology*, **63**:140–148.

[34] Keating, T. and Cork, R., (1994). Improved spatial resolution in ratio images using computational confocal techniques. In *A Practical Guide to the Study of Calcium in Living Cells*, R. Nuccitelli, ed., Vol. 40, pp. 221–241. San Diego: Academic Press.

[35] Uttenweiler, D., Mann, S., Steubing, R., Veigel, C., Haussecker, H., Jähne, B., and Fink, R., (1998). Actin filament sliding velocity in the motility assay analyzed with the structure tensor method. *Jour. Muscle Res. and Cell Motil. (*Abstract*) in press.*

13 Electron Microscopic Image Acquisition

Heiko Stegmann[1], Roger Wepf[2], and Rasmus R. Schröder[3]

[1]II. Physiologisches Institut, Universität Heidelberg, Germany
[2]Beiersdorf AG, Hamburg, Germany
[3]MPI für medizinische Forschung, Heidelberg, Germany

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

13.1 Introduction

Since its development in the 1930s, electron microscopy (EM) has been used as an imaging technique for the ultrastructural investigation of biological as well as inorganic specimens. Its magnification and spatial resolution capabilities are more than a 1000 times greater than those of light microscopy and are being steadily improved. Electron microscopy has been used for imaging of object areas of several millimeters to individual atoms. Today's advances in electron optics and computer technology have led to completely computer-controlled electron microscopes and fully digitized handling and analysis of the image data produced. Modern electron microscopy utilizes the whole palette of image processing tools, including contrast enhancement, digital filtering, image alignment, object measurement and classification, etc.

This article introduces the three basic concepts of electron-mediated microscopic image acquisition: transmission (TEM); scanning transmission (STEM); and scanning (SEM) electron microscopy.

One uses EM to reveal the 'true' 3-D structure of an object in the microns to Ångstrom range. By parallel illumination of the whole object (TEM) or by scanning it with a focused illumination spot (STEM) and reconstructing the image from that, transmission EM performs a 2-D projection of the object's 3-D spatial information. Scanning EM (SEM) produces topographical information of the object surface by scanning it with an illumination spot. In both cases, the 3-D object information is lost and has to be restored by analysis and interpretation of the image data.

In discussing the nature of the electron specimen interaction, it will become obvious that imaging in the TEM and STEM can be described as a phase and amplitude modulation of the incident electron wave by the transmission function of the object. This function contains the complete information of the specimen. However, imaging in a real transmission microscope is not realized as a simple projection of that function. Electron wave propagation and imaging aberrations introduced by imperfections of the real instrument can be described by introducing a contrast transfer function (CTF) that alternates object information. Correction for this CTF in quantitative image analysis and image reconstruction is indispensable. A brief introduction to the underlying theoretical concepts as well as application examples with regard to corresponding image processing methods will be given.

Image formation in electron microscopic devices is carried out by fast-moving electrons that are deflected in magnetic and electrostatic fields of the constituent atoms of the object and the lenses in the microscope column. During the 1930s, Glaser and Scherzer [1] developed the basic theory of electron optics, and Knoll and Ruska [2] built the first prototypes of TEMs. This resulted in the introduction of the first

commercial TEM by the end of that decade. Invented by von Ardenne [3] in the 1930s, SEM was not developed for routine usage before the 1960s by Oatley [4], and it was commercialized in 1965. For a complete review of historical aspects of electron microscopy see Hawkes [5].

13.2 Electron-specimen interactions

Electrons entering a material interact with its constituent atoms via Coulomb force. Due to this force, some electrons are scattered, resulting in an energy transfer to the atomic nucleus or electrons and a change in their momentum vectors. Scattering events are divided into the two categories *elastic* and *inelastic* and are usually described in terms of scattering cross sections. In electron microscopy, the electron energy is generally measured in electronvolts (eV); 1 eV is the amount of kinetic energy that an electron gains when being accelerated in the electric field that is generated by an electric potential gradient of 1 V.

Elastic scattering comprises electrons that are deflected by the electric field of the atomic nucleus. Electrons that pass close to the center of the atom are scattered through large angles or are even backscattered by the intense field in the immediate vicinity of the nucleus (Rutherford scattering). Most electrons travel far from the nucleus where its electric field is less intense and partially screened by the atomic electrons. They are forward scattered through small angles (typically 10–100 mrad for electrons with an initial energy of $E_0 = 100$ keV). The energy transfer in such small-angle scattering events is limited to fractions of an electronvolt and can therefore be treated as negligible, thus meriting the term 'elastic.'

Inelastic scattering results from the interaction between the incident electrons and the atomic electrons of the target. Excitation of single atomic electrons as well as collective excitations are possible. A fast incident electron may remove an *inner shell* electron (also called *core* electron) of the target atom. This process is called *inner shell ionization*. As a result of the conservation of the total energy, the fast electron loses an amount of energy equal to the binding energy of the removed electron (some ten to several hundred electronvolts) and is scattered through angles of the order of 10 mrad (for $E_0 = 100$ keV). Another atomic electron will fill the vacant core hole. Excess energy is set free as an x-ray photon or is transferred to another atomic electron that may in some cases gain enough kinetic energy to escape from the solid (*Auger emission*).

Interaction of the incident electrons with *single outer-shell* electrons (valence- or conduction-band electrons) leads to smaller energy losses (typically a few electronvolts) and scattering through smaller angles

(1–2 mrad for E_0 = 100 keV). The excited electron may be emitted as a so-called *secondary electron*.

Collective excitations involving the outer-shell electrons of many atoms (*plasmon scattering*) results in energy losses of 5–30 eV for most materials. These events are responsible for the very short *mean free path* of the beam electrons in most materials. Thus for TEM and STEM objects have to be very thin.

In many materials the energy deposited by the beam electrons is only to a small extent converted into x-ray or Auger radiation. Most of it appears as heat. Especially in organic materials, permanent disruption of chemical bonds may also appear. These mechanisms are the main causes of beam-induced damage to the object.

13.3 Transmission electron microscopy (TEM)

The *transmission electron microscope* (TEM) uses parallel illumination of a thin specimen to image it into the final image plane, similar to the well-known light microscope. 'Thin' means that there is negligible absorption of the beam electrons in the specimen so that their larger fraction is transmitted. Specimen thickness has therefore to be in the range of the mean free path for electrons of the respective initial energy (some nanometers to some hundred nanometers).

Electron energies E_0 range from 60 to 140 keV for conventional transmission electron microscopes (CTEM) or from 200 keV to 3 MeV for high-voltage electron microscopes (HVEM). Due to the small mean free path for electrons in air, the microscope column has to be evacuated to pressures around 10^{-6} mbar or better. The electron gun has to deliver a stable, brilliant electron beam, the brightness β of which is given by

$$\beta = \frac{4I_0}{\pi^2 d^2 \theta} \tag{13.1}$$

with the beam current I_0, the beam diameter d and the beam convergence (or divergence) angle θ.

Electrons are emitted thermionically from a tungsten filament (β = 10^5 A cm^{-2}sr^{-1}) or, for higher brightness, from a LaB_6 tip (β = 10^6 A cm^{-2}sr^{-1}); LaB_6 tips, however, require a better vacuum in the gun chamber ($< 10^{-7}$ mbar). For maximum brightness and beam quality, field-emission cathodes ($\beta = 10^8 - 10^9$ A cm^{-2}sr^{-1}) are used that require an even better vacuum ($< 10^{-10}$ mbar) for proper functioning.

The actual image formation is carried out by electromagnetic lenses consisting of wire coils and metal pole pieces. The alternative electrostatic lenses are hardly used today. The currents driving the lens coils are required to be highly stable. A two- or three-lens condenser system

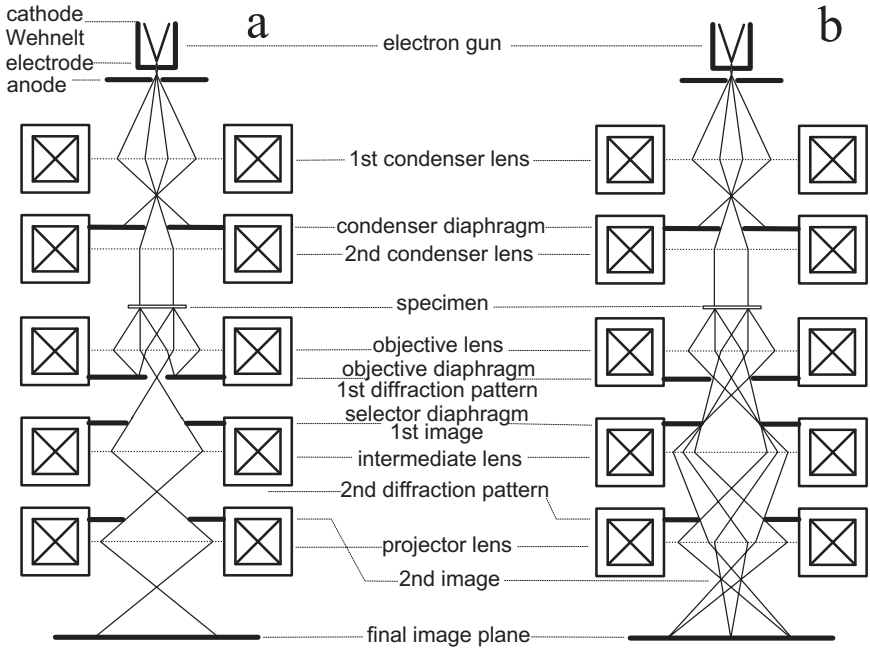


Figure 13.1: *Schematic ray path in an TEM in* ***a*** *bright field imaging mode and* ***b*** *diffraction mode.*

allows the variation of illumination aperture and illuminated area. An objective lens, an intermediate lens and a one- or two-lens projector system image the intensity distribution of the specimen plane into the final image plane (Fig. 13.1).

The electron wavelength λ is given by the relativistic formula

$$\lambda = \frac{h}{\sqrt{2m_0eV\left(1+\frac{eV}{2m_0c^2}\right)}} \tag{13.2}$$

with the Planck constant h, the electron rest mass m_0, the electron charge e, the vacuum speed of light c, and the accelerating voltage V. Although λ is in the picometer range, the large aberrations of the lenses require the use of small objective apertures (5 - 25 mrad) to achieve resolutions in the subnanometer range. Magnifications of some 100,000 times can be achieved routinely. Modern microscopes reach a resolution limit of 0.1 nm for periodic and 0.2–0.3 nm for aperiodic objects.

13.3.1 Ideal TEM

Taking into account the particle-like and the wave-like properties of electrons, the main processes in TEM image contrast formation can be described as elastic scattering and phase shifts introduced to a portion of the beam electrons by the specimen. Contrast formation of a bright-field image can be regarded either as due to absorption of electrons elastically scattered into angles larger than the objective aperture in the particle approach (scattering contrast) or as due to interference between incident and scattered waves in the wave approach (phase contrast).

The theory describing image formation by scattering is referred to as charged-particle (or electron) optics and is a relativistic particle theory. The electron trajectories in an electrostatic field $\boldsymbol{E}$ and a magnetic field $\boldsymbol{B}$ are determined by the Lorentz equation for electrons:

$$\frac{d}{dt}(\gamma m_0 \boldsymbol{v}) = -e(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}) \tag{13.3}$$

with the electron velocity $\boldsymbol{v}$ and the relativistic factor $\gamma = (1-\boldsymbol{v}^2/c^2)^{-1/2}$. Because the electrons travel in close vicinity to the optic axis, only the lowest-order terms in the resulting equations of motion are taken into account (paraxial approximation). Where phase effects are negligible, as in imaging of amorphous specimens in an 'ideal,' that is, aberration-free TEM with perfectly coherent illumination, the formation of scattering contrast can be described in terms of electron optics in the paraxial approximation for low and middle magnifications.

The wave-like aspect of charged particles permits a far more general approach for EM image formation. The incident electron beam is considered as a plane wave that is modified by the object. This modification can be described as a change of amplitude and phase of the electron wave by a 2-D complex transmission function $t_{OP}(\boldsymbol{r}_0)$, where $\boldsymbol{r}_0$ denotes the 2-D projected spatial coordinates (x_0,y_0) of the specimen in the object plane $OP(z = 0)$; $t_{OP}(\boldsymbol{r}_0)$ contains amplitude and phase of the beams that emerge from each point $\boldsymbol{r}_0$, each of which can be seen as a point source for a Huygens spherical elementary wave. The interference between these elementary waves generates a diffraction pattern in the back focal plane (BFP) of the objective. The amplitude distribution $f_{BFP}\,(t_{OP}(\boldsymbol{r}_0), z_{BFP})$ of this pattern corresponds to the Fourier transform of the specimen transmission function. The diffraction pattern itself can again be seen as a source of Huygens spherical elementary waves that interfere to form an enlarged image f_{FIP} of the transmission function in the final image plane (FIP). This amplitude distribution is the inverse Fourier transform of $f_{BFP}\,(t_{OP}(\boldsymbol{r}_0), z_{BFP})$. The image intensity is obtained as the square of the wave amplitude in the final image

plane. In summary, imaging in the ideal microscope reproduces the object by a double Fourier transform times a linear magnification:

$$F_{FIP}(\zeta,\eta) = \frac{1}{M} T_{OP}(\zeta,\eta) \tag{13.4}$$

where F_{FIP} and T_{OP} denote the Fourier transforms of f_{FIP} and t_{OP}, M the magnification, and ζ and η the spatial frequency components in x- and y-direction.

On first sight, this looks like an ideal situation for imaging, providing a simple projection of the phase and amplitude changing potentials of the sample studied. Unfortunately, lens aberrations in a real TEM as well as the necessity for defocus to visualize phase contrast lead to contrast transfer modulations that change the sample information recorded in one single image.

13.3.2 Real TEM

For real magnetic lenses, the paraxial approximation of the particle optics model is no longer valid and higher-order terms in the equations of motion for the electrons have to be considered. These terms account for geometrical aberrations in a real microscope; chromatic aberrations can be introduced via consideration of small changes in electron energy and lens strength.

In the wave-optical approach, a thorough analysis of the situation leads to the concept of the contrast transfer function (CTF) [6, 7], the electron optical analogue to the optical transfer functions of light optical devices (Chapter 20). Independent of specimen structure, the CTF describes the imaging properties of the microscope.

The double Fourier transform times linear magnification as mentioned in Eq. (13.4) has to be multiplied by the CTF that contains the effects of limited illumination and objective apertures, lens aberrations and the energy spread of the electrons:

$$F_{FIP}(\zeta,\eta) = \frac{1}{M} CTF(\zeta,\eta) T_{OP}(\zeta,\eta) \tag{13.5}$$

Thus, image formation in a TEM can be regarded as a filtering operation, the filter function CTF not being equal to unity in a real microscope. Different spatial frequencies will be transferred with different weights (Fig. 13.2).

In the following, the most important imperfections of real electron optical systems and their contributions to phase shifting and resolution limiting will be listed.

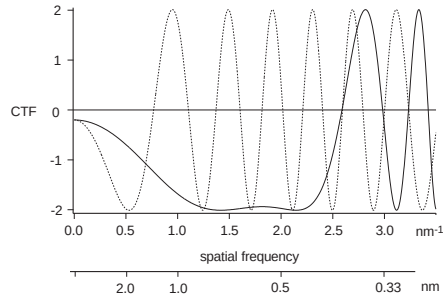


Figure 13.2: *Electron optical contrast transfer function (CTF) for a weak phase-weak amplitude object (10 % amplitude contrast). Assumed are 120-keV electrons and a spherical aberration of the objective lens of C_s =2.7 mm. Shown are the CTFs for two underfocus values: solid line for Δz =100 nm (Scherzer focus with broad transfer band) and dashed line for Δz =500 nm.*

Spherical aberration. Spherical aberration can be described as an additional phase shift to $f_{BFP}\,(t_{OP}(r_0), z_{BFP})$ depending on the spherical aberration coefficient C_S (an instrumental parameter characteristic for every microscope design, typically C_S =0.5-3 mm), the defocus Δz and the scattering angle θ. In bright field mode, the maximum positive phase contrast is reached at the Scherzer defocus $\Delta z = (C_s\lambda)^{1/2}$.

The radius ρ_s of the confusion disk in the object plane caused by spherical aberration for a ray at the angle θ to the optical axis is given by

$$\rho_s = C_s\theta^3 \tag{13.6}$$

Spherical aberration causes a phase shift $\Delta\varphi_A$ for electrons beams not parallel to the optical axis:

$$\Delta\varphi_A = \frac{\pi}{2} C_s\theta^4/\lambda \tag{13.7}$$

Diffraction (Abbe) limit. Magnetic lenses have a large spherical aberration that has to be reduced by the use of small objective diaphragms. This limits the achievable resolution with a given objective aperture θ_A, the Abbe limit well known from light microscopy. An object point is imaged as a disk of confusion with the radius

$$\rho_A = 0.61\lambda/\theta_A \tag{13.8}$$

Thus, two points in the object have to be separated at least by this distance to be observable as two distinct points in the image.

Defocus. For weak amplitude-weak phase objects, phase contrast vanishes in the exact focus to leave only a faint scattering or amplitude contrast. Image contrast increases by phase contrast in underfocus ($\Delta z > 0$). Dark structures become darker in their center, thus providing maximum contrast. In overfocus ($\Delta z < 0$), phase contrast reverses and dark structures appear bright in the center with a dark rim. However, shifting of the object plane by defocusing leads to a disk of confusion of

$$\rho_D = \Delta z \theta \tag{13.9}$$

and to a phase shift of

$$\Delta\varphi_D = \pi \frac{\Delta z}{\lambda} \theta^2 \tag{13.10}$$

Chromatic aberration. Due to instabilities in high voltage E and lens current I as well as inelastic scattering in the object, the focus distance f of the objective lens is smeared out over a range Δf. This leads to a disk of confusion in the object plane given by

$$\rho_C = \Delta f \theta \tag{13.11}$$

with

$$\Delta f = C_C \sqrt{\left(\frac{\Delta E}{E}\right)^2 + 4\left(\frac{\Delta I}{I}\right)^2} \tag{13.12}$$

where C_C denotes the chromatic aberration constant of the microscope, ΔE the energy spread due to high-voltage instability and inelastic scattering, and ΔI the lens current fluctuations.

Resolution limit. Further imperfections include 'parasitic' aberrations such as radial and axial astigmatism, which can be minimized by electron-optical correction elements (e. g., quadrupole lenses), and the effect of incident beam divergence. These can be neglected compared to the major deviations from ideal imaging conditions listed in the foregoing. The radius of the resulting disk of confusion is given by

$$\rho = \sqrt{\rho_S^2 + \rho_A^2 + \rho_D^2 + \rho_C^2} \tag{13.13}$$

with ρ_S, ρ_A, ρ_D, and ρ_c as already defined. For the quantitative analysis of images and their reconstruction to form a 3-D visualization of the

original object, it is absolutely necessary to correct for the CTF either by combining information from defocus series (Section 13.8) or—most recently—by improving the optical properties of the lens, for example, by correction of the spherical aberration [8]. The latter correction leads to an almost perfect transfer of the amplitude contrast. In the case that samples provide strong amplitude contrast (as, e. g., for metals or typical samples in solid state physics), this leads to almost perfect images that are close to the ideal projection of specimen potentials, and thus are being easily interpreted.

13.3.3 Imaging modes in the TEM

In principle, two modes of imaging can be distinguished for the conventional, nonanalytical TEM: first, the imaging of the specimen itself to visualize directly the spatial distribution of an object; and second, the imaging of electron diffraction patterns of a given sample, thus recording the amplitudes of Fourier structure factors of the spatial distribution only. These two modes are realized in the TEM by changing the excitation current of the first projective lens that either magnifies the first intermediate image formed by the objective lens or the first diffraction pattern in the back focal plane of the objective lens (Fig. 13.1).

On first sight, electron diffraction seems not to be very useful compared to direct imaging, as all the phase information of the structure factors is lost. However, it is relatively easy to obtain high-resolution diffraction patterns of crystalline samples that give unique information about the 3-D structure of the crystal, lattice excitations, or multiple elastic-inelastic scattering (Kikuchi bands in the case of dynamic scattering on thick samples). Also for biological samples it is advantageous to study crystalline arrays. Such specimens are extremely susceptible for beam damage, thus electron diffraction is the method of choice to obtain high-resolution structure factor amplitudes because it has a much better signal-to-noise ratio than comparable imaging (Section 13.9).

For the normal imaging of an object TEM electron optics allows a large variety of different techniques to obtain various aspects of the specimen information contained in the transmitted electron beam.

Bright-field mode (BF). In conventional imaging in the BF mode, a centered objective diaphragm (the corresponding objective apertures are 5-20 mrad) that rejects electrons scattered into large angles leads to scattering contrast. Therefore, the amount of transmitted and collected electrons depends on the objective aperture, the electron energy and the mean atomic number and mass thickness (the product of density and thickness) of the specimen. In a thin specimen, it decreases

exponentially with increasing mass thickness. This fact can be used for the measurement of mass thickness.

For very beam-sensitive specimens, modern electron microscopes are equipped with a minimal dose focusing (MDF) aid. For focusing, it only illuminates an object region adjacent to the one to be actually recorded and then switches the illumination to the desired region for image acquisition.

Dark-field mode (DF). The electrons scattered into small angles are rejected, and the image is formed by the electrons scattered into large angles thus producing a reverse contrast (negative image). This can be achieved by shifting the objective diaphragm out of center, by tilting the incident beam or by introducing an objective diaphragm with a central beam stop. This mode yields higher contrast when imaging structures with very low mass thickness, but also needs higher electron doses than the bright-field mode.

Spectrum mode in the EFTEM, energy filtered imaging. Provided the TEM is equipped with an energy filter that separates the transmitted electrons according to their kinetic energy, the electron energy loss spectrum of a selected area of the specimen can be imaged and recorded. In the case of a corrected energy filter that can be used for imaging, setting a slit aperture in the energy-dispersive plane allows formation of images by electrons that lie in a defined energy interval only (Section 13.5.2). If, for example, the slit is centered at zero energy loss, no inelastic electrons contribute to the image, thus providing increased bright-field contrast and resolution and increasing the useful specimen thickness.

Electron holography. A successful means to overcome the resolution limit imposed by spherical aberration is electron holography. Its principle is identical to that of light holography that was accomplished with the invention of the laser as a highly coherent light source. The electron beam coming from a highly coherent field emission gun is split into two half-beams, one of which images the specimen while the other one serves as a reference beam.

In the final image plane, the image wave is brought to interference with the reference beam by means of an electrostatic biprism and the resulting pattern is recorded on a film. The reconstruction of the image from the hologram is either carried out with a laser as a reference wave illuminating the film or digitally in a computer. Simpler forms of electron holography use the unscattered part of the electron beam as a reference wave and can do without additional prisms.

13.3.4 TEM image detectors

Besides the classical methods of viewing and storing an electron micrograph (fluorescent screen and photographic film), today there are a number of alternatives such as intensified TV cameras and cooled CCD-arrays (charge coupled devices) with scintillator screens for on-line acquisition and processing of the TEM image. Because these electronic image detectors are more sensitive than the combination fluorescent screen-human eye, they are indispensable for specimen examination and focusing under low-dose conditions.

Fluorescent screens. The classical viewing device in the TEM is a fluorescent screen that converts the impinging electron's kinetic energy into visible light. In most cases it is made of ZnS and CdS powder, usually green fluorescent, but occasionally with the addition of other metals to change its color. The fluorescent screen is essential for adjustment of the microscope, selecting the desired specimen area and focusing. The light intensity L of a fluorescent screen is proportional to the incident electron current density and has also a weak dependence on the electron energy. The spatial resolution δ lies in the range of 30 to 50 μm. The quality of fluorescent layers can be compared by means of the ratio L/δ^2.

Photographic emulsions. Photographic emulsions containing silver halide particles are not only sensitive to photons but also to electrons. Thus they can be directly exposed to the electrons in the TEM. A short exposure to a high electron density results in the same photographic density of the developed emulsion as a long exposure to a low electron density. However, before being used in the EM the photographic material has to be dried in a desiccator to remove the water content of the emulsion that would otherwise deteriorate the microscope column vacuum.

In addition to the size of the silver halide grains, the resolution of a photographic emulsion depends on the diameter of the electron diffusion halo. This electron cloud is produced when the impinging electrons are scattered at the grains. Unlike for light exposure, the diameter of the halo is independent of the grain size and subject only to the mean emulsion density and electron energy. Typically, the resolution is limited to about 10-20 μm.

Thus, for the common film size of 6×9 cm^2, a storage capacity of 2.4×10^7 image points can be expected. The dynamic range is restricted to less than 10^4.

Image plates. Image plates consist of a plastic sheet coated with a phosphorescent storage layer. Such storage layers can be made, for example, from small grains of BaFBr:Eu^{2+}. An incident electron will

generate multiple electron-electron hole pairs in such an active layer that are trapped in F-centers in the crystalline structure of BaFBr:Eu^{2+}. Therefore, electrons are recorded in the form of electron-electron hole pairs that can later be activated to recombine by red laser light. The energy stored in the F-center is then converted into blue luminescent light. The detection system itself consists of the image plate and a read-out device that scans red light over the plate, simultaneously detecting the blue luminescence signal.

Comparing image plates with fluorescent screens, photographic emulsions and CCDs, they show a very high quantum detection efficiency, a medium spatial resolution, and the disadvantage of not being an online detection medium. Further advantages of image plates are their ease of use, their large detection size (both comparable to photographic negatives), and their unsurpassed dynamic range of up to 10^6. Except for their non-online handling, they are ideal detectors especially for electron diffraction patterns.

TV cameras. For recording of electron micrographs, TV camera tubes have to be equipped with a fluorescent screen coupled to the tube with a fiber-optic plate. Silicon intensifier target (SIT) tubes or the combination of a regular TV tube with an image-intensifier tube allow detection of single electrons. Frame-grabbing cards permit an easy digitalization of the image. TV tubes are advantageous for real-time observation of the TEM image. However, they have a very low dynamic range ($< 10^3$).

Semiconductor detectors. Modern CCD chips are made with up to 2048×2048 pixels and pixel sizes of about 19-24 μm^2. Since direct exposition to the electron beam results in long-term damage to the chip, scintillator screens made of plastic, yttrium-aluminum garnet (YAG) crystals or phosphor powder are used to stop the electrons and to convert their energy into photons, thereby somewhat deteriorating the resolution due to the lateral point spread of the scintillator screen. Dynamic ranges are typically $< 10^5$. Cooling of the chip provides a low noise level. In a computer-controlled microscope equipped with digital image acquisition and on-line processing capabilities, the image detector may be used for automated microscope adjustment, comprising beam alignment, correction of astigmatism and focusing.

13.4 Scanning transmission electron microscopy (STEM)

In the *scanning transmission electron microscope* (STEM), the image is formed by scanning a thin specimen with a highly convergent beam focused to the smallest possible spot and detecting the transmitted electrons downstream by various electron detectors. The objective lens is used to demagnify the electron source crossover formed by the con-

denser lenses into the object plane. Below the specimen, no further electron lenses are needed. The beam is deflected in x- and y-direction by scanning coils that are driven by saw-tooth currents. Simultaneously, these currents deflect the electron beam of a cathode-ray tube. To visualize the image, the tube's electron beam is modulated by one or a mixture of the electron detector signals. The detector signals can as well be processed digitally for pixelwise acquisition of the image data. Because brightness, probe size and scanning speed depend on each other, a high-resolution STEM with probe diameters of only some ångstroms is only possible using high brightness field emission guns. Some TEMs are equipped with additional scanning attachments that allow it to be run in the STEM mode, however, the performance power usually does not reach that of a dedicated STEM.

Because in thin specimens virtually no electrons are absorbed, the STEM can make use of all beam electrons in their different portions after they have passed through the specimen: unscattered, elastically and inelastically scattered. As elastically scattered electrons are on average deflected through larger angles than inelastically scattered ones, the former can be separated by an anullar detector that may consist of a metal plate with a center hole. Unscattered and inelastically scattered electrons will pass through that hole and can be detected by a central detector or be separated according to their energy by an energy dispersive filter (Section 13.5.2). All electrons can be detected so that choosing between various portions of unscattered and large-angle-scattered electrons to contribute to the image allows simultaneous imaging modes such as bright-field and dark field imaging. The ratio $I_{\text{elastic}}/I_{\text{inelastic}}$ delivers an increased contrast for atoms of different atomic number (*Z-contrast imaging*). By additional special detectors, backscattered and secondary electrons can be detected to image the surface structure of the specimen as in dedicated SEMs.

The beam path in a STEM is reciprocal to that in a TEM. Image formation in the STEM can therefore be described in analogy to the theory developed for the TEM. This fact is known as the theorem of reciprocity. The central detector of the STEM corresponds to the TEM electron gun, while the large detection area of the TEM is equivalent to the STEM source if the scanning probe is traced back (Fig. 13.3).

Besides the forementioned metal plate detector, photomultipliers (PMT) with scintillator screens or semiconductor detectors are also used as STEM electron detectors. Semiconductor detectors are easy in operation, but offer less gain and smaller bandwidth than scintillator/PMT combinations. Higher gains can be achieved by microchannel plates.

Generally speaking, STEM imaging allows better control of the applied radiation dose for beam-sensitive specimens. Faraday cups in combination with an electrometer are used for direct quantitative measurements of electron currents. For recording of energy loss spectra,

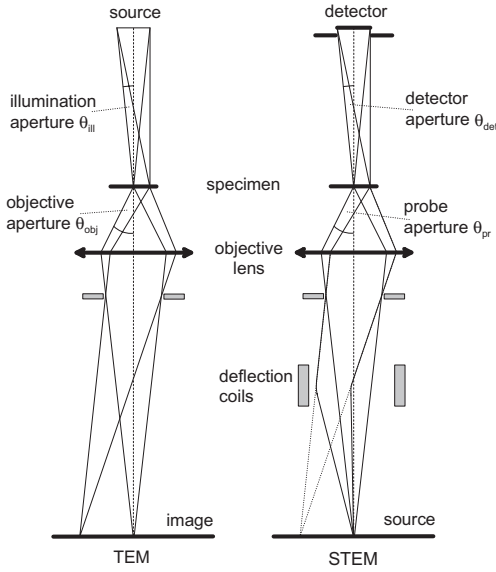


Figure 13.3: *Schematic ray path in a TEM and a STEM. Redrawn from [9].*

linear photodiode or CCD arrays (parallel detection) or photomultipliers over which the spectrum is shifted (serial detection) are employed.

The electron current distribution in the detector plane is a far-field diffraction pattern of the illuminated object area. Therefore, recording that pattern with an array of small detectors rather than one large detector allows a wide range of structural and phase information to be extracted from each object point by digital data processing.

Despite being superior to TEMs in most respects, dedicated STEMs are today only very rarely used for some special applications due to their complicated use and maintenance.

13.5 Analytical transmission electron microscopy

As a general trend, transmission electron microscopy moves more and more away from simple imaging towards analytical methods, above all element analysis. *Electron energy loss spectroscopy* (EELS) in the *energy filtering transmission electron microscope* (EFTEM) and *electron-probe microanalysis* (EPMA) in the STEM provide a sensitive means of obtaining image data specific to chemical elements or chemical phases present in the sample, thereby introducing another 'element dimension' to the two-dimensional image.

13.5.1 Electron probe x-ray microanalysis

During the past decades, electron-probe x-ray microanalysis (EPMA, also called EDX (energy dispersive x-ray microanalysis) has become a standard technique for the qualitative and quantitative analysis of the element composition of very small samples. Here, the characteristic and continuum x-ray quanta generated in the sample by the electron beam are counted by a semiconductor detector and the resulting x-ray spectra processed quantitatively to obtain absolute concentrations or concentration ratios of the elements present in the specimen [10, 11]. When carried out in a scanning transmission electron microscope (STEM), a high spatial resolution of a few nanometers and a sensitivity of 10^{-15} g of an element can be reached, limited by the counting statistics, the diameter of the electron probe and the thickness of the object to be analyzed. Digitally storing the x-ray spectrum from every pixel in a scanning image allows computation of compositional maps.

However, EDX imposes some severe disadvantages that limit its usefulness especially for biological specimens: in biological samples one is generally interested in the detection of light elements with an atomic number up to 20 such as O, Na, Mg, P, S, Cl, K and Ca. Since the x-ray quantum yield decreases with falling atomic number in favor of non-radiative processes and since only a small portion of the x-ray quanta can be detected due to the limited size of the detector entrance window (covering only about 10% of the solid angle into which x-ray radiation is emitted), there is a need for long acquisition times and high beam intensities. This also results in high electron doses that produce radiation damage of the delicate specimen, thus causing drift problems, loss of the elements to be measured, and overall mass loss.

13.5.2 Energy-filtering electron microscopy

Facing these disadvantages, it seems natural to look for the primary process of the electron beam-target interaction, that is, the elastic and inelastic electron scattering that results in a respective energy loss of the beam electrons, rather than to observe the ineffective secondary process, that is, the generation of x-rays. By collecting and analyzing the electrons transmitted through the sample spectroscopically up to 80% of the inelastic collision events can be detected. This method—known as electron energy loss spectroscopy (EELS)—results in smaller electron doses needed to obtain the same amount of information as with EDX methods. Therefore, shorter acquisition times, less beam damage to the specimen or higher sensitivity especially for the light elements can be achieved [12].

After the incident beam electrons with an initial kinetic energy E_0 have passed the sample, they are separated according to their kinetic

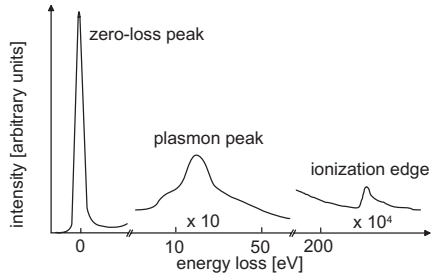


Figure 13.4: *Schematic electron energy loss spectrum.*

energy E by means of an energy dispersive electron spectrometer to produce what is called an *electron energy loss spectrum*, showing the scattered intensity as a function of the energy loss $\Delta E = E_0 - E$ of the beam electrons.

A schematic energy loss spectrum is showed in Fig. 13.4. The first *zero-loss* or elastic peak at 0 eV represents those electrons that are elastically scattered into small angles.

The second *plasmon-loss* or *low-loss* peak in the region 5-50 eV represents the inelastic scattering from outer-shell electrons and shows discrete energy losses in multiples of the plasmon energy of the respective material. Towards higher energy losses, the plasmon peak decreases smoothly according to a power of the energy loss.

Core-loss or *ionization* edges are superimposed on the plasmon loss background at higher energy losses: a sharp rise in the scattered intensity occurs on the *ionization threshold* of an inner-shell excitation—the energy loss that approximately equals the binding energy of the corresponding atomic shell—and decreases in a long tail. Since the binding energies depend on the atomic number of the atom, the resulting edges in the energy loss spectrum are characteristic for the elements present in the specimen. Measuring the area under these edges allows for quantitative element analysis. The chemical binding type of the specimen atoms results in a fine structure of the edges that can give information on the chemical environment.

Similar to EDX, EELS can be carried out using a highly focused electron probe, for example, in the STEM, which allows the acquisition of the electron energy loss spectrum from the illuminated area by a serial detector such as a photomultiplier (*serial EELS*) or a parallel detector similar to a photodiode array, a CCD or TV camera (*parallel EELS*). If the electron probe is directed to scan over the specimen, the acquired two-dimensional set of spectra can be processed to obtain an element mapping of the scanned area (*spectrum-imaging*). Using an imaging energy filter and parallel illumination, EELS offers the possibility of recording images to which only electrons of a defined energy loss con-

tribute, so-called energy-selective or energy-filtered images. Taking 2, 3 or more energy-filtered images at different energy losses around element specific features of the spectrum (*electron spectroscopic imaging* (ESI) allows qualitative and semiquantitative element-specific imaging. If whole series of energy-selective images at constant energy intervals are recorded, spectra can be extracted from these series by means of digital image processing (*image EELS*). For that purpose, *regions of interest* (ROI) are outlined in one of the images or in a separately recorded high-contrast image. The pixel gray values from within a ROI have then to be averaged and these average values arranged along the energy-loss axis according to the energy intervals used during recording to end up in a spectrum from that ROI.

With the advance in electron optics, electron detector and computer technology during the last ten years, the routine use of EFTEM techniques in EM image acquisition has become more and more feasible. However, the large and often highly varying background underlying the characteristic energy-loss signals demands considerable knowledge and experience from the operator and takes far more effort for data processing than in EDX microanalysis.

13.6 Scanning electron microscopy (SEM)

An image formation mechanism completely different from transmission microscopy is used in *scanning electron microscopy* (SEM). Mainly the surface topology of thick objects is imaged. While probe forming and scanning is carried out in the same way as in the STEM (Section 13.4), there are no transmitted electrons that could be detected from bulk specimens. The object information is contained in the various shares and emission directions of secondary and backscattered electrons. Modern SEMs achieve a resolution in the range of up to 1 to 3 nm. SEM offers a high depth of field and delivers topographical, magnetic, chemical and electronic state information of the sample.

13.6.1 Signal generation

The signal is generated at the specimen surface or within the specimen by scanning it with a fine electron probe. The signal consists of *backscattered electrons* (BSE), *secondary electrons* (SE, Section 13.2) and *Auger electrons* [13, 14]. Compared to SE, Auger electrons are emitted from the specimen in such low numbers that, because of the low signal-to-noise (S/N) ratio, they can be neglected in a conventional SEM; BSE and SE are generated in sufficient numbers if high brightness electron sources—LaB_6 or field emitter—are used. Due to the low energy of SE

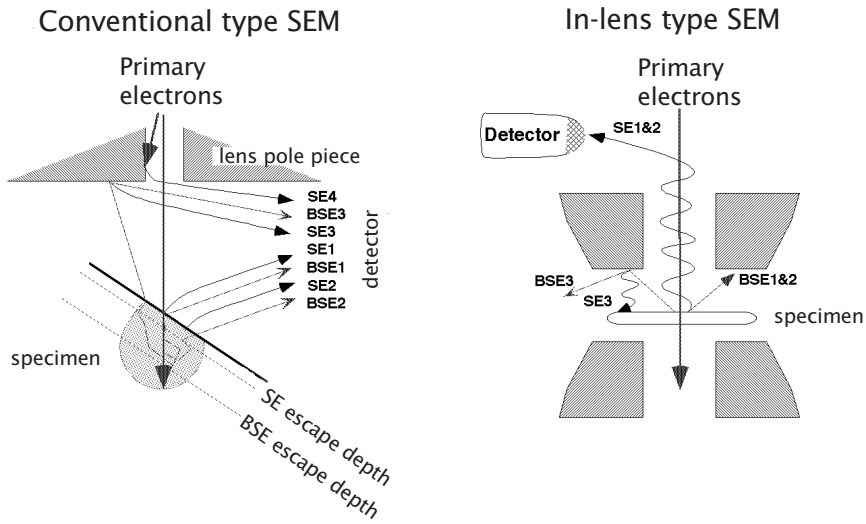


Figure 13.5: *Signal detection in the SEM. Schematic detection of SE and BSE in a conventional below-the-lens type of SEM (redrawn from Reimer and Pfefferkorn [14]) compared to a high resolution "in-lens" type SEM.*

compared to BSE, they can be separated by applying a small positive or negative bias to the detector front end.

Another prerequisite for high-resolution SEM is a probe size of about 1 nm in order to deliver a highly localized signal. Analogous to the minimum disk of confusion in the TEM, the minimum achievable spot size is given by Eq. (13.13). A probe size of less than 1 nm as obtained in STEM can be produced in an "in-lens" field emission SEM (FESEM) [15]. The different arrangements of the specimen and the detectors in an "in-lens" FESEM compared to a conventional (FE)SEM affect the signal registration and contrast in the "in-lens" type FESEM (Fig. 13.5).

Electrons emitted from the sample with an energy below the arbitrary value of 50 eV are classically called *secondary electrons* (SE) with a further classification into SE produced by the primary beam (SE I) and SE produced by BSE (SE II). The volume from which SE I emerge is given by their escape depth and the probe size. This volume being very small, SE I carry the high spatial resolution information; SE II may be emitted from an area of several microns squared (cross section of the SE escape depth with the BSE escape volume), thus reducing the SE I signal contrast; SE generated by BSE at the lower part of the column (pole piece) and at other parts of the specimen chamber (*SE-III*), and SE generated at the final aperture (*SE-IV*) contribute disturbingly to the SE-I and SE-II signal emitted from the sample. The SE-III, which account for 60-70 % of the SE signal in a conventional type SEM [16] do not contribute to the

collected signal in an in-lens type SEM, because the SE detector is positioned behind the condenser/objective lens and the specimen between the two pole pieces of the immersion lens [17]. In other words, the S/N ratio and the contrast in an 'in-lens' type SEM depend therefore mainly on the SE-I and SE-II signal and, besides the different geometrical arrangements, may be quite different compared to the contrast obtained in a conventional SEM.

Secondary electron imaging. The high resolving SE-I cannot be separated from the SE-II. However, to obtain high-resolution images (< 3 nm) with the SE-imaging mode, a way has to be found to enhance the signal generated at the spot of incidence against the SE-II signal. Especially on biological specimens, the SE-II signal is even enlarged due to a larger electron interaction volume in low atomic number specimens.

The insufficient SE emission and its unsatisfactory S/N ratio from specimens with low atomic numbers (e. g., biological samples) and induced charging effects during irradiation made metal coating a powerful tool for SE imaging of biological samples at low and high magnification [18, 19]. One way to increase the high resolution topographic signal on biological samples is to coat the surface with a thin metal film (1–2 nm). Such a thin metal film localizes the signal at the surface of the specimen and reduces charging effects. A way to separate the high resolution signal (SE-I) from the diffuse SE-II signal is to use a coating film that has a low BSE coefficient, hence no or very few BSE are produced. Only the SE-II produced by the BSE from the interaction volume in the biological sample below such a coating film will contribute to the SE-image. Various light metals fulfill this condition, and thin chromium films have proven to reveal a high SE-I signal and therefore high resolution in SEM. Higher resolution can be achieved with the help of fine-grained tungsten (W) films [20] due to a clear distinction and hence localization of the fine metal grain (1–3 nm) against the vacuum or biological material.

Backscattered electron imaging. BSE have energies of 50 eV to the full primary beam energy with the major fraction around the primary energy and emerge from an area of up to a few microns squared. They therefore also contain information from beneath the surface. High-resolution BSE (BSE-I), generated in the specimen with low energy loss in the area of probe incidence differ from other BSE (BSE-II) that emerge after multiple scattering and high energy loss at some distance from the probe. With the BSE-I signal collected from highly tilted bulk-metal or gold-coated specimens, high resolution (2 nm) in SEM with short focal lenses has been demonstrated in the low-loss imaging mode [21]. Walther and Hentschel [22] demonstrated that even from an untilted bulk biological specimen shadowed with 2 nm platinum-carbon and sta-

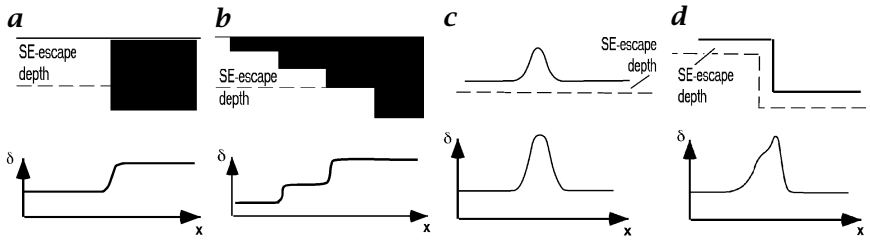


Figure 13.6: *High-resolution SE-contrast from thin coating films. Types of surface contrast produced by specimens coated with a thin film.* ***a*** *Nontopographic contrast types are the atomic number contrast and* ***b*** *the mass-thickness contrast.* ***c*** *Topographic contrast types are the particle contrast* ***d*** *and the edge brightness contrast.*

bilized with 10 nm carbon, sufficient BSE are generated to form an image with high resolution.

13.6.2 Contrast mechanisms

The basic contrast-forming mechanism in SEM is topographic contrast: the larger the angle between incident beam and surface normal, the larger the number of secondary electrons that lie within their escape depth and thus contribute to the SE signal. The resulting SE image resembles a photograph of the object taken from the direction of the electron beam with the object illuminated from the direction of the detector.

Contrast in the SEM of thin coated specimens (<2 nm) is obtained because the SE and BSE signals vary with the film thickness of thin coating films parallel to the incident beam. The different SE contrast mechanisms in high-resolution SEM of thin coated specimens are shown in Fig. 13.6. The so-called mass thickness contrast (Fig. 13.6b) allows imaging structures of a few nanometers in size [19, 23]. Different mass thicknesses can also be seen by the electron probe on a coated slope, revealing signal variations according to the steepness of the slope. Another contrast mechanism is the different SE- and BSE-yield with varying atomic number Z (material or atomic number contrast, Fig. 13.6a), which is stronger for BSE than for SE [24]. Special contrast effects appear if the electron probe is close or within the SE-escape depth at a border (Fig. 13.6d). The contrast of such a border is enhanced due to the emission of a higher portion of SE that still can leave the close surface. This signal can exceed the maximum signal of a film with a thickness corresponding to the SE-escape depth. The same effect also enhances the contrast of small particles (particle contrast, Fig. 13.6c), if their diameter is smaller than twice the SE-escape depth.

The magnetic state of the specimen surface as well as electric surface potentials alter SE- and BSE-yield and allow magnetic contrast and voltage contrast imaging. Energy-dispersive x-ray (EDX) detectors are routinely used in the SEM to detect the x-rays that are generated by the incident electrons inside the specimen. Counting the characteristic x-ray quanta allows mapping and quantification of the chemical elements present in the specimen.

Recent improvements in electron optics made it possible to work with low acceleration voltages of 200 eV to 5 kV (LVSEM, low-voltage scanning electron microscopy) without decreasing resolution, thus allowing control of the penetration depth of the incident electrons and therefore varying the depth of the imaged surface layer even for uncoated specimens. Another development, the *environmental scanning electron microscope* (ESEM), allows microscopy of fully hydrated biological specimens under normal conditions, imaging of water layers, dynamic events and chemical reactions. Reconstruction of the 3-D object topography is possible by taking images under different directions of illumination and digitally recovering the 3-D information.

13.7 Preparation techniques

Specimen preparation is the most crucial step in electron microscopic image acquisition. Poor preparation will lead to restricted image information or even to the constitution of artifacts. As, with a few exceptions, electron microscopes have to be operated at high vacuum, it is not possible to use living biological specimens as in a light microscope. Living specimens have to be adequately processed by fixation, dehydration and coating or embedding in plastic resins, which also raises their ability to withstand beam-induced structural damage [25]. The need for very thin, electron-transparent specimens requires these embedded preparations to be cut into ultrathin sections by ultramicrotomy using glass or diamond knives. These sections have thicknesses of typically 50 to 150 nm. In HVEM, sections a few hundred nanometers thick can be studied. Hard inorganic solids such as ceramics or metals may be crushed or first cut into slices with diamond saws and then thinned by mechanical polishing, electropolishing or ion milling. Ultramicrotome sections and small crunching chips must be supported by small metal grids, if necessary with a supporting carbon or polymer film, to be introduced into the microscope column.

Staining with heavy metals (“negative staining”) allows fine structures, supramolecular assemblies and single macromolecules to be visualized at a high resolution. Specific macromolecules can be highlighted by various labeling techniques. High-resolution imaging of specimen surfaces in the TEM can be achieved by the use of replica tech-

niques, where thin metal films are evaporated onto the sample to be examined in the microscope after removing the original specimen.

Conventional chemical fixation techniques for EM specimens very often lead to preparation artifacts that can be avoided using cryofixation techniques that are of increasing importance especially in biological applications. Here, rapid freezing of the specimen is used to achieve an optimum conservation of the native ultrastructure. Freeze-fracture and freeze-etching, two other freeze-preparation techniques, use cleaving of the frozen specimen to reveal the interior of cells and cell membranes.

13.8 Digital image processing of electron micrographs

The *electron micrograph* can be made available as a matrix of pixel gray values either from direct digital image acquisition with a CCD camera or adequate transformation from a photographic film or TV image in the TEM, or by direct digital acquisition in a scanning microscope [25, 26]. It is obvious that all conventional image processing techniques can then be applied to such image data. The whole palette of digital image enhancement tools can be used for visual improvement of the images acquired: adjusting brightness and contrast, gamma correction, sharpening, deblurring, or removing background structures (Chapters 5 and 6). It becomes especially necessary to use these methods to preprocess very low signal-to-noise ratio images for image alignment and extensive image averaging.

Electron micrographs recorded with very low electron doses, for example, tend to be very noisy. If the image contains a sufficient number of identical structures such as cellular organelles or macromolecules, it is possible to average over those to obtain a noise-reduced image (Chapter 7). This *averaging of low-dose images* requires motif-detection procedures to select similar structures (Chapter 10) and cross-correlation algorithms to position, orient and align them. *Image alignment* can be achieved by computing the necessary shift vectors (Chapter 9) from the 2-D cross-correlation between two subsequent images. Proper alignment is particularly essential to the processing of electron micrograph series. For the evaluation of periodic structures in the object, *fast Fourier transform* (FFT) (Chapter 3) can be used, because 2-D FFT of the image provides the diffraction pattern of the imaged specimen area. It can also be used to control aberrations, defocus, astigmatism and other lens defects. The image may also be filtered in Fourier space and then inversely transformed to perform digital filtering without loss of phase information.

As in light optical microscopy it is also necessary to correct the image information for the point spread function of the microscope.

As mentioned in the foregoing, in electron microscopy the effect of the point spread function is usually modeled by the *contrast transfer function* (CTF) (Chapter 20).

To recover the complete object information from the given image data and to end up with an ideal 2-D projection of the specimen potentials the microscope contrast transfer function (CTF) is determined and corrected. This is equivalent to *retrieving lost phase information*. One is especially interested in overcoming the spatial frequency transfer gaps introduced by the CTF. This means separation of amplitude and phase distribution in the image plane and reconstruction of the complete image information from that. As mentioned earlier, the image and diffraction pattern are related by Fourier transformation. If both the image and diffraction pattern of a periodic specimen are recorded, the unknown phase information can also be restored from those two by an iterative procedure (*Gerchberg-Saxton algorithm*). Since the phase shifts to the electron waves depend on defocus, a series of two or more micrographs at different defocus values can be recorded (*defocus series*, Chapter 20) and the missing phase information be calculated from that.

Finally, electron microscopy needs special reconstruction algorithms for *retrieving the 3-D information*: regular TEM imaging being a two-dimensional projection of a certainly thin but nevertheless 3-D object, one needs to acquire at least two of these projections taken under different tilt angles of the specimen to recover 3-D spatial information. Two images under two tilt angles are required for stereographic image pairs (Chapter 17). Series of more than two micrographs covering a larger angle range (*tilt series*) are needed for a *3-D reconstruction* of the object structure. A tilt series is equivalent to a tomographic series of central sections in Fourier space. The 3-D structure in real space can therefore be calculated by inverse Fourier transformations. In practice, however, beam damage to the specimen often limits the number of micrographs that can be recorded in a tilt series, so that low-dose conditions have to be applied.

13.9 Imaging examples

Today's applications of electron microscopy can be classified as imaging of specimens in its original optical meaning and as analytical imaging. In both cases, electron scattering potentials of the specimen are imaged that are converted into structural information about the sample. The obtainable spatial and analytical resolution is largely dependent on the preparation of the sample and its sensitivity to beam damage. Only at very high resolution, that is, for imaging of single atoms in material sciences, the instruments become the limiting factor. Typical structural resolutions obtained routinely today are 1-2 nm for biologi-

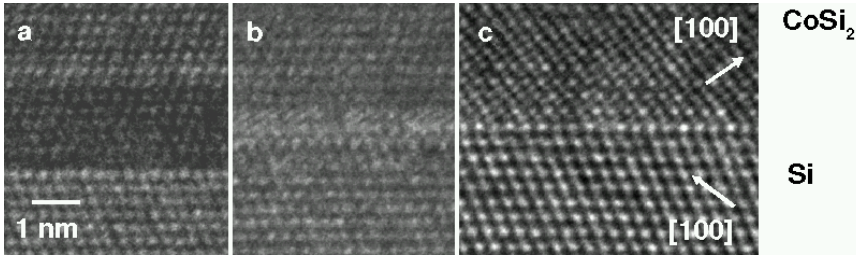


Figure 13.7: *Image of an epitaxial Si(111)/CoSi interface illustrating the contrast delocalization as image artifact due to spherical aberration. Images* ***a*** *and* ***b*** *are taken without C_s correction at different defocus values:* ***a*** *Scherzer focus;* ***b*** *Lichte focus, C_s = 1.2 mm). Image* ***c*** *shows the identical sample area in the corrected TEM at Scherzer defocus with a remaining C_s = 0.05 mm. Figure from Haider et al. [8] courtesy Nature (London), Macmillan Magazines Ltd.*

cal samples embedded in glucose or vitrified ice or imaging near atomic resolution on thin inorganic samples. In a few cases with dedicated instrumentation it was possible to resolve biological molecules to the molecular level (e. g., Bacteriorhodopsin [27], Light Harvesting Complex LHC II [28], and Tubulin [29]).

As a complement to direct imaging, the diffraction of electrons at crystalline samples is also widely used.

13.9.1 Imaging at atomic resolution

Imaging of atom columns in crystalline samples is one application of EM necessary to find solutions to problems in modern solid state physics or material sciences. To overcome the optical flaws of conventional instrumentation, different methods have been alternatively studied. These include electron holography, high-energy TEM, and the correction of lens aberrations in the TEM. Whereas the first two methods try to work around the intrinsic optical problems of conventional TEMs, the work on a corrected TEM tries to reduce the spherical aberration of the objective lens to get a more ideal, directly interpretable image.

Figure 13.7 shows the differences between imaging using a conventional objective lens and a lens system with vanishing C_s (C_s corrector). As expected from theory and light optical devices, finite spherical aberration leads to an increased error disk of each image point and thus to contrast delocalization, contrast decrease, and lower resulting resolution and structural interpretability. The recent development of the electron optical corrector elements must therefore be seen as a major milestone in electron optical imaging.

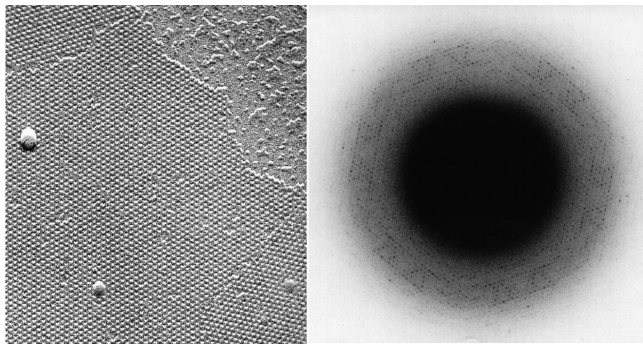


Figure 13.8: *Light Harvesting Complex II, electron micrograph (left, from Kühlbrandt [30], courtesy Nature (London), Macmillan Magazines Ltd.) and electron diffraction pattern (right, from Wang and Kühlbrandt [31], by copyright permission of Academic Press, San Diego).*

13.9.2 Imaging of biological samples

In contrast to inorganic material, biological samples cannot be imaged directly without great preparative efforts and dedicated instrumentation. The reason for this is the atomic composition of biological samples that almost completely consist of light atoms such as H, C, O, N, and minor contributions of P, S and others. As was discussed in Section 13.7, conventional preparation techniques have been developed that use heavy atom staining of the biological sample and the imaging of this stain instead of the biological sample itself. Obtainable maximum structural resolutions for such samples vary with specimen thickness and preparation technique between about 2 and 10 nm. Even if this resolution seems to be very limited, applying these methods to biological material many interesting questions have been answered and will be answered in the future.

To achieve higher structural resolution as is needed, for instance, for the determination of molecular structures in structural biology new preparation methods were necessary. With the development of cryotechniques it is possible to study cooled samples in dedicated cryomicroscopes. In such instruments samples are cooled down to liquid nitrogen (-196 °C) or liquid helium temperature (-269 °C). At this temperature samples are less susceptible to beam damage and even the light atoms of biological material can then be imaged. However, compared to material sciences samples, a much lower electron dose has to be used for imaging, which results in very noisy images. Therefore the obtainable resolution from such images is largely dependent on good imaging strategies to average over large ensembles of individual particles.

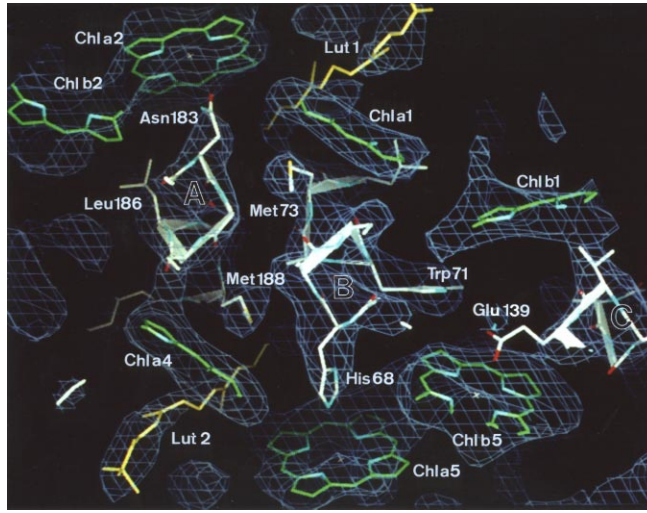


Figure 13.9: *Light Harvesting Complex II, 3-D representation. From Kühlbrandt et al. [28], courtesy Nature (London), Macmillan Magazines Ltd.*

Such averaging is performed best on crystalline samples. All the molecular structures obtained by EM resulted from studies of 2-D protein crystals, that is, crystalline monolayers of proteins. Such crystals can either be conventionally imaged or electron diffraction patterns can be recorded from them. Electron diffraction patterns of well-ordered protein crystals diffract very often to a resolution better than 2 Å. Unfortunately, imaging of such crystals does not in general yield phase information to the same resolution. Images are affected by specimen drift, specimen charging and the low signal-to-noise ratio of the images. Up to now, typical resolution limits for imaging on dedicated TEMs are of the order of 3-4 Å. Combining data from higher angle specimen tilt series, it is then possible to calculate 3-D density maps at this resolution. This resolution is sufficient for the determination of the molecular structure, that is, the known amino acid sequence can be built into such densities as a molecular amino acid chain. For a fine example that shows the application of all these techniques see the work on bacteriorhodopsin crystals [27]. Work on the Light Harvesting Complex II is shown in Figs. 13.8 and 13.9, which illustrate typical rotary shadowing images of crystals, electron diffraction patterns from such crystals, and the finally obtained 3-D density maps together with the built-in amino acid chain.

Besides the successful studies on 2-D crystalline assemblies of biological material other methods for analyzing objects with lower symmetry have been developed. Such systems are helical objects (actin filaments), icosahedral objects (viruses), or—without any symmetry—

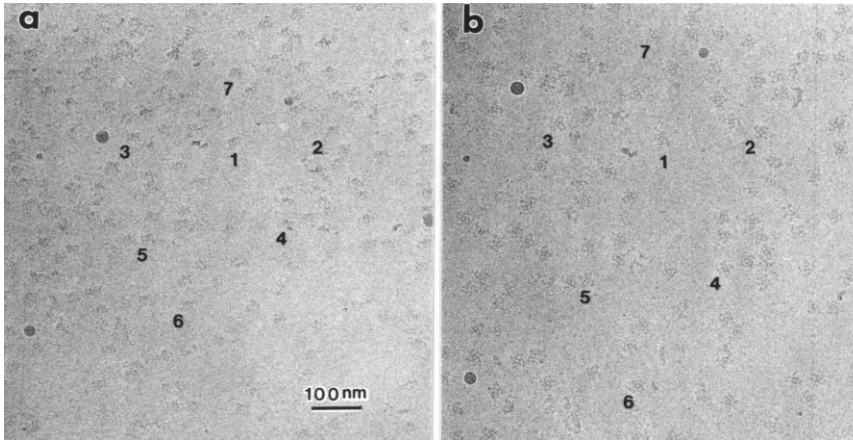


Figure 13.10: *Cryomicrographs showing top-view images of the calcium release channel as* ***a*** *tilted and* ***b*** *untilted specimen. Corresponding molecules are labeled with numbers. From Radermacher et al. [32], by copyright permission of Rockefeller University Press.*

assemblies of a nonsymmetric particle. Such "single particles" are special in the sense that no averaging methods can be applied unless a thorough classification and alignment procedure was applied to the image data set. Appropriate methods have been developed by Frank and van Heel; for a rigorous discussion of these methods see [33]. Because data averaging and alignment is much more tedious for single particles than for periodic objects, the structural resolution achieved up to now is not as high as for 2-D crystals. The best examples are reconstructions of viruses at 9 Å resolution (still applying icosahedral symmetry) [34] or 20-25 Å for ribosomes as large, asymmetric assemblies of protein and RNA [35, 36].

A general problem for all 3-D reconstruction in electron microscopy is the collection of the third dimension structural information from 2-D projection images. This problem has in general been solved using tomographic methods. For oriented samples such as 2-D crystals or other specimens that have a preferential orientation, this is realized by single axis tilting of the specimen in the microscope. Because the tilt axes for individual objects are randomly oriented, merging data from single axis tilted images give complete coverage in all three dimensions (random conical tilt for single particles, tilt series of 2-D crystals). Figures 13.10 and 13.11 illustrate this situation for a membrane channel protein. This flat biological object orients in only two views in the microscopical preparation (Fig. 13.11a,b, side view c does not occur in micrographs). For technical details of the random conical tilt method used here see [39]. In the case of completely random orientation of

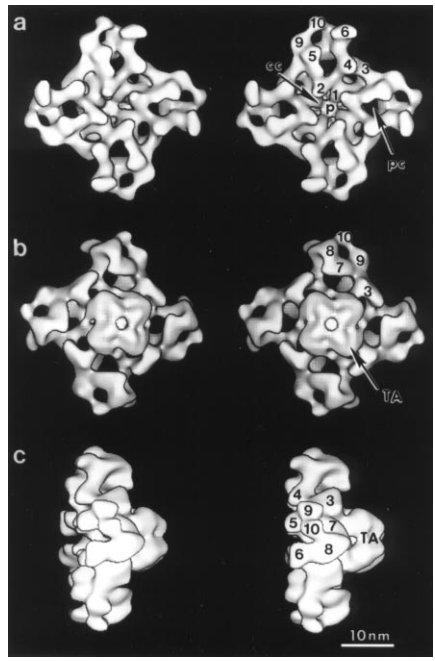


Figure 13.11: *Stereorepresentation of the reconstruction of the calcium release channel (Fig. 13.10)* ***a*** *cytoplasmic side* ***b*** *sarcoplasmatic side of the channel,* ***c*** *side view. From Radermacher et al. [32], by copyright permission of Rockefeller University Press.*

the object in the micrograph, for example, Fig. 13.12 for E. coli ribosomes, images of untilted objects can be aligned and merged to obtain the 3-D structure [35] (see van Heel [40] for technical details on direct alignment of objects of unknown Eulerian angles).

13.9.3 Electron diffraction in material sciences

As has been pointed out for biological samples, electron diffraction on crystalline samples is a powerful method to obtain additional high-resolution information about the specimen. In contrast to the normal parallel beam electron diffraction, convergent beams are often used in material science. In contrast to the diffraction on biological samples (Fig. 13.8), the corresponding diffraction patterns are not simple Bragg spot patterns. Different techniques have been developed, and from convergent beam patterns today many different kinds of information on the 3-D structure can be collected. Typical examples of convergent beam diffraction patterns are shown in Fig. 13.13 and Fig. 13.14.

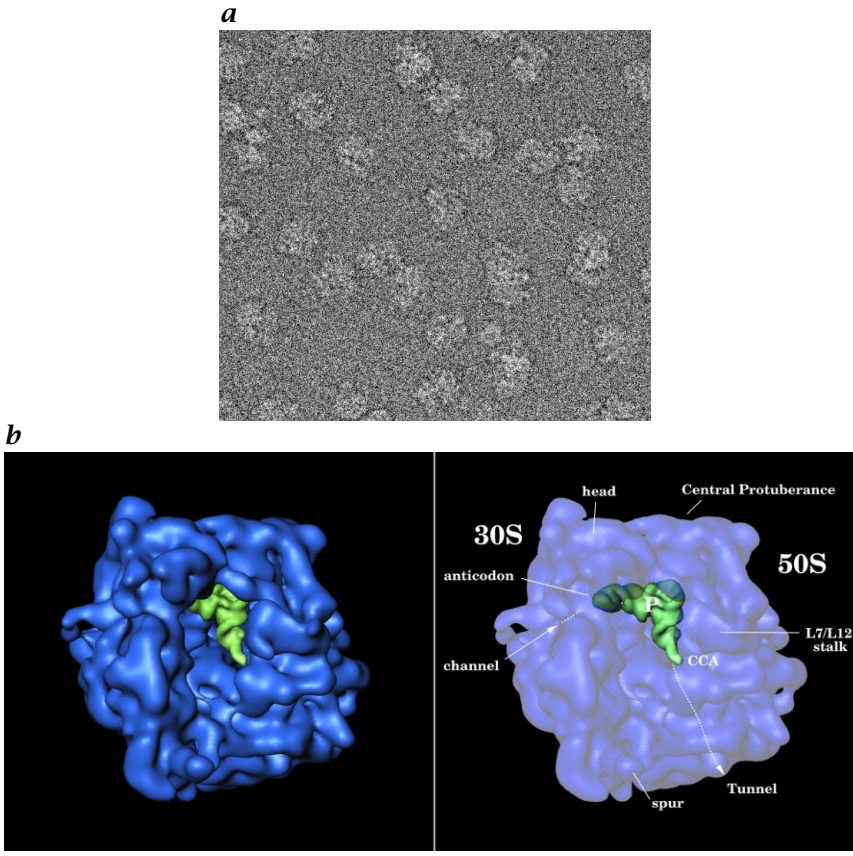


Figure 13.12: ***a*** *70S ribosomes from Escherichia coli, visualized by cryo-electron microscopy. Electron optical magnification 50,000×.* ***b*** *The ribosome at 15 Å, reconstructed from 30,000 projections obtained by cryoelectron microscopy, shown with a tRNA in the P-site position as experimentally found [37]. The anticodon of the tRNA is in the vicinity of the channel that is thought to conduct the messenger RNA [35], while the acceptor end (marked CCA) is seen to point toward the opening of the tunnel that is believed to export the polypeptide chain [38] [Prepared by Amy Heagle and Joachim Frank, Laboratory of Computational Biology and Molecular Imaging, Wadsworth Center]; (see also Plate 8).*

13.9.4 Element mapping in biology and material science

Often a quick and convenient way of producing chemical element maps is ESI in the EFTEM (Section 13.5.2). Here one or more energy-filtered images are acquired just before the onset of the interesting ionization edge (pre-edge images) and another one just after the onset (post-edge image).

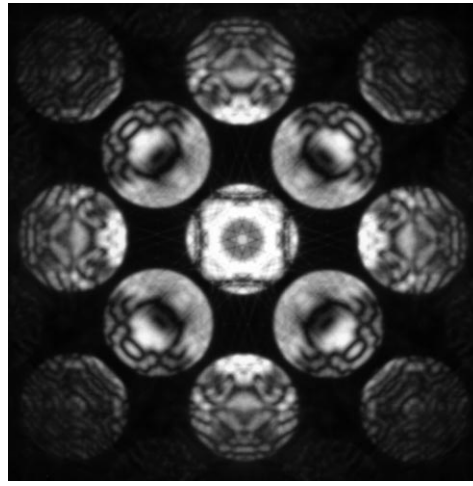


Figure 13.13: *Convergent beam electron diffraction pattern from a $MgAl_2O_4$ spinel single crystal obtained along the <100> zone axis. From such patterns information can be obtained on crystal structure and space group, structure factors and temperature factors (Debye-Waller-factors), charge density distribution and bonding charge densities [Prepared by Joachim Mayer, MPI für Metallforschung, Stuttgart, Germany].*

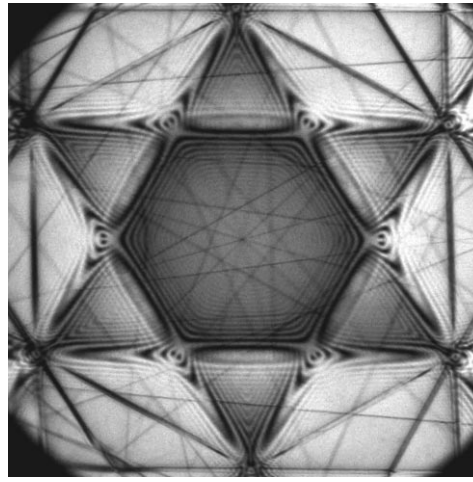


Figure 13.14: *Large-angle convergent beam electron diffraction pattern obtained with the Tanaka-technique on a LEO EM 912 microscope with zero loss energy filtering. The pattern was obtained from an Al single crystal in <111> orientation [Prepared by Joachim Mayer, MPI für Metallforschung, Stuttgart, Germany].*

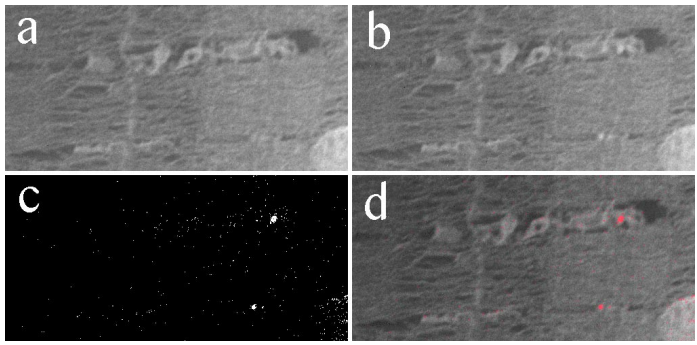


Figure 13.15: *Calcium mapping from selected images of an Image-EELS series:* ***a*** *pre-$Ca_{L2,3}$-edge (at 344 eV energy loss);* ***b*** *post-$Ca_{L2,3}$-edge (354 eV);* ***c*** *jump-ratio image (post-edge divided by pre-edge);* ***d*** *Ca-mapping (pre-edge + jump ratio (red)). Size of the imaged area 1.5 × 3.0 μm^2.*

The most simple method for element mapping from only two images (a pre-edge and a post-edge one) is jump-ratio imaging. The pre-edge image is divided by the post-edge image and the result is binarized by setting an appropriate gray value threshold. The resulting mask can be overlaid to a conventional bright-field image to show a pseudocolor distribution of the assayed element.

Figure 13.15 shows a biological example, the calcium-map of a freeze-dried section of murine skeletal muscle quick-frozen after a procedure that loads the cell's calcium stores, the terminal cisternae. Calcium is an extremely important "second messenger" substance that mediates signal transduction in a large class of cells such as muscle fibers and neurons. Therefore biologists and biophysicists are highly interested in revealing subcellular calcium distributions. The images were taken at a magnification of 6300× using an energy window of 8 eV centered at 344 eV energy loss to record the pre-$Ca_{L2,3}$-edge image (Fig. 13.15a) and at 354 eV for the post-$Ca_{L2,3}$-edge image (Fig. 13.15b). The jump-ratio image (Fig. 13.15c) was used to obtain the calcium mapping (Fig. 13.15d). In the right image half, two terminal cisternae containing high calcium concentrations are visible as red spots.

Despite the simplicity of the jump-ratio method it is quite insensitive to the production of mass thickness artifacts, that is, false positive element signals due to local thickness variations of the sample, provided the overall thickness is not higher than about half the mean free path for electrons of the given energy in that material. However, on biological specimen one often works at the detection limit of ESI, thus requiring more sophisticated processing methods.

An example from materials science using the three-window method is shown in Fig. 13.16. Here, two pre-edge images are used to extrap-

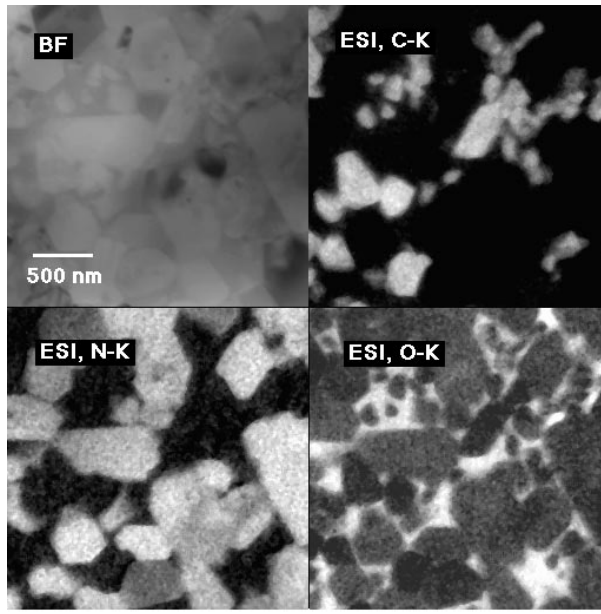


Figure 13.16: *Bright field image and elemental distribution images for C, N, O from a Si_3N_4-SiC composite ceramic prepared form polysilazan polymer precursors. The distribution of the Si_3N_4-, SiC-grains and the amorphous oxide can be seen in the individual elemental distribution images. Elemental distribution images obtained with the three window technique on the LEO EM 912 microscope. Figure from Mayer [41], by copyright permission of Elsevier Science Ltd., Oxford, England.*

olate the pre-edge background to the post-edge region according to a function of the energy loss, the parameters of which are calculated from the pre-edge image intensities for every pixel. The net element distribution is then acquired by subtracting the calculated background intensity from the post-edge image. Our example shows a conventional bright-field image of a sintered polymer derived Si_3N_4/SiC-composite ceramic (top left), a carbon map (top right), a nitrogen map (bottom left) and an oxygen map (bottom right) obtained from the same sample area by this three-window method.

13.9.5 SEM image examples

The advantage of SEM is that large bulk samples can be imaged without or only slight fragmentation of the specimen and hence disturbing its natural context. In addition, the large depth of focus in SEM allows to image large and complex surface structures in focus and hence makes SEM a powerful tool for the determination of surface structures.

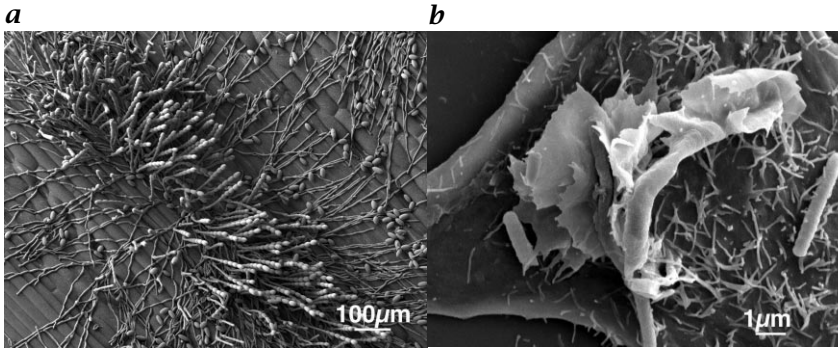


Figure 13.17: *Examples for the large depth of focus in SEM images:* ***a*** *Barley leaf infected by plant fungi (by courtesy of Dr. U. Hässler);* ***b*** *bacteria infecting HeLa cells.*

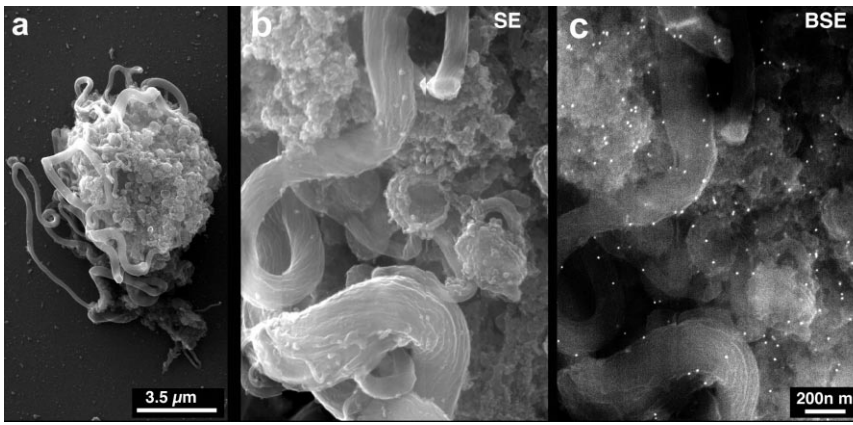


Figure 13.18: *Example for the different SEM imaging modes:* ***a*** *insect cell labeled with Au antibodies against a membrane receptor (by courtesy of Dr. M. Cyrclaff);* ***b*** *detail of* ***a*** *in the SE imaging mode;* ***c*** *same detail in the BSE mode. Single Au clusters (10 nm) can be visualized and hence the receptor localized on the membrane.*

Two examples for the large depth of focus in SEM images are shown in Fig. 13.17. Figure 13.17a shows a cryo-SEM image of a Barley leaf infected by plant fungi. Fungal hyphae with large fruit bodies pointing out several millimeters from the leaf surface are imaged in focus, allowing to see the wax structures on the leaf as well as details of the fruit bodies. In Fig. 13.17b, a cryo-SEM image of the complex defense reaction of a HeLa cell, the formation of several microns high membrane ruffles against a bacterial attack of shigella can be imaged in focus. The different signals that are generated on a specimen during imaging at high acceleration voltages allow imaging of the surface by *secondary*

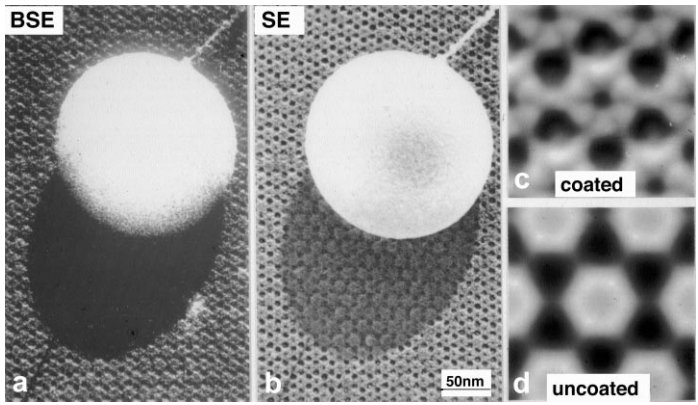


Figure 13.19: *Effects of thin metal coating of the specimen:* ***a*** *BSE image of a metal-coated 2-D protein crystal partly covered by a latex bead;* ***b*** *corresponding SE image;* ***c*** *averaged core structure from coated; and* ***d*** *from uncoated protein.*

electrons (SE) and Z-number contrast by *backscattered electrons* (BSE). An example for the different signals generated on a specimen from biological science is shown in Fig. 13.18. By using Au-labeled antibodies against a surface protein (membrane receptor) on a insect cell in culture (Fig. 13.18a overview, Fig. 13.18b zoom-in image of a membrane surface area in the SE imaging mode), it is possible to localize the position of the receptors in the BSE image (Fig. 13.18c, the corresponding image recorded in the BSE mode) by the intense BSE signal generated at the high Z number material and to correlate it with the surface features obtained in the SE image.

In Fig. 13.19 secondary and backscattered electrons generated on a metal coated thin 2-D protein crystal (*HPI-layer*) are not only used to determine the protein structure by cross correlation, but also allow demonstration of the intrinsic difference in signal localization and resolution. Backscattered electrons (Fig. 13.19a) are only generated on specimen areas which are metal coated (1 nm W) whereas SE (Fig. 13.19b) are generated on uncoated as well as on coated areas.

The SE yield on coated areas is roughly 5 times higher than on the uncoated protein area. In the BSE mode the structures are imaged with a higher local precision and hence higher resolution than in the SE mode. A comparison of the averaged core structure from the coated area of the SE image with the core structure from the uncoated area proves that the additional metal film of 1 nm W not only localizes the signal in the lateral but also in the vertical direction.

On uncoated proteins the resolution is limited at about 5 nm due to the lateral diffusion of low energy electrons (<50 eV), whereas the distinct and limited expansion of a W grain (1-2 nm) helps to keep the

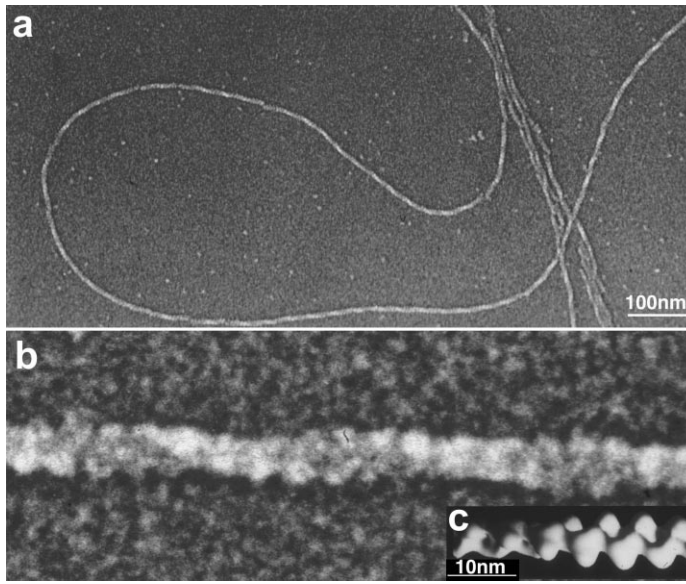


Figure 13.20: *SEM images of F-actin filaments:* ***a*** *overview;* ***b*** *zoom-in raw data;* ***c*** *3-D model from STEM imaging.*

generated SE signal localized to the metal grain position. The image is therefore built up by small illuminated particles (the metal grains) and the signal intensity depends on the mass thickness of these particles as well as on the slope of the specimen on which these flat metal grains lay. Therefore the surface structure can only be extracted after averaging of several subunits. The obtainable maximum structural resolution for such samples varies with the coating film thickness and preparation technique between 1.5 and 2 nm.

Due to the high SNR in the SE image, direct imaging of molecular structures such as the actin subunits in a F-actin filament becomes possible (Fig. 13.20a and b). The SE image of the F-actin filament is astonishingly similar to the surface representation of the 3-D model from STEM data of negatively stained F-actin filaments (Fig. 13.20c). Since the contrast contribution of the protein is not negligible (see Fig. 13.19b and d), caution in image interpretation is necessary because the topographic contrast in SEM at molecular level is not yet completely understood.

Acknowledgments

We would like to thank the following persons for their contribution of figures: Dr. J. Frank and Dr. T. Wagenknecht, Wadsworth Center, Albany, NY, USA, Dr. J. Mayer, MPI für Metallforschung, Stuttgart, Ger-

many, Dr. M. Haider, CEOS GmbH, Heidelberg, Germany, Dr. W. Kühlbrandt, MPI für Biophysik, Frankfurt, Germany.

13.10 References

[1] Glaser, W., (1952). *Grundlagen der Elektronenoptik*. Wien: Springer.

[2] Knoll, M. and Ruska, E., (1932). A contribution to geometrical electron optics. *Ann. Phys.*, **12**:607–640.

[3] von Ardenne, M., (1938). Das Elektronen-Rastermikroskop. Theoretische Grundlagen. *Z. Phys.*, **109**:533–572.

[4] Oatley, C. W., Nixon, W. L., and Pease, R. F. W., (1965). Scanning electron microscopy. *Adv. Electronics Electron Phys.*, **21**:181–247.

[5] Hawkes, P. (ed.), (1985). *The Beginnings of Electron Microscopy. Advances in Electronics and Electron Physics, Suppl. 16*. Orlando, FL: Academic Press.

[6] Thon, F., (1966). Defocus dependence of the phase contrast in the electron microscopic image. *Z. Naturforschung*, **21a**:476–478.

[7] Thon, F., (1971). Phase contrast electron microscopy. In *Electron Microscopy in Material Science*, U. Valdr, ed. New York: Academic Press.

[8] Haider, M., Uhlemann, S., Schwan, E., Rose, H., Kabius, B., and Urban, K., (1998). Development of the first spherical aberration corrected 200 kV transmission electron microscope. *Nature*, **392**:768–769.

[9] Reimer, L. (ed.), (1997). *Transmission Electron Microscopy,* 4th edition. Berlin, Heidelberg, New York: Springer Verlag.

[10] Heinrich, K. and Newbury, D. (eds.), (1991). *Electron Probe Quantization*. New York: Plenum Press.

[11] Scott, V., (1995). *Quantitative Electron Probe Microanalysis*. New York: Ellis Horwood.

[12] Egerton, R. F., (1996). *Electron Energy-loss Spectroscopy in the Electron Microscope,* 2nd edition. New York: Plenum Press.

[13] Joy, D. C., (1984). Beam interactions, contrast and resolution in the SEM. *Jour. Microsc.*, **136**:241–258.

[14] Reimer, L. and Pfefferkorn, G., (1977). *Rasterelektronenmikroskopie*. Berlin, Heidelberg, New York: Springer Verlag.

[15] Nagatani, T., Saito, S., and et al., (1987). Development of an ultra high resolution scanning electron microscope by means of a field emission source and in-lens system. *Scanning Microscopy*, **1(3)**:901–909.

[16] Peters, K. R., (1984). Generation, collection and properties of an SE-I enriched signal suitable for high resolution SEM on bulk specimen. In *Electron Beam Interactions with Solids*, D. F. Kyser, D. E. Newbury, H. Niedrig, and R. Shimizu, eds., pp. 363–372. AMF O'Hare.

[17] Koike, H., Ueno, K., and Suzuki, M., (1970). Scanning device combined with conventional electron micoscope. In *Proc. 29th Ann. Meeting of EMSA*, p. 28.

[18] Echlin, P., (1979). Thin films for high resolution conventional scanning electron microscopy. *Scanning Electron Microsc.*, **2**:21–30.

[19] Hermann, R. and Müller, M., (1991). High resolution biological scanning electron microscopy: A comparative study of low temperature metal coating Techniques. *Jour. El. Mic. Techn.*, **18**:440–449.

[20] Wepf, R., Bremer, A., Amrein, M., Aebi, U., and Gross, M., (1992). *Surface imaging of F-actin filaments: a comparative study by SEM, TEM and STM*, Vol. III. Secretariado del Publicaciones de la Universidad de Granada.

[21] Wells, O. C., (1974). Resolution of the topographic image in the SEM. *Scanning Electron Microsc.*, **I**:1–8.

[22] Walther, P. and Hentschel, J., (1989). Improved representation of cell surface structures by freeze substitution and backscattered electron imaging. *Scanning Microsc.*, **3**:201–211.

[23] Wepf, R., Amrein, M., et al., (1991). Platinum/iridium/carbon: a high-resolution shadowing material for TEM, STM and SEM of biological macromolecular structures. *J. Microsc.*, **163(1)**:51–64.

[24] Reimer, L., (1979). Electron-specimen interaction. *Scanning Electron Microsc.*, **II**:111–124.

[25] Glauert, A. (ed.), (1981). *Practical Methods in Electron Microscopy*. Amsterdam: North-Holland.

[26] Hawkes, P. (ed.), (1980). *Computer Processing of Electron Microscopic Images*. Berlin, Heidelberg, New York: Springer Verlag.

[27] Henderson, R., Baldwin, J., Ceska, T., Zemlin, F., Beckmann, E., and Downing, K., (1990). Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *Jour. Mol. Biol.*, **213**:899–929.

[28] Kühlbrandt, W., Wang, D., and Fujiyoshi, Y., (1994). Atomic model of the plant light-harvesting complex by electron crystallography. *Nature*, **367**: 614–621.

[29] Nogales, E., Wolf, S., and Downing, K., (1998). Structure of the ab tubulin dimer by electron crystallography. *Nature*, **391**:199–203.

[30] Kühlbrandt, W., (1984). Three-dimensional structure of the light-harvesting chlorophyll a/b-protein complex. *Nature*, **307**:478–480.

[31] Wang, D. N. and Kühlbrandt, (1991). High-resolution electron crystallography of light-harvesting chlorophyll a/b-protein complex in three different media. *Jour. Mol. Biol.*, **217(4)**:691–699.

[32] Radermacher, M., Rao, V., Grassucci, R., Frank, J., Timerman, A., Fleischer, S., and Wagenknecht, T., (1994). Cryo-electron microscopy and three-dimensional reconstruction of the calcium release channel/ryanodine receptor from skeletal muscle. *Jour. Cell Biol.*, **127**:411–423.

[33] Frank, J., (1996). *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. San Diego: Academic Press.

[34] Böttcher, B., Wynne, S., and Crowther, R., (1997). Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. *Nature*, **386**:88–91.

[35] Frank, J., Zhu, J., Penczek, P., Li, Y., Srivastava, S., Verschoor, A., Grassucci, R., Lata, R., and Agrawal, R., (1995). A model of protein synthesis

based on cryo-electron microscopy of the E. coli ribosome. *Nature*, **376**: 441–444.

[36] Stark, H., Mueller, F., Orlova, E., Schatz, M., Dube, P., Erdemir, T., Zemlin, F., Brimacombe, R., and van Heel., M., (1995). The 70S Escherichia coli ribosome at 23 Å resolution: fitting the ribosomal RNA. *Structure*, **3**: 815–821.

[37] Malhotra, A., Penczek, P., Agrawal, R. K., Gabashvili, I. S., Grassucci, R. A., Juenemann, R., Burkhardt, N., Nierhaus, K. H., and Frank, J., (1998). Escherichia coli 70S ribosome at 15 A resolution by cryo-electron microscopy: localization of fMet-tRNA(f/Met) and fitting of L1 protein. *Jour. Mol. Biol.*, **in press**.

[38] Beckmann, R., Bubeck, D., Grassucci, R. A., Penczek, P., Verschoor, A., Blobel, G., and Frank, J., (1997). Alignment of conduits for the nascent polypeptide chain in the ribosome-Sec61 complex. *Science*, **278**:2123–2126.

[39] Radermacher, M., (1988). Three-dimensional reconstruction of single particles from random and nonrandom tilt series. *Jour. Electr. Microsc. Tech.*, **9**:359–394.

[40] van Heel, M., (1987). Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy*, **21**:111–124.

[41] Mayer, J., (1996). Polymer-derived Si-based bulk ceramics: 2. Microstructural characterization by electron spectroscopic imaging. *Jour. of the European Ceramic Society*, **15**:717–727.

14 Processing of Ultrasound Images in Medical Diagnosis

Werner Albert[1], and Madhukar Pandit[2]

[1]Westpfalz-Klinikum, Kaiserslautern, Germany
[2]Regelungstechnik und Signaltheorie, Universität Kaiserslautern

14.1 Introduction

Since its inception five decades ago, *ultrasound imaging* has become an invaluable and versatile tool with an increasing sphere of applications in medical diagnosis. The increasing capabilities of signal processing hardware and algorithms lead to a steady enhancement of performance and utility of commercially available ultrasound equipment. Improved spatial resolution and image quality resulting from electronically controlled aperture and focus allow the physician to use an ultrasound imaging system as a sensitive probe in the diagnosis of ailments by associating *image features* with organ and tissue characteristics of the patient.

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

Most clinical ultrasound imaging systems operate on the impulse-echo principle Cho et al. [1], Hill [2]. A pulse train of acoustic waves consisting of an rf sinusoidal carrier modulated by impulses of appropriate form is launched into the region under investigation by means of one or more piezocrystals. The reflected and backscattered acoustic waves are collected by the same crystals in the intervals interspersed between the transmitted impulses and converted into voltage signals. By processing the received signal, an image of the variations of the acoustic impedance of the area under investigation is formed. Figure 14.6 shows typical B-mode images. Multiple reflections and diffraction are accompanying phenomena that affect the received signal and deteriorate the ultrasound imaging. Furthermore, the acoustic signal is attenuated by the tissue—with increasing frequency, the attenuation increases. Thus, the depth up to which ultrasound imaging is feasible is limited by the frequency. As, on the other hand, resolution increases with frequency, trade-off between depth and resolution of imaging is inevitable.

To counteract the effects of attenuation, multiple reflection, refraction etc., techniques are developed with degrees of sophistication, which ever increasingly go hand-in-hand with the availability of powerful digital electronics. These techniques employ the following measures:

1. Several (up to a few hundred) piezocrystals are employed in the transmitting and receiving transducers.
2. Electronic beam forming and focusing techniques are employed.
3. Attenuation is compensated.
4. Several images are acquired, stored, and processed to obtain a resulting image with a higher signal to noise ratio (SNR).

The image quality achievable is limited by the physics of ultrasound and its interaction with tissue. One limiting factor of the fidelity of imaging is the amplitude and phase aberration of the ultrasound waves. A current topic of research and development is the compensation of the effects of these aberrations. Generally, a huge volume of data is generated and has to be handled. The processing of the beam-formed signal is incoherent, that is, the voltage signal corresponding to the superposition of received sinusoidal acoustic echo signals is envelope-detected with no consideration being given to the phase. This leads to interference of the echoes and gives rise to the familiar but disturbing speckles in the image.

Ultrasound imaging was first employed in clinical diagnosis in the "*A-mode*" in 1945. Here, the acoustic echoes (which depict the acoustic impedance variations) in an organ or tissue along a scanning straight-line direction were imaged. The technique became popular with the advent of "B-mode" 2-D imaging systems, which were introduced subsequently in the 1950s. These are the systems currently in widespread

Table 14.1: *Some areas of applications of ultrasound imaging*

Organ	Indication	Type of ultrasound equipment and frequency
Blood vessels	Stenosis, Thrombosis	B-mode, 3–7.5 MHz, Duplex sonography, Color Doppler
Thyroid	Tumors, Enlargement, Cysts	B-mode, 5–7.5 MHz, Color Doppler
Heart	Pathological condition, Enlargement, Blood flow	B-mode, 2–7.5 MHz, M-mode, Color Doppler, Transoesophagial echo cardiography
Liver	Enlargement, Tumors, Cysts	B-mode, 2–7.5 MHz, Color Doppler
Gall bladder and Bile ducts	Enlargement, Calcification, Stones	B-mode, 2–7.5 MHz
Spleen and Lymph nodes	Enlargement, Tumors	B-mode, 2–7.5 MHz
Pancreas	Enlargement, Tumors, Inflammation	B-mode, 2–7.5 MHz, Endosonography, 5–12 MHz
Gastro-intestinal tract	Tumors, Ulcers, Inflammation	B-mode, 2–7.5 MHz, Endosonography, 5–12 MHz
Kidneys and Urinary tract	Tumors, Obstructions, Stones	B-mode, 2–7.5 MHz, Color Doppler
Prostata	Enlargement, Tumors	B-mode, 2–7.5 MHz, Endosonography, 5–12 MHz
Uterus	Fetal physiology	B-mode, 2–7.5 MHz
Joints and muscles	Calcification, Defects, Inflammation	B-mode, 5–7.5 MHz

use for imaging sections of organs and tissues in clinical practice. Further developments include *M-mode imaging* employed for monitoring cardiac movement and Doppler systems for imaging blood flow, especially in the heart, kidney, and thyroid glands. Ultrasound endoscopes are being developed for various areas of application. The steady increase in the capabilities of modern digital microelectronics and software makes advanced signal and image processing in ultrasound imaging systems feasible and leads to a new generation of machines every 7 to 10 years. The popularity of ultrasound imaging systems is due to the following advantages:

1. real-time imaging capability;

2. flexibility and ease of manipulation with regard to the selection of the desired section and region of interest to be imaged;
3. noninvasiveness; it is generally acknowledged that ultrasound waves do not cause any tissue damage at densities lower than 100 mW cm^{-2}; clinical B-mode imaging systems employ densities less than 1 mW cm^{-2}; and
4. convenience and low operating costs.

The main disadvantage of ultrasound is that, by and large, it offers a qualitative and not a quantitative method for assessing organ and tissue characteristics.

In clinical practice, ultrasound imaging systems find widespread applications in investigations pertaining to abdomen and thorax, particularly in prenatal checks and checks of digestive tracts. Imaging systems employing ultrasound in conjunction with Doppler techniques are used for detecting functional disorders of the heart by monitoring blood flow. These systems are also used for monitoring the blood flow activity in other organs such as kidneys, thyroid glands, and blood vessels. Table 14.1 shows details of some applications of ultrasound imaging in clinical diagnosis.

14.2 Ultrasound imaging systems

Ultrasound imaging systems work on the echo principle. In the generic system, an acoustic wave in the form of a short rf impulse is launched into the body. Typically, the wave could have a duration of 1 μs and a center frequency of 3.5 MHz; it is generated by exciting the piezoelectric crystals of a transducer by a voltage impulse. It is partly reflected and/or scattered back when it meets continuous and discontinuous variations of the acoustic impedance of the tissue in which it is propagated. Reflection refers to the phenomenon that occurs when the geometrical dimensions of the boundaries are larger than the wavelength; scattering refers to the phenomenon when the dimensions are equal to or less than the wavelength. The component of the waves reflected and backscattered in the direction of the transducer is converted by the same piezoelectric crystals into an electrical signal. This received signal has a duration which is several hundred times the duration of the impulse launched.

The variations of the amplitude of the received signal bear the information regarding the changes of acoustic impedance of the tissues along the direction of propagation of the impulse. To enhance the SNR, not one pulse but a pulse train is launched into the body and the corresponding individual echo responses delayed and superposed to form the received signal. This signal is processed and, finally, the varia-

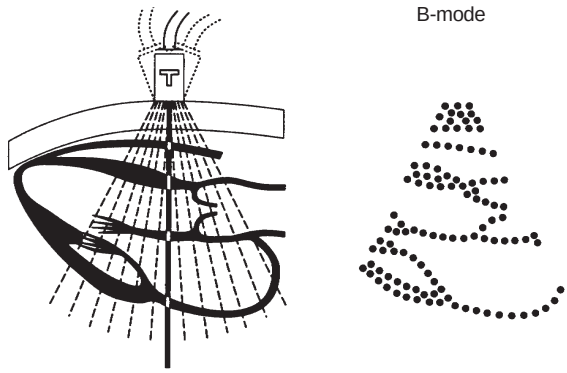


Figure 14.1: *B-mode imaging of the heart.*

tions of the amplitude are depicted as a function of the corresponding depths at which these occur. This is the "A-mode" image. An alternative method of display is to make the brightness of a point moving in a straight line corresponding to the propagation direction proportional to the amplitude of the received signal. By successively sweeping the direction of propagation in a plane and recording the series of "A-mode" images as straight lines with varying brightness next to one another, one obtains the "B-mode" image. Figure 14.1 shows the formation of a B-mode image of the heart

To be able to interpret the image and study the relation between the images and the acoustic properties of the organs and tissues that are imaged, it is necessary to understand the mechanism of wave propagation. An exact analysis of the mechanism is the subject of continuing research and is complex. In the following, basic equations according to Cho et al. [1] are given.

14.2.1 Mechanisms of ultrasound wave propagation

The acoustic wave launched by the transducer is propagated in the tissue as a longitudinal pressure wave that causes local variations of pressure density and velocity of the medium. The tissue is assumed to act like an isotropic nonabsorbing homogeneous fluid in which shear forces are negligible. These and further assumptions make the problem mathematically tractable; however, they are valid only as approximations. The plane wave equation of acoustic waves in a homogeneous

lossless medium is:

$$\frac{\partial^2 p}{\partial x^2} = \frac{1}{c_0^2}\frac{\partial^2 p}{\partial t^2} \tag{14.1}$$

with the local pressure p, the spatial coordinate x, and the velocity of sound c_0.

If the medium is nonhomogeneous, the density ρ is a function of the spatial coordinate x as is the compressibility κ. Then we have:

$$\frac{\partial^2 p}{\partial x^2} = \frac{1}{c_0^2}\frac{\partial^2 p}{\partial t^2} = \frac{\gamma(x)}{c_0^2}\frac{\partial^2 p}{\partial t^2} + \frac{\partial}{\partial x}\left(\mu(x)\frac{\partial p}{\partial x}\right) \tag{14.2}$$

with

$$\mu(x) = \frac{\rho(x) - \rho_0}{\rho(x)} \qquad \text{and} \qquad \gamma(x) = \frac{\kappa - \kappa_0}{\kappa_0}$$

The 3-D version of the preceding equation is, with the spatial coordinate vector r,

$$\nabla^2(p(r,t)) - \frac{1}{c_0^2}\frac{\partial^2 p(r,t)}{\partial t^2}$$

$$= \frac{1}{c_0^2}\frac{\partial^2 p(r,t)}{\partial t^2}\left[\frac{\kappa - \kappa_0}{\kappa_0}\right] + \nabla\left\{\left[\frac{\rho - \rho_0}{\rho}\right]\nabla p(r,t)\right\} \tag{14.3}$$

At time $t = 0$, an impulse is applied at $r = 0$. At an instant $t_R = R/c_0$, the pulse encounters a volume V located at a distance R as an incident pressure wave and is scattered. To determine the scattered wave, scattering is considered to be weak, that is, it is assumed that the amplitude of the incident wave is much larger than that of the scattered wave. This holds true if $|\rho - \rho_0| \ll \rho_0$ and $|\kappa - \kappa_0| \ll \kappa_0$. Then one has the homogeneous equation for the scattering region:

$$\nabla^2 p_0(r,t) - \frac{1}{c_0^2}\frac{\partial^2 p_0(r,t)}{\partial t^2} = 0 \tag{14.4}$$

with the incident field, $p_0(r,t)$. One can approximate the solution of the homogeneous equation in the region of focus

$$p_0(r,t) = A\,(R + z - c_0 t)\,B(x,y) \tag{14.5}$$

where A and B represent the axial pulse and the beam characteristics, respectively. The solution of the inhomogeneous equation yields

the expression for the backscattered wave $p_s(r,t)$ at the transducer ($r = R$):

$$p_s(R,t) = \frac{1}{4\pi R}\int A(2R + 2z - c_0 t)H(z)\,\mathrm{d}z \tag{14.6}$$

with

$$H(z) = \frac{1}{4}\frac{\partial^2}{\partial z^2}\int\left[\frac{\rho_1(x,y,z)}{\rho_0} - \frac{\kappa_1(x,y,z)}{\kappa_0}\right]B(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

The direction of propagation is along z. Thus, the amplitude of backscattered signal can be interpreted as the limit of the weighted sum of infinitesimal component signals weighted by $H(z)$. The function $H(z)$ itself is expressed as

$$H(z) = \frac{1}{4}\frac{\mathrm{d}^2}{\mathrm{d}z^2}Z_{\text{eff}}(z) \tag{14.7}$$

where $Z_{\text{eff}}(z)$ represents the equivalent acoustic impedance averaged over the beam cross section. Equation (14.7) indicates that the amplitude contributions are large from those regions where $Z_{\text{eff}}(z)$ exhibits large changes of the acoustic impedance. The voltage generated at the transducer by the pressure variations is the integral over the transducer surface:

$$V(R,t) = c_T\int_{S_\xi} p_s\left((R+\xi),t\right)\,\mathrm{d}s_\xi \tag{14.8}$$

where c_T = a constant. The tissue impulse response is defined as

$$g_R(b,t) = V(R,t)\big|_{H(z)=\delta(z-b)} = \frac{c_T}{4\pi R}\cdot\int_{S_\xi} A(2R + 2b - c_0 t)\,\mathrm{d}S_\xi \tag{14.9}$$

so that equations 14.6, 14.8, and 14.9 can be combined to obtain

$$V(R,t) = \int g_R(z,t)H(z)\,\mathrm{d}z \tag{14.10}$$

This equation indicates the factors involved in the signal components in the various scan directions and is valid for an elementary impulse and volume. The actual voltage of the transducer has to be determined by taking the superposition of the responses of the delayed elementary impulses into consideration. Apparently it is not easy to estimate $\rho(x,y,z)$ and $\kappa(x,y,z)$ (which would be a mapping of the acoustic parameters of the tissue from the equation). Simulation models are employed in the design of transducers.

14.2.2 Implementation aspects

Implementation of ultrasound imaging techniques to produce commercial machines involves application using hardware currently available. As this is a continually changing scenario one can only sketch the general framework of implementation aspects. First, we deal with B-mode imaging equipment and then ultrasound equipment with the capability of measuring and/or imaging blood flow.

An ultrasound imaging system for B-mode images consists of the basic unit, which provides for the generation of pulse pattern, signal processing, display and output, and the transducers, which serve to convert the electrical signals from the basic unit into acoustic signals and the backscattered acoustic signals back to the electrical signals.

Signal acquisition and processing. The interrogating pulse has a carrier frequency in the range of 2 to 7 MHz, duration of 2 to 3 cycles, and an envelope that is approximately a Gaussian function. The pulse repetition rate lies in the range of 1 to 3 KHz. Considering that the velocity of acoustic waves in tissues is approximately that in water (viz. 1540 m/s), the wavelengths lie in the range of 0.20 to 0.75 mm. This figure also represents the axial resolution of the imaging. The lateral resolution is also dependant on the array dimensions and the focusing. One major problem with ultrasound imaging is the attenuation of the acoustic waves caused by the tissues. The attenuation depends on the tissue and frequency; however, as a rule of thumb it can be taken to be 2 db/(depth in cm $\times$ the frequency in MHz). Thus a 100-db attenuation is typical in abdominal examinations. Compensation of attenuation is a prerequisite for a uniformly bright image. This is achieved by "Time gain control" (TGC)—sometimes also called "Depth gain control" (DGC) of amplification. An important aspect is the tradeoff between axial resolution and the attenuation. Lateral resolution is enhanced by focusing the beam using acoustic lenses and electronic means. The latter is realized by employing timing control of the transmitted pulses and delaying the received signals in a precise predetermined pattern.

Modern ultrasound imaging systems use digital electronics and signal processing. The sampling rate for the AD converters and delay lines must be chosen at a value at least twice the highest bandwidth of the incoming signals, that is, at least 6 times the carrier frequency. Usually a value of 8 times the carrier frequency is chosen. Digital electronics permit the achievement of high precision and stability of the amplifiers and delay lines. Furthermore, they allow flexible advanced signal processing algorithms to be incorporated. These result directly in a high image quality and reproducible settings of the imaging equipment. The latter is especially important for detecting tissue and organ changes with time.

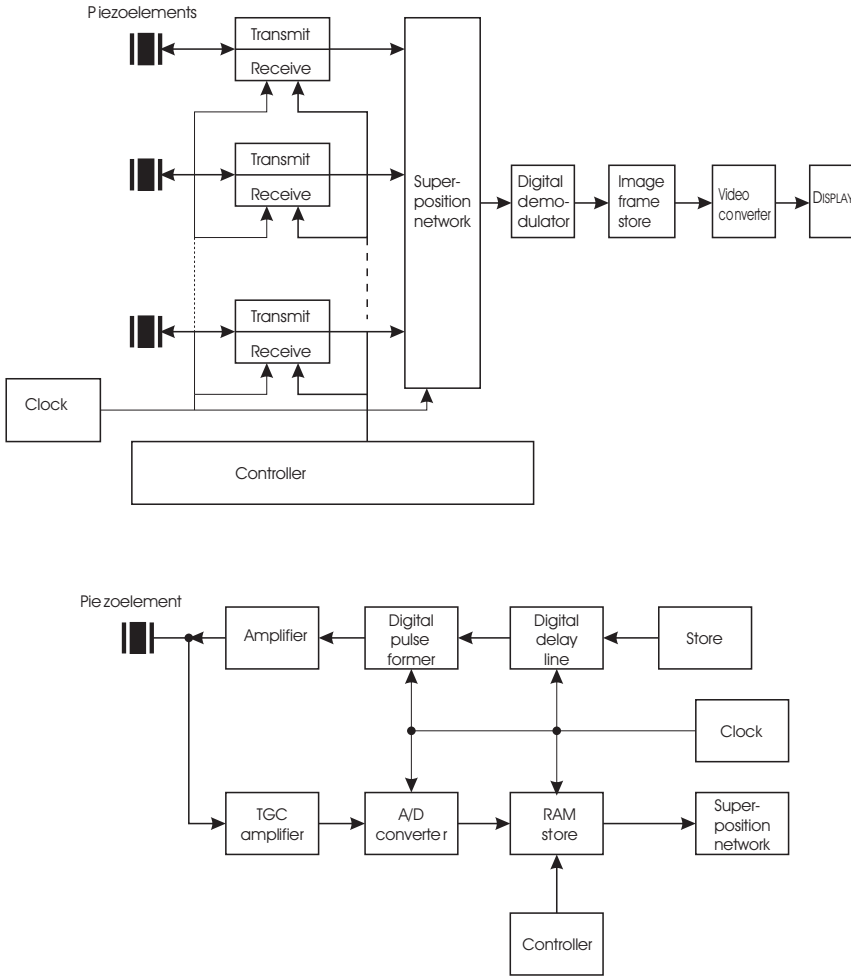


Figure 14.2: *Block-diagram of a B-mode imaging system.*

Figure 14.2 shows the block diagram of a B-mode imaging system. The transmitting and receiving piezoelements are activated in groups. Timing and delay are controlled in relation to the depth of the region imaged. By doing this, the beam is focused region-wise in several regions. The implementation of this technique of "dynamic focusing" can be performed elegantly by employing digital signal processing. The received signals in the individual groups are digitized by means of the AD converters, delayed appropriately and finally summed up.

The sum is then demodulated using a digital envelope detector, thus achieving a high SNR even for the low-amplitude signals (which are especially important). The SNR is further enhanced by employing a

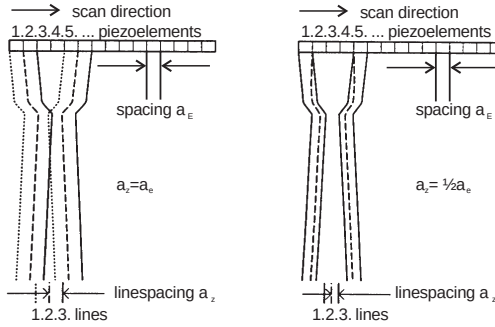


Figure 14.3: *Scanning with a linear array.*

digital filter with the envelope detector. The signal is processed to compose the B-mode image, which consists of lines and pixels as in TV frames. Finally, the signal samples are stored in a 2-D array. The gray level of a pixel is coded with 6 to 8 bits. Provision is made for selecting a *region of interest* (ROI), which can be zoomed out to observe details better. This is converted into a video norm signal and displayed in the monitor. The store also facilitates the choice of the desired ROI.

Transducer types. A single stationary piezocrystal maps the acoustic impedance along the scanning direction, that is, the direction of transmission and reflection of the acoustic wave. In order to map the acoustic impedance in a plane section, the scanning direction is successively rotated by controlling the position of the crystal mechanically, and the impedance mapped along each of the corresponding straight lines. Furthermore the signal strength is enhanced by using several (up to 5 crystals) instead of a single one. An alternative often implemented in modern apparatus consists in using an array of crystals and electronically controlling the activation of the crystals. Although mechanically operated transducers are prone to wear, they have the advantage that they allow higher carrier frequencies (up to 50 MHz). Electronic operation eliminates moving parts but limits the frequency to 10 MHz. As this frequency is also the upper limit due to attenuation considerations, it is not a serious issue.

In electronically controlled transducers, the crystals are arranged along a straightline (linear array) or on the outer periphery of an arc (curved array or convex array). Typically, 50 to 500 crystals are used. Figure 14.3 shows a linear array. Here, the crystals are activated in groups of 4.

First, crystals 1 to 4, next 2 to 5, and so on, are activated. By delaying the signals from the individual elements by appropriate amounts, a

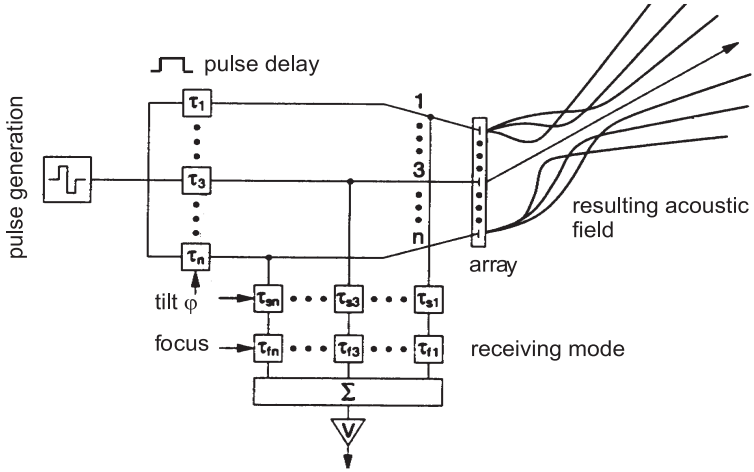


Figure 14.4: *Beam forming in an array transducer.*

plane of focus is created. Intensifying the focusing effect reduces the effective size of the region of interest, but increases lateral resolution. Lateral resolution is made to lie in the same range as axial resolution.

Figure 14.4 shows a phased-array transducer in which the piezo-elements are not fired in groups but consecutively from the first to the last.

By delaying the firing of the individual elements appropriately, both while transmitting and while composing the received signal, a beam that sweeps across a sector is generated for mapping the tissue or organ.

14.2.3 Ultrasound Doppler imaging

Ultrasound imaging systems that offer the additional capability of imaging blood flow exploit the Doppler effect. These operate in the so-called Duplex mode in which the blood velocity is measured and displayed. Simultaneously, the B-scan of the volume element in which the velocity is measured is also displayed. Thus, the scanner is equipped with velocity measuring and B-mode imaging subsystems.

The various forms in which Doppler techniques are implemented are:

a. continuous wave (CW-) Doppler;
b. pulsed wave (PW-) Doppler; and
c. color Doppler.

The *continuous wave (*CW-*) Doppler* employs not pulses but a continuous sinusoidal signal. Two separate piezocrystals, one for transmission and one for receiving the reflected sinusoidal signal, are thus necessary. The frequency shift of the received signal with respect to the transmitted frequency bears information on the blood velocity. As the transit time is not measured, the location of the depth at which the velocity is measured cannot be assessed, rather the mean velocity along the line of transmission is measured. An advantage of the CW-Doppler is that the measured velocity range is not limited as is the case with pulsed wave doppler described in the following.

The *pulsed wave Doppler* (PW-Doppler) employs a pulse train as in the A-mode or B-mode imaging. The shift of the pulse frequency of the received pulses is measured and displayed. By appropriate gating of the received signal, only the pressure wave reflected by a selected volume element is considered for determining the frequency. The selection of the volume element is facilitated by the B-mode image, which is simultaneously displayed. Means are also provided to display the temporal run of the blood velocity. The advantage of the PW-Doppler over the CW-Doppler, viz. the selection and display of the volume element in which the flow is measured is offset by the disadvantage that the velocity range of PW apparatus is limited by the pulse repetition frequency of the transmitted pulse. One method often employed to overcome this disadvantage consists in using two pulse repetition frequencies: A lower one for the usual B-mode image and a higher one for the velocity measure. The lower frequency facilitates the imaging of deeper layers and the higher frequency increases the velocity range.

Modern ultrasound scanners are equipped with the necessary hardware and software for the spectral analysis of the Doppler signal using *fast Fourier transform* (FFT) algorithms. Thus the frequency distribution—which corresponds to the velocity distribution—in the sample volume can be determined and, using this data, quantities such as the mean velocity and the pulsatile index can be estimated.

Of particular interest with reference to image processing is the Color-Doppler (alias color flow) mapping of blood velocities in the B-mode image. Here, a color-coded image of the blood velocities is superimposed on the B-mode image. The velocities in the various regions are determined using the PW-Doppler principle. A color scheme often employed is: flow towards the transducer $\hat{=}$ red, flow away from the transducer $\hat{=}$ blue, turbulent flow $\hat{=}$ green; darker shades for lower velocities and brighter shades for the higher velocities. The real-time color Doppler mode is a very useful tool for diagnosis of the state of thyroids, heart, and kidneys.

14.3 Processing the B-mode image

The *B-mode image*, which depicts variations of the acoustic impedance of tissues in the plane of propagation as a visual output of gray intensities, is the most frequently employed form of ultrasound imaging in clinical practice. Exact quantitative mapping of the spatial distribution of acoustic parameters of the tissue is not practical. This is due to the following reasons: The tissues scatter the ultrasound waves by virtue of their structural inhomogeneities with respect to compressibility and density. The acoustic properties of the tissues scattering the pressure wave are space and time dependent and nonisotropic. The correlation lengths that characterize the spatial distribution of the inhomogeneities lie in a very wide range. The exact determination of the spatial distribution of the acoustic characteristics of the tissue from the received signals leads to the solution of an ill-posed inverse problem. Attempts to exploit the information content of the received rf signal in the imaging system by coherent detection have not been successful. Extraction of tissue characteristics using the rf and A-mode signal has attracted a large number of researchers; the results, however, have not found widespread application. Thus, one has to be satisfied with the mapping of the spatial distribution of the acoustic parameters on a qualitative basis. This is the ramification of the practice of depicting organ and tissue properties as B-mode images.

Primarily, it is interpreted on the basis of a visual inspection by the examining physician or radiologist. There is no dearth of efforts to automate the diagnosing process using a computer and image processing software, especially as the image is generally easily accessible for processing at the video output of the imaging system. Computer-based processing of the B-mode image with the aim of characterizing tissues has also caught the interest of a number of research groups. Generally, there are two categories of work:

a. work directed towards improving the visual quality of the B-mode image; and
b. work directed towards feature extraction and texture analysis of the images with a view to correlate image features with properties of tissues. Note that the received RF signal is subjected to envelope is detection and its amplitude is represented as the pixel intensity.

14.3.1 Speckles and artifacts in B-mode image

The texture of the B-mode image is influenced by interference effects. The wavelengths of the ultrasound in biological tissues are in the submillimeter range. The consequence of this wavelength range is that in the B-mode image, not only specular reflections caused by smooth

anatomical surfaces of larger organs appear as bright boundary interfaces, but also fine structures of the parenchymal tissues cause diffuse scattering and appear as areas with granular texture in different shades of gray. The same texture is also caused by scattering occurring due to sound interactions with rough tissue interfaces. The image texture under this condition is often referred to as ultrasound speckle pattern. Most tissues consist of a combination of dominant nonresolvable scatterers (small vessels and microscopic anatomy interfaces) and a smaller component of resolvable (major vessels and connective tissue interfaces) scatterers. When a very large number of backscattered echoes are received simultaneously due to diffuse scattering, the statistics of the amplitude of the demodulated RF signal is described by a Rayleigh probability distribution. The mean and standard deviation variance of this distribution bear a constant ratio, viz. 1.91:1. As this ratio also represents the signal to noise ratio (SNR), when speckle is considered to be noise, the SNR of an image cannot be enhanced by increasing the signal amplitude. Thus, speckle behaves like multiplicative (i.e., signal-dependent) noise. The second-order statistics of a speckle pattern (i.e., the average dimensions of the speckle) can be specified by the dimensions of the 2-D auto-covariance function [3]. This function is largely determined by the parameters of the imaging system and partly by that of the tissue. Algorithms for speckle reduction are based on one of two principles: averaging over the areas of the speckles or replacing a speckle by a representative gray value. For eliminating speckles and at the same time retaining sharp contours non linear algorithms have been proposed, for example, by Greiner et al. [4].

Artifacts in B-mode images are a direct result of the physics of ultrasound imaging. They are basically structures in the B-mode that do not correspond to imaged tissue structures but are caused by a phenomenon related to transmission, scattering, and reflection of acoustic waves.

A common artifact is caused by the boundary between tissues and air, which can be mistaken for the reflection from an organ. The most common artifact is a shadow caused by strongly reflecting tissue boundaries. As a typical example one observes as in Fig. 14.5 that a calcified gall bladder stone is imaged as a crescent moon even though the stone is spherical. Here, most of the energy is reflected back at the surface so that only a small portion of the energy is left to be reflected from inside the sphere.

A second type of artifact is a bright region in the “shadow” of a weakly reflecting boundary caused by regions of high acoustic impedance. This is sometimes referred to as amplification.

Multiple reflections give rise to further artifacts. Here, a region with two parallel strongly reflecting boundaries yields a B-mode image with several parallel bright boundaries.

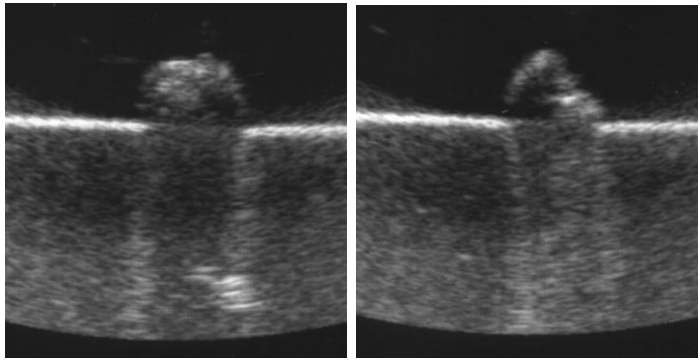


Figure 14.5: *B-mode images of a calcified gall bladder stone.*

14.3.2 Ultrasonic tissue characterization

Ultrasonic *tissue characterization* involves the process of extracting information of the properties of tissues using ultrasound waves. Often, one is interested in registering changes of tissue characteristics as these are more accessible than absolute characteristics. The bases for this are typical signatures of the echo signal or features of the B-mode image. To be useful or significant, such signatures or measures must be related to tissue pathology or physiology. Typical tasks of tissue characterization are:

a. detection of sclerosis of the liver;
b. detection of tumors;
c. determination of the composition of gall bladder stones;
d. detection of tissue changes occurring in renal transplants as a result of immune reactions.

The purely quantitative methods of ultrasonic tissue characterization (not restricted to B-mode imaging) involve the estimation of characteristic quantities of tissues on the basis of a quantitative processing of the transmitted acoustic signal or reflected and backscattered echo (see, e. g., Hill [2], Greenleaf [5], Shung and Thime [6]). A property that is often quoted as a basis for characterization is the frequency dependence of attenuation of ultrasound waves in tissues, which is approximately linear. The attenuation coefficient, which denotes the proportionality between the attenuation and frequency, is a quantitative measure often employed for this purpose. As the ultrasound wave propagates through the tissue, the higher frequency components get weaker at deeper depths. Thus, in an ultrasound imaging system it is theoretically possible to estimate the attenuation coefficient by analyzing the spectra of the received signal at various depths. However, that

this is practically not feasible in a B-mode imaging system should be apparent when one remembers that the signals are subjected to TGC and logarithmic signal compression aimed at obtaining a nice picture rather than in determining the parameters of signals; thus, such quantitative methods are rarely feasible *in vivo*.

A further method proposed for quantitative tissue characterization is the technique of "quantitative imaging." Here, the spatial distribution and strength of a single ultrasound/tissue interaction parameter is mapped quantitatively. By doing this one opens the avenues to tomography. Two major approaches to quantitative imaging are transmission imaging and scatter imaging. These methods are largely experimental and are in the research stage.

B-mode image analysis for tissue characterization. The visual interpretation of B-mode image as a means of tissue characterization is basically a qualitative method of ultrasound tissue characterization. Even though this is the avenue followed almost always in practice, it has the great disadvantage that the diagnosis requires expertise on the part of the radiologist—which is not always available. Also, the diagnosis is often subjective.

The *via media* approach often taken consists in applying computer aided image processing techniques to the B-mode image, in order to obtain quantitative, image-based parameters leading to a more objective interpretation. A technique often dealt with is tissue characterization by image texture analysis. Theoretical studies of the relationship between tissue and its texture in B-mode images have been conducted using scattering analysis. The speckle pattern and dimensions in the image of a tissue are dependent on the ultrasound interrogating beam characteristics and the scattering properties of the tissue. In this context, for theoretical investigations, the tissue is modeled either by discrete scatterers distributed randomly in a medium or by the random variations of the compressibility and density of a continuous inhomogeneity [6]. In the first case, the scatterer density and, in the second, the mean correlation length are employed as quantitative measures. With the former model, it has been established that if the density of scatterers per volume exceeds a certain limit, the speckle is fully developed [7]. This means that the speckle is not affected by further increase in the scatterer density. However, the mean gray level depends on the scatterer density. Generally, using models for deducing tissue properties from the B-mode image has met with only modest success, because the effects of scattering are complicated and not reproducible.

"Parameter imaging" as a means of tissue characterization involves a redisplay or recombination of pulse-echo image information in a way that is more pathology-specific. Here, a parametric image of the original B-mode is generated that maps a local image parameter onto a new

image. The objective followed here is based on the idea that tissue changes show up more distinctly in the parameter image rather than in the original. Parameters that have been proposed are the local variance of the gray level [8] and the local fractal dimension [9]. Although these methods definitely offer an expansive field for research, they seldom lead to a sustaining practical application accepted by a wide circle of practitioners.

One simple method of tissue characterization consists of trying to connect the autocorrelation function of the B-mode image with tissue properties by experiment and empirical means. It is disillusioning but of practical utility to establish that the autocorrelation function usually does not do much worse than more sophisticated methods of analysis of the B-mode image [10].

Quantifying the image features. The goal of quantifying image features is to obtain a set of parameters describing the salient features of the image as they correspond to tissue characteristics. The features of the image one strives to parameterize are first of all those that are implicitly or explicitly assessed by the expert physician to form a diagnosis. In the next stage, additional parameters are sought in order to enhance the reliability of the diagnosis. The following principles summarize the experience acquired from various studies:

a. Not one parameter but a set (vector) of parameters should be used to represent the image features.
b. Fluctuations of tissue characteristics and their images thwart all efforts to find absolute parameters. One has to resort to *changes* in parameters or *relative* parameters.
c. The technique of parameter imaging should be considered to achieve invariance with respect to settings of the apparatus and image scale.
d. The structural stochastic nature of the original ultrasound image as also the parameter image offer the possibilities to parameterize features.
e. By choosing homogeneous areas as regions of interest, one can characterize the region of interest with first-order and second-order statistics of the gray levels. These give reliable results only if the statistics are related to those of a reference area.

A characteristic feature of applied image processing is that the developed methods are specific to the image acquisition hardware. The success of the methods is dependent on the reproducibility of the conditions of imaging. Thus, it is necessary to define the image acquisition and standardizing procedures before examining the image features or developing algorithms for image processing.

14.4 Examples of image processing of B-mode images

In this section, two examples are presented to illustrate the techniques of applying the methods of image processing to process B-mode images. These examples depict typical research activity in the area.

14.4.1 Detection of immune reactions in renal transplants

In certain cases, renal transplants are liable to immunological rejection. Timely detection of these reactions goes a long way towards initiating successful therapy. Various complementary methods are available for the diagnosis of immune reactions. Sonographic examination offering a noninvasive mode of examination is one. Here, an expert physician inspects the B-mode image visually and forms a diagnosis based on his experience. Extrapolating from this situation, efforts were made to develop objective methods to evolve a reliable diagnosis. A knowledge-based classification method that processes a vector of three parameters characterizing the ultrasound image features of the transplant has been proposed in Albert et al. [11], Hoefer [12]. The parameterized image features are chosen first on the basis of the criteria employed by the expert physician for diagnosis and second by correlation of various parameters with the result of diagnosis by other (invasive) means. The criteria used by the expert physician are:

1. An acute reaction is accompanied by *an enlargement of the ultrasound image of the renal transplant*. The geometrical dimensions of the image are measured using standard facilities in modern ultrasound scanners and will not be further elaborated upon here.
2. In a healthy transplant the echogenity of the cortex differs from that of the surrounding tissue. *The lowering of the difference in the mean gray level is an indication of an immune reaction*.
3. As the immune reaction affects both the cortex and the surrounding tissue in the same manner, a reaction is accompanied by *a loss of sharpness of the boundary between the two regions*.
4. A further visual criterion is the texture of the region depicting the cortex. *Whereas a fine structure indicates a healthy renal transplant, a coarse structure indicates immune reactions*.

Image acquisition. The ultrasound images of renal transplants referred to were obtained with an Acuson 128 XP Ultrasound Scanner and a sector transducer (V 328) with a center frequency of 3.5 MHz. The scans were collected in the clinic with a PC and a frame grabber on hard disk and transferred to a Sun workstation equipped with Khoros image processing software. Figure 14.6 shows B-mode scans of a healthy and a sick transplant.

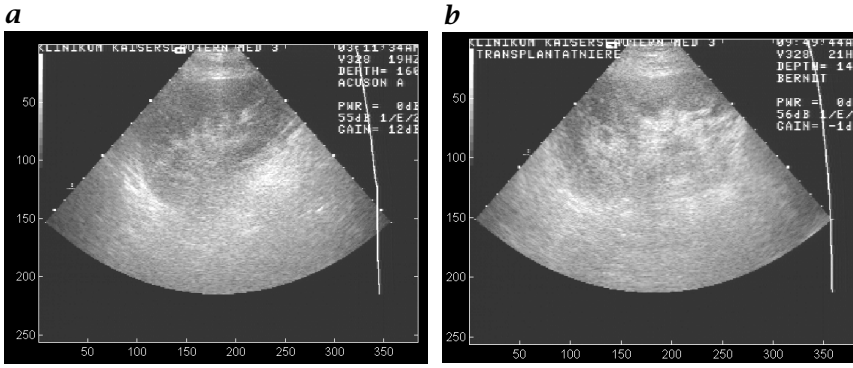


Figure 14.6: *B-mode images of renal transplants:* ***a*** *healthy;* ***b*** *sick.*

Although the renal cortex is more or less homogeneous, a visual inspection of the images reveals that the statistical properties of the image texture vary over the region. This is due to distortions in the imaging with the phased array, the depth dependence of the signal, multiple reflections and anomalies caused by acoustic shadows. To ensure standardization of the imaging conditions and the selection of the region of interest all settings are selected suitably and held constant for the entire session. Heart-beat caused cyclic fluctuations at first sight seem to be avoidable by synchronizing the imaging process with the heart beat. However, actual tests revealed that the advantage gained is marginal. Ultimately it should be noted that the fluctuations cannot be eliminated beyond a certain degree. The rest has to be taken care of by the robustness of the algorithm.

Selection of image features to be parameterized. The image features in the region of interest chosen to represent the state of health of the transplant pertain to the texture of the region surrounding the boundary between the cortex and the surrounding tissue. In accordance with the features taken into account by the physician, the following characteristics are considered:

a. the *variations* of the mean gray level in the neighborhood of the boundary;
b. the distinctness of the boundary; and
c. the variations of the coarseness of the texture in the neighborhood of the boundary.

Whereas characteristic a is examined directly on the basis of the original image, characteristics b and c are investigated with the aid of parameter imaging. The parameter image is obtained by mapping

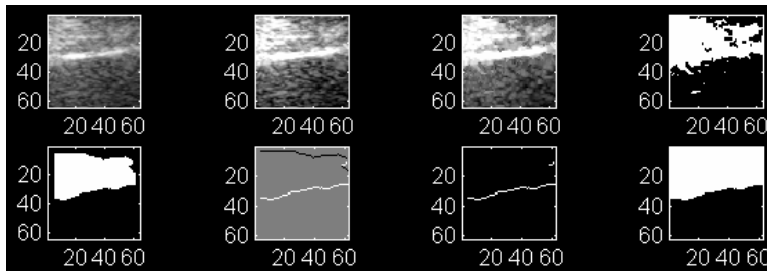


Figure 14.7: *Identifying and marking out the boundary.*

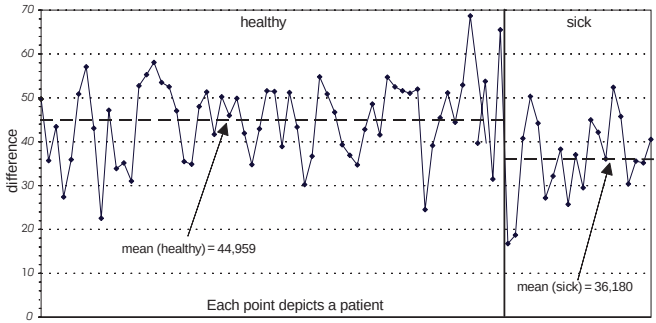


Figure 14.8: *Difference of mean gray values (surrounding tissue—cortex) for a group of patients.*

points of the original image by the Hurst coefficient of local regions surrounding the points.

Parameter of the first-order statistics: difference of mean gray level in original image. First, the boundary region has to be identified and segmented. To do this, starting from the original ROI, a histogram transformation to normalize the gray level is performed. This is followed by a median filtering with a 3×1 window. Next the image is reduced to binary form and subjected to a second median filtering with a 9×9 window. Using a region growing process, the image is segmented. The boundary between the two regions is smoothed and taken as the uncorrupted boundary between the cortex and the surrounding tissue. Figure 14.7 shows the various steps.

The parameter $\boldsymbol{p}_1$, the difference of the mean gray levels on either side of the uncorrupted boundary, is calculated and taken to be the one of the characteristic parameters. Figure 14.8 shows this parameter for a group of 21 patients.

Parametrizing second-order statistics of the image by means of the Hurst coefficient. A parameter often used in signal processing to depict the second-order statistics of a stochastic signal is the width of its autocorrelation function. As this parameter is not independent of the scaling of the image, the fact that the intensity of the points of the image can be approximately represented by a fractional Brownian motion process $B_H[k,l]$ is exploited. In such a case the variance of the difference of intensity at two points $[k,l]$ and $[n,m]$, is given by Hoefer et al. [13] and Hoefer [12]

$$E\{\left|B_H[k,l]-B_H[n,m]\right|^2\} \propto \sqrt{(k-n)^2+(l-m)^2}^{\,H} = \Delta^H \qquad (14.11)$$

where H denotes the Hurst coefficient, a parameter which is independent of scale. To determine H, the mean of the squares of the intensity increments for all pairs of points at the distance Δ from each other for various values Δ in the region of interest is calculated. The ordered set of values for increasing Δ is termed as the Brownian feature vector. Then, the Hurst coefficient H, which is given by the slope of the Brownian feature vector, completely specifies the second-order statistics of the fractional Brownian process.

Parameter imaging with the Hurst coefficient. The parameter image of the region of interest with the local Hurst coefficient as parameter offers a scale-independent representation of texture [9]. Figure 14.9 shows the original (a,b) and the parameter images (c,d) of two of the regions of interest presented earlier. An abrupt transition in the original image yields a transition area that consists of a bright strip in the parameter image. The boundary area is marked off with two regression straight lines enclosing the strip in the parameter image.

The parameter p_2 is the Hurst coefficient of the original image in the boundary region is a measure of the abruptness of the transition and is chosen as a further parameter for classifying an image.

An alternative for quantifying the abruptness of the transition is based on the Brownian feature vector of *the parameter image.* The maxima of the increments in the two directions represent parameters for the quality of the transition region. The parameter p_3 is the product of the maxima is taken as a further measure for classifying the images.

A knowledge-based learning classifier as diagnosis aid. The image parameters specified in the preceding are incorporated in a learning classifier. First, a knowledge base is set up. This consists in locating the normalized representant parameter vectors $\boldsymbol{p}_H = [p_{H1}, p_{H2}, p_{H3}]^T$ and $\boldsymbol{p}_S = [p_{S1}, p_{S2}, p_{S3}]^T$ of the classes of vectors corresponding to the healthy and sick cases and the spread of the parameters about the representants. The representants are formed using the components $\boldsymbol{p}_{Hi}$

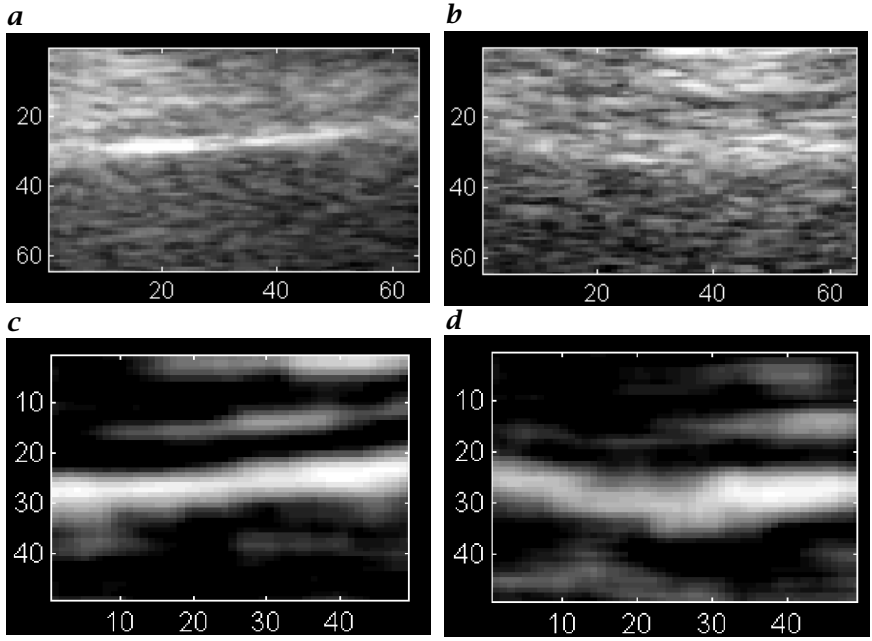


Figure 14.9: *Original (**a** and **b**) and parameter images (**c** and **d**)of the regions of interest.*

and $\boldsymbol{p}_{Si}$ obtained by averaging over the available data of n_H healthy n_S sick transplants, respectively. The spread of the parameters is used as weighting indices $\boldsymbol{w}_{Hi}$ and $\boldsymbol{w}_{Si}$. This is the knowledge base with which a diagnosis is formed for a measured parameter vector $\boldsymbol{p} = [p_1, p_2, p_3]$. The distances d_h and d_s of the parameter vector under consideration from the representant of the classes corresponding to the healthy and sick transplants, respectively, are calculated. The class at a shorter distance is selected.

Results obtained in clinical tests. The parameters of the ultrasound scan of 21 patients were calculated and plotted. Figure 14.10 shows distance difference $\boldsymbol{d}_S\boldsymbol{d}_H$ for each of the image parameter vectors. Ideally the transplants corresponding to the distance differences below 0 should be classified as healthy and those corresponding to distance differences greater than 0 should be classified as sick.

14.4.2 Determination of composition of gall bladder stones

The determination of the composition of gall bladder stones on the basis of (noninvasive) ultrasound scans has been a subject of intense research. One good reason for this is that it is advantageous to select

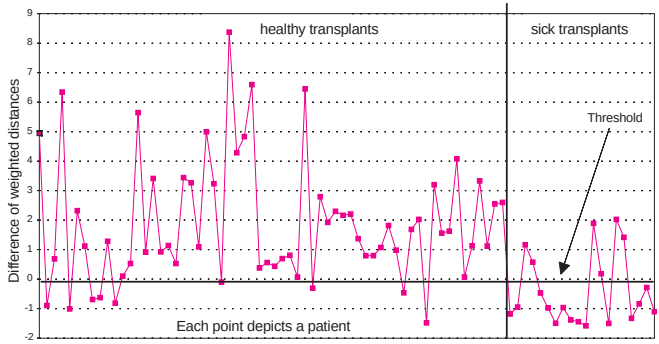


Figure 14.10: *Difference of distance from representants.*

therapy on the basis of the composition. For example, detection of calcification is especially important in order not to employ lithotripsy, as the pulverized calcium compounds cannot be dissolved and washed away with solvent medicines. Opinion is divided as to whether a reliable method of composition determination is possible at all with ultrasound, especially on the basis of in vivo (and, therefore, also of in vitro) B-scans. In Kocherscheidt et al. [14], the authors undertook to check methods published in the literature and to devise and try out new methods for detection of calcification in gall bladder stones. To make the investigation as comprehensive as possible, both the high frequency signal and the B-scans were examined in vitro, in the latter case in vivo as well. Finally, calcification was established using x-ray imaging in order to check the correctness of the diagnosis.

Signal acquisition was carried out in three phases: (a) The stones were suspended in a water bath and the RF echo signal was sampled; (b) the in vivo and in vitro B-scans were obtained via the video output; and (c) the stones were checked for calcification using x-rays.

The following classification procedures were applied:

1. RF signal analysis: The group delay was calculated from short-term FFTs of the sampled signal. It was used because of its relation to the sound velocity, which has been claimed to be an indicator of calcification in the literature.
2. B-scan analysis: (a) Using a thresholding technique and a combination of morphological and labeling methods, the area consisting of the front and rear echoes of the stone was segmented in the steps shown in Fig. 14.11. (b) Several texture based features such as those using the co-ocurrence matrix and the auto correlation function were calculated for the segmented regions. Their classification abilities were tested with regard to calcification. (c) As suggested in the literature, the visible area of the stones was divided into two sub-

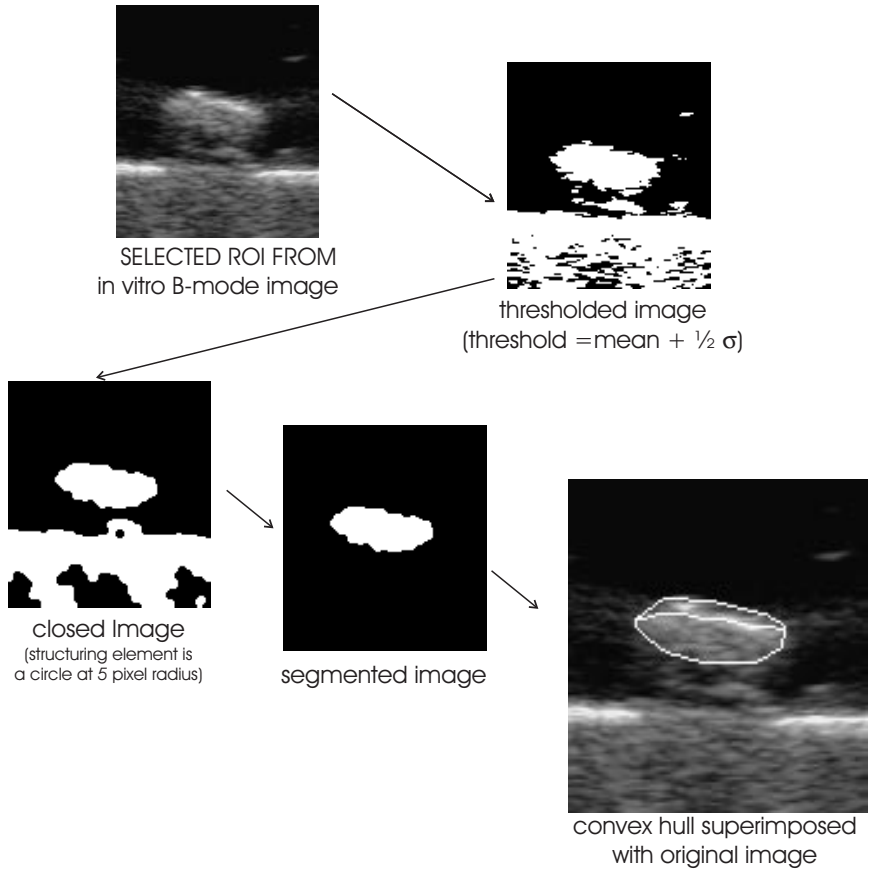


Figure 14.11: *Filtering and segmentation of the gallbladder stone image.*

areas corresponding to the front and the main echoes (the shadows were not considered). Mean and standard deviation were calculated for the two subareas and were analyzed with respect to each other, expecting a kind of clustering related to the calcification.

(a) The group delays obtained for the calcified and uncalcified stones lay within the same range of values and therefore allowed no classification. The same was found by inspection for the variations in one and the same stone. (b) The evaluation of the texture parameters of the segmented areas did not show any calcification dependent clustering. (c) Features proposed in the literature, for example relative intensity of front and main echoes, also failed as indicators of calcification. However, the described features show some correlation with the surface geometry rather than the composition.

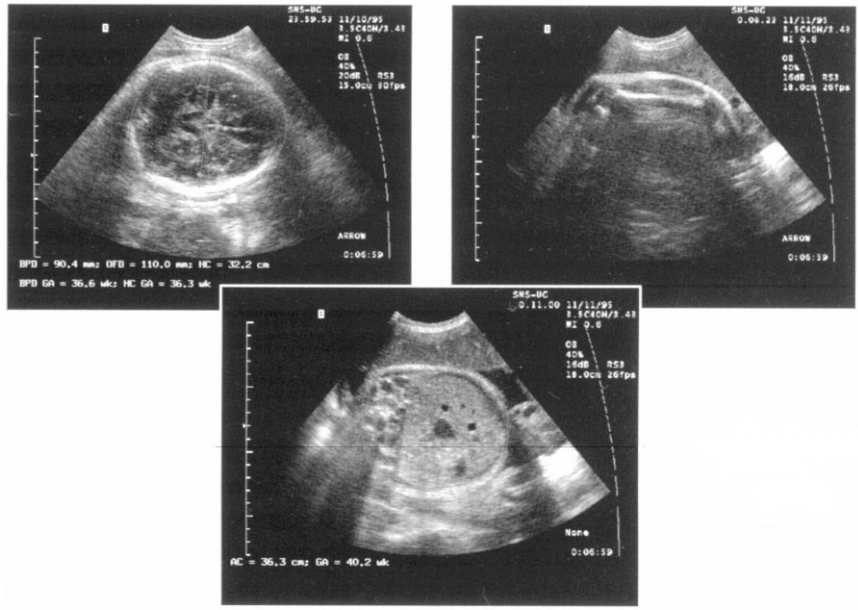


Figure 14.12: *Image processing for measuring organ dimensions.*

14.5 Conclusions and perspectives

Processing ultrasound B-mode images offers scope for applying methods of image processing for improving the quality of the image for characterization of tissues. The latter is still in the research stage or is applied in individual clinics. Recent and current developments show the trend of more sophisticated processing algorithms for the post-processing of image signals.

Fast and high-resolution digital microelectronics have led to commercially viable ultrasound imaging systems in which the generation, acquisition and processing of the rf and envelope signals and the "raw" B-mode image have been perfected. Consequently, manufacturers have gone on to develop and apply advanced algorithms for the post-processing of the images in real-time in order to enhance the utility of the systems. For real-time implementation, specially developed image processors and multimedia video-processors are employed. Applications that are cited are refined algorithms for edge detection and segmentation as a precursor to measuring dimensions foetal cranium. The process is interactive: The examining gynecologist sets markings with a cursor in the neighborhood of the edge that has to be filtered and activates the algorithm. The algorithm constructs by filtering and extrapolation a continuous line as the edge. The result of such a construction is

shown in Fig. 14.12. A second application involves an adaptive filter for filtering objects in motion. The adaptive filtering algorithms permit the averaging of images to reduce speckle and noise without blurring the edges. This is made possible by a two-stage process: In the first stage, motion is estimated by comparing areas and in the second stage, filtering is performed. A current development concerns the enlargement of the field of vision. To extend the field a composite ("panorama") image is formed from several individual images obtained by moving the transducer and coupling the movement with the imaging. Image processing algorithms are employed to estimate the position coordinates of the transducer from the image itself, thereby eliminating the need for an additional mechanical position sensor.

Acknowledgments

The authors acknowledge with thanks the support and courtesy of Siemens Medizintechnik Erlangen, in particular Dipl.-Ing. Gert Hetzel for information and the figures in Sections 14.2 and 14.5 and Plenum Press Publishers for the figures in Section 14.4. Thanks are due to Dr. M. Insana of the University of Kansas Medical Center for useful suggestions. The interest and help of present and former colleagues and students of the department, especially Dr.-Ing. Thomas Greiner, Dr.-Ing. Stefan Hoefer, and Heiko Hengen were instrumental in producing the manuscript.

14.6 References

[1] Cho, Z., Jones, P. J., and Singh, M., (1993). *Foundations of Medical Imaging*. New York: Wiley.

[2] Hill, C. R. (ed.), (1986). *Physical Principles of Medical Ultrasonics*. Chichester: Ellis Horwood/Wiley.

[3] Wagner, R., Smith, S., Sandrik, J., and Lopez, H., (1983). Statistics of speckle in ultrasound B-Scan. *IEEE*, **SU-30**(3):156–164.

[4] Greiner, T., Loizou, C., Pandit, M., Maurschat, J., and Albert, F. W., (1991). Speckle reduction in ultrasonic imaging for medical applications. In *Proc. International Conference on Acoustics, Speech and Signal Processing*. Toronto, Canada.

[5] Greenleaf, J. E., (1986). *Tissue Characterization with Ultrasound*. Boca Raton, USA: CRC Press.

[6] Shung, K. K. and Thime, G. A. (eds.), (1993). *Ultrasonic Scattering in Biological Tissues*. Boca Raton, FL: CRC Press.

[7] Thijssen, J. M., (1989). Ultrasonic tissue characterization and echographic imaging. *Phys. Med. Biol.*, **34**(11):1667–1674.

[8] Verhoeven, J. T. M., Thijssen, J. M., and Theuwes, A. G. M., (1991). Improvement of lesion detection by echographic image processing: signal to noise imaging. *Ultrasonic Imaging*, **13**:238–2516.

[9] Chen, C. C., Daponte, J. S., and Fox, M. D., (1989). Fractal feature analysis and classification in medical imaging. *IEEE Trans. on Medical Imaging*, **8** (2):133–142.

[10] Greiner, T., (1994). *Methoden der digitalen Bildverarbeitung zur computergestützten Gewebecharakterisierung mit Ultraschall unter besonderer Berücksichtigung der hierarchischen Texturanalyse*. Aachen: Shaker-Verlag.

[11] Albert, W., F., Berndt, N., Schmidt, U., Hoefer, S., Keuser, D., and Pandit, M., (1996). Detection of immune reactions in renal transplants based on ultrasonic image analysis. In *Acoustic Imaging*, pp. 251–256. New York: Plenum Press.

[12] Hoefer, S., (1995). *Statistische Analyse medizinischer Ultraschallbilder u. Berücksichtigung der physikalischen Effekte der Ultraschallbildgebung sowie der fraktalen Struktur der abgebildeten Organe*. Aachen: Shaker-Verlag.

[13] Hoefer, S., Hannachi, H., Pandit, M., and Kumaresan, R., (1992). Isotropic two-dimensional fractional Brownian motion and its application in ultrasonic analysis. In *Conference of the IEEE Engineering in Medicine and Biology Society, Paris*. New York: IEEE.

[14] Kocherscheidt, C., Schmidt, U., Albert, W., Racky, J., M., P., and Pandit, M., (1997). Determination of gallbladder stone composition with ultrasound — in vivo and in vitro studies. In *23rd Conference on Acoustic Imaging, Boston*. New York: Plenum Press.

15 Acoustic Daylight Imaging in the Ocean

Michael J. Buckingham

Scripps Institution of Oceanography, La Jolla, CA, USA

15.1 Introduction

Seawater is essentially transparent to sound and opaque to all other forms of radiation, including light. Acoustic techniques are thus a preferred choice for probing the ocean depths. Two types of acoustic systems, passive sonar and active sonar, are commonly used as detection devices in the ocean [1]. A passive sonar simply listens for the sound radiated by a target such as a submarine, whereas an active sonar transmits an acoustic pulse and listens for returning echoes from objects of interest. Apart from their military applications, active sonars are used in various configurations for numerous purposes, from simple echo sounding to mapping the seafloor.

Besides the signals of interest, both passive and active sonars respond to ambient noise in the ocean, which tends to degrade the performance of such systems. The noise is generated partly by natural sources including wave breaking [2], precipitation [3], and, of course, a variety of biological organisms ranging from marine mammals [4, 5, 6, 7] to snapping shrimp [8, 9] and croakers [10]; and also by anthropogenic sources, notably surface shipping, and offshore exploration and drilling for hydrocarbons. Considerable effort has been devoted to suppressing the effects of the noise on active and passive sonars, with a view to enhancing the signal-to-noise ratio at the output of the system.

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

Acoustic noise propagates through the ocean in much the same way that light propagates through the atmosphere, and in so doing acquires information about objects it encounters in its path. This suggests that an alternative to noise suppression is noise exploitation, in which information, in the form of an image of the environment, is obtained from the ensonification provided by the noise field. A reasonable optical analog is conventional photography in the atmosphere using daylight as the illumination; a photographic image of an object can be obtained, even though the object itself radiates no light and no artificial source of light, such as a photoflood or flashgun, is employed to illuminate the scene. Evidently, photography with daylight is neither "passive" nor "active," but instead relies on the ambient light scattered from the object space to create the image.

Beneath the sea surface the *ambient noise* field provides a form of "acoustic daylight." The noise is a stochastic radiation field with energy traveling in all directions, and as such has much in common with the diffuse ambient light field filtering through a cloud layer. This suggests that the noise has the potential for *acoustic imaging*, although the highest usable frequency is expected to be around 100 kHz. (At higher frequencies the propagating noise is likely to be masked by the localized thermal noise [11, 12] of the water molecules, which carries no imaging information.) The wavelength of sound in seawater at a frequency of 100 kHz is 1.5 cm, which is orders of magnitude greater than optical wavelengths, implying that the resolution of the images obtained from the ambient noise field will not match that of the pictures from a photographic camera.

The first issue to address, however, is not the quality of acoustic daylight images but whether such images can be formed at all. In this chapter, some early experiments on scattering of ambient noise from targets on the seabed are briefly discussed, the first acoustic daylight imaging system is described, and some results from deployments of the system in the ocean off southern California are presented. These images clearly demonstrate that, in certain ambient noise environments, it is possible to create stable, recognizable acoustic daylight images of targets at ranges out to 40 m.

15.2 The pilot experiment

In 1991, a simple experiment [13] was conducted off Scripps pier, southern California, where the nominal water depth is 7 m. The experiment was designed to determine whether ambient noise scattered by an object in the ocean is detectable using a low-noise, directional acoustic receiver. The targets were three rectangular ($0.9 \times 0.77\,\mathrm{m}^2$) wooden panels faced with closed-cell neoprene foam, which, being loaded with

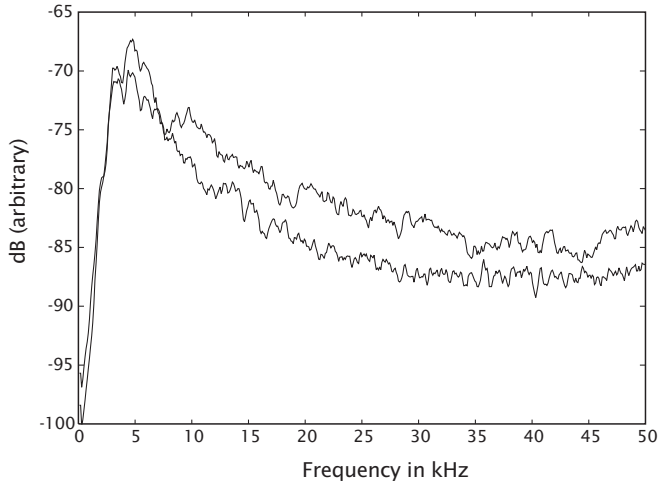


Figure 15.1: *Noise spectra obtained in the pilot experiment with the targets "on" (upper trace) and "off" (lower trace).*

air, is an efficient acoustic reflector. A single hydrophone mounted at the focus of a parabolic reflector of diameter 1.22 m was used as the directional acoustic receiver. The frequency response of the system extended from 5 to 50 kHz, and at the highest frequency the beamwidth was $\approx 3.6^{\circ}$.

Divers deployed the targets and the receiver on the seabed, separated by a horizontal distance of 7 m. The target panels were placed vertically and could be rotated about a vertical axis, allowing them to be oriented either normal to or parallel to the line of sight, the "on" and "off" positions, respectively. In the "on" configuration, the three target panels formed, in effect, a single rectangular target of height 0.9 m and width 2.5 m. At a frequency of 9 kHz, the angular width of the main lobe of the receiver, as measured at the −6 dB points, matched the angle subtended by this extended target, and hence at higher frequencies the angular width of the target was greater than that of the beam.

Ambient noise data were collected sequentially with the target panels in the "on" and "off" positions. Typically, it took 15 to 30 min for the divers to rotate the panels from one configuration to the other. Figure 15.1 shows an example of the noise spectra that were obtained in the experiment. Over the frequency band of interest, it can be seen that the spectral level is higher by some 3 dB with the panels in the "on" position, a result that was found consistently throughout the experiment. The excess noise level observed with the panels "on" suggests that the noise originated behind the receiver, a situation analogous to taking a photograph with the sun behind the camera. In fact, a subsequent ex-

periment [14] established that the dominant noise sources were snapping shrimp located around the pilings of the pier, which was indeed immediately behind the acoustic receiver.

With most of the noise sources located behind the receiver, little if any noise radiation was blocked by the target but a significant portion was scattered back towards the dish. Evidently, the presence of the panels modified the observed noise intensity significantly. Such behavior is potentially important, as it provides the basis for incoherent imaging with ambient noise. In effect, what had been achieved with the single-beam receiver was just one pixel of an image. Based on this result from the pilot experiment, it seemed possible that, by increasing the number of beams and mapping the noise intensity in each into a pixel on a computer monitor, a visually recognizable image of the target space could be achieved. The objective of the next stage of the acoustic daylight imaging project [15] was to design and build a multibeam, broadband acoustic receiver with the capability of resolving meter-sized objects at ranges of about 50 m.

15.3 ADONIS

In designing the multibeam acoustic receiver, it was decided that a minimum of about 100 pixels would be needed to produce an acoustic daylight image. The question then arose as to what type of detector should be used for the purpose? A possible design solution to the problem would have been a phased array with an aperture in the region of 3 m. This approach, however, was not pursued for several practical reasons, one of which was the heavy computational load imposed by the beamforming. Instead, a reflector technology was again employed, this time in the form of the *Acoustic Daylight Ocean Noise Imaging System* (ADONIS). The design of ADONIS is very much more complex than the parabolic system used in the pilot acoustic daylight experiment. Before it was constructed, the imaging performance of ADONIS was simulated using a numerical algorithm based on the Helmholtz-Kirchhoff scattering integral [16]; and the effects of noise directionality on the acoustic contrast between pixels were examined in a theoretical analysis of a multibeam imaging system [17].

ADONIS consists of a spherical dish faced with neoprene foam, with a diameter of 3 m and a radius of curvature also of 3 m. An array of 130 ceramic *hydrophones* arranged in an elliptical configuration (Fig. 15.2) occupies the focal region of the dish. Each sensor is of square cross section and the center-to-center distance between sensors is 2 cm. The major and minor axes of the array are approximately 0.28 m and 0.22 m, giving an approximate field of view of 6° horizontally and 5° vertically.

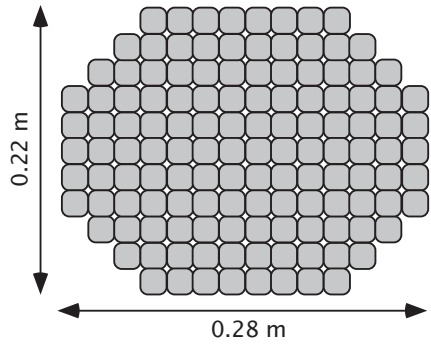


Figure 15.2: *Schematic of the 130-element elliptical array head.*

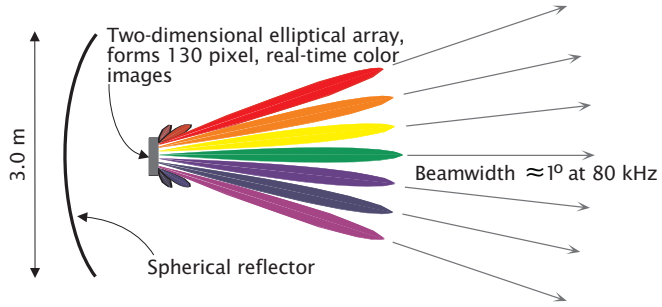


Figure 15.3: *Schematic showing the spherical reflector, array head, and fan of beams; (see also Plate 9).*

A spherical dish was chosen because the aberrations associated with the off-axis receivers are less severe than would be the case with a parabolic reflector. An advantage of the reflector design is that the *beamforming* is performed geometrically, with each hydrophone, by virtue of its position relative to the center of the dish, having its own unique "look" direction. This feature reduces the computational load significantly, as beamforming by the application of phase or time delays to the individual receiver elements is unnecessary. Because the array head contains 130 hydrophones, ADONIS forms a fan of 130 beams (Fig. 15.3) distributed in the horizontal and the vertical. The system operates in the frequency band from 8 to 80 kHz, and at the highest frequency the beamwidth is slightly less than 1°.

To produce the images, the signal in each channel is processed in real time through a custom-built, hybrid analog-digital electronics package. The analog circuitry performs a *spectral analysis* by sweeping a *bandpass filter* with a Q of 4 across the decade of frequency occupied by the signal, and the noise intensity is recorded at sixteen frequency

points uniformly distributed logarithmically through the band. This function is repeated 25 times per second by a switched capacitor filter in each of the 130 channels. The spectral data are then digitized and displayed on a computer monitor as moving, color images, each consisting of 130 pixels. Various signal conditioning options have been developed to improve the quality of the raw images, including interpolation, time-averaging and normalization.

Color in the acoustic daylight images is used in two ways. In most cases, color simply represents the intensity level of the signal in a particular beam, or pixel, averaged over some or all of the sixteen frequency cells. Because the sharpest images are obtained at the highest frequencies, where the beams are narrowest, we have tended to concentrate on the top three frequency cells, covering the band from 58 to 77 kHz.

However, when the lower frequencies are neglected, the broadband capability of ADONIS is not fully exploited. On occasion, the full bandwidth of the system can be used to advantage to yield information about a target, even though spatial resolution may be sacrificed by including lower-frequency data in the image. For instance, the spectral shape of the scattered field from various targets in the object space may differ because of their composition or surface properties, that is to say, they show different acoustic "color." Such targets could, in principle at least, be distinguished from one another through the differences in their scattered spectra. These differences have been visualized in the acoustic daylight images through the second way of using color. By assigning a color to a frequency cell and weighting each color component with the intensity in the cell, the color in the resultant image provides a visual indication of the acoustic and physical properties of the objects in the image. A hollow sphere, for example, may appear as red whereas a solid sphere could be blue. In fact, target recognition through a mapping of the *acoustic color* of an object into visual color in the broadband acoustic daylight image has been demonstrated successfully with certain simple targets of similar shape but different composition.

15.4 Acoustic daylight images

The construction of ADONIS was completed in August 1994. Since then, two major series of acoustic daylight imaging experiments, known as ORB 1 and ORB 2, have been performed with ADONIS mounted on the seabed in San Diego Bay, southern California. In this location, the ambient noise field in the 8 to 80 kHz band is dominated by the sound from colonies of snapping shrimp. These creatures, which are about the size of a thumbnail, generate a random series of disproportionately loud pulses of sound. The duration of each pulse is less than

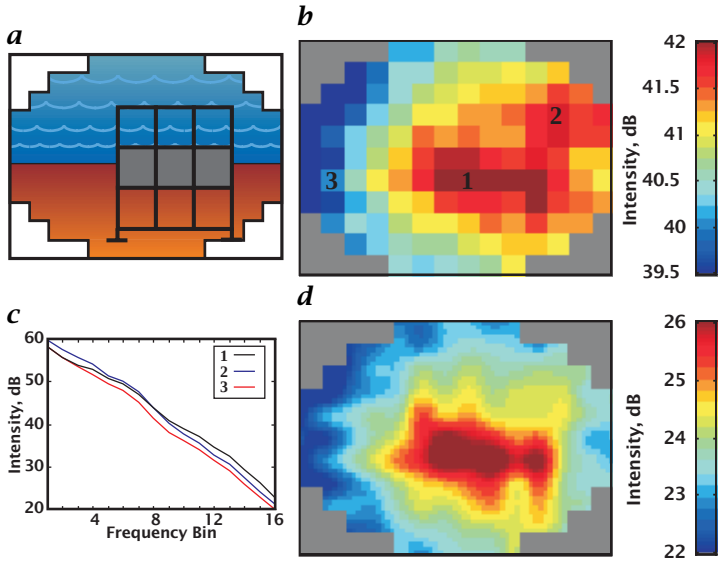


Figure 15.4: ***a*** *Simulated view of the rectangular bar target mounted vertically on the seabed;* ***b*** *acoustic daylight image of the bar target from raw intensity data;* ***c*** *spectra of the pixels labeled 1, 2, and 3 in* ***b****;* ***d*** *interpolated acoustic daylight image of bar target from the same data used in* ***b****. For movies of the same and other targets see* `/movies/15` *on the CD-ROM (*`BarTarget`*,* `BotAirDrum`*,* `BotDrum`*,* `BotSandDrum`*,* `HngDrums`*,* `Sphere`*, and* `Credits`*); (see also Plate 10).*

10 μs and the bandwidth is in excess of 100 kHz. Clearly, such a noise environment is not temporally stationary and in fact gives rise to stability problems in the acoustic daylight images. An effective though not necessarily optimal solution to the difficulty is temporal averaging, which, when performed over about 1 s, corresponding to 25 frames, produces reasonably stable images.

Several different types of targets were used in the ORB experiments, including planar aluminum panels faced with closed-cell neoprene foam, corrugated steel, or a number of other materials. Volumetric targets in the form of 113-liter polyethylene drums, 0.76 m high by 0.5 m diameter, and filled with syntactic foam (essentially air), wet sand or sea-water, have been imaged both in mid-water column and partially buried in the seabed, which consists of a fine-grained silt. A hollow titanium sphere, 70 cm in diameter, has also been imaged in mid-water column. A more mobile target, in the form of a diver with closed breathing system, has been recorded in a series of acoustic daylight images as he swam through the field of view of ADONIS. In most of these imaging experiments, the range between ADONIS and the targets was nominally 40 m.

Figure 15.4 shows two versions of an acoustic daylight image of a horizontal, bar-like target, 1 m high by 3 m wide, and formed from three square aluminum panels faced with neoprene foam. The panels were mounted on a target frame, as illustrated schematically in Fig. 15.4a, and the basic image formed from the raw intensity data is shown in Fig. 15.4b. Each of the square pixels shows the noise intensity in one of the ADONIS beams, averaged over the top three frequency bins (58 – 77 kHz). The noise spectra of the pixels labeled 1, 2, and 3 are shown in Fig. 15.4c, where it can be seen that the background noise level is about 3 dB below the on-target level across the whole frequency band of the system. Figure 15.4d, showing a post-processed version of the same data as in Fig. 15.4b, illustrates the visual improvement over the raw image that can be obtained by interpolation between pixels, in this case by a factor of five.

The acoustic daylight image of the bar target in Fig. 15.4, like many of the images obtained in the ORB experiments, is strongly ensonified from the front. Most of the snapping shrimp sources were located on pier pilings behind ADONIS, which gave rise to a very directional noise field. The strong directionality accounts for the relatively high acoustic contrast between the on-target and off-target regions in the image. Occasionally, a passing vessel ensonified the targets from behind, in which case a silhouette effect was observed in the images. This and other effects appearing in the images are discussed in some detail by Epifanio [18], who also presents acoustic daylight images of all the targets used in the ORB experiments, including the swimming diver.

15.5 Concluding remarks

ADONIS and the ORB experiments have served to demonstrate that ambient noise can be used for incoherent imaging of objects in the ocean. The acoustic contrast achieved in these experiments was typically 3 dB, which is sufficient to produce a recognizable visual realization of the object space. Moreover, ADONIS is unique among underwater acoustic systems in that it produces images continuously in real time at a frame rate of 25 Hz. When shown as a movie, the images show fluid movement and, with appropriate interpolation, averaging and normalization, the objects depicted appear reasonably clearly. Nevertheless, each image contains only 130 pixels and consequently appears rather coarse by optical standards, but it should be borne in mind that ADONIS is a prototype and that more channels could be incorporated into future systems.

Of course, an increase in the number of pixels would represent just one contribution towards improved image quality. Another important factor is the angular resolution that can be achieved, which is governed

by the aperture of the detector measured in wavelengths. In many applications, it would not be practical to extend the physical aperture beyond 3 m, as used in ADONIS. An alternative is to operate at higher frequencies, which may be feasible in areas where snapping shrimp are the main contributors to the noise field, as appears to be the case in most temperate and tropical coastal waters [19, 20]. Signal processing is another means of enhancing image quality. The current acoustic daylight images, such as those in Fig. 15.4, are created by applying simple operations to the noise intensity in each channel. It may be possible to do better, at least in some circumstances, by turning to higher-order statistics; Kalman filtering would appear to have advantages for tracking moving objects [21].

Acknowledgments

Broderick Berkhout generated the spectral data shown in Fig. 15.1 and Chad Epifanio produced the images in Fig. 15.4. John Potter was central to the design and construction of ADONIS. The acoustic daylight research described here is supported by the Office of Naval Research under contract number N00014-93-1-0054.

15.6 References

[1] Urick, R., (1983). *Principles of Underwater Sound*. New York: McGraw-Hill.

[2] Kerman, B., (1988). *Sea Surface Sound: Natural Mechanisms of Surface Generated Noise in the Ocean*. Dordrecht: Kluwer.

[3] Nystuen, J. and Medwin, H., (1995). Underwater sound produced by rainfall: Secondary splashes of aerosols. *J. Acoust. Soc. Am.*, **97**:1606–1613.

[4] Watkins, W. and Schevill, W., (1977). Sperm whale codas. *J. Acoust. Soc. Am.*, **62**:1485–1490.

[5] Watkins, W., Tyack, P., Moore, K. E., and Bird, J. E., (1987). The 20-Hz signals of finback whales (Balaenoptera physalus). *J. Acoust. Soc. Am.*, **82**:1901–1912.

[6] Cato, D., (1991). Songs of humpback whales: the Australian perspective. *Memoirs of the Queensland Museum*, **30**:277–290.

[7] Cato, D., (1992). The biological contribution to the ambient noise in waters near Australia. *Acoustics Australia*, **20**:76–80.

[8] Readhead, M., (1997). Snapping shrimp noise near Gladstone, Queensland. *J. Acoust. Soc. Am.*, **101**:1718–1722.

[9] Cato, D. and Bell, M., (1992). Ultrasonic ambient noise in Australian shallow waters at frequencies up to 200 kHz, DSTO Materials Research Laboratory, Sydney, Report No. MRL-TR-91-23.

[10] Kelly, L., Kewley, D., and Burgess, A., (1985). A biological chorus in deep water northwest of Australia. *J. Acoust. Soc. Am.*, **77**:508–511.

[11] Callen, H. and Welton, T., (1951). Irreversibility and generalized noise. *J. Acoust. Soc. Am.*, **83**:34–40.

[12] Mellen, R., (1952). The thermal-noise limit in the detection of underwater acoustic signals. *J. Acoust. Soc. Am.*, **24**:478–480.

[13] Buckingham, M., Berkhout, B., and Glegg, S., (1992). Imaging the ocean with ambient noise. *Nature*, **356**:327–329.

[14] Buckingham, M. and Potter, J., (1994). Observation, theory and simulation of anisotropy in oceanic ambient noise fields and its relevance to Acoustic Daylight™ imaging. In *Sea Surface Sound '94*, pp. 65–73. Lake Arrowhead, CA: World Scientific.

[15] Buckingham, M., Potter, J., and Epifanio, C., (1996). Seeing underwater with background noise. *Scientific American*, **274**:86–90.

[16] Potter, J., (1994). Acoustic imaging using ambient noise: Some theory and simulation results. *J. Acoust. Soc. Am.*, **95**:21–33.

[17] Buckingham, M., (1993). Theory of acoustic imaging in the ocean with ambient noise. *J. Comp. Acoust.*, **1**:117–140.

[18] Epifanio, C., (1997). *Acoustic Daylight: Passive Acoustic Imaging using Ambient Noise*. PhD thesis, Scripps Institution of Oceanography, University of California, San Diego, CA.

[19] Chitre, M. and Potter, J., (1997). Ambient noise imaging simulation using Gaussian beams. *J. Acoust. Soc. Am.*, **102(5/2)**:3104.

[20] Potter, J. and Chitre, M., (1996). Statistical models for ambient noise imaging in temperate and tropical waters. *J. Acoust. Soc. Am.*, **100(4/2)**:2738–2739.

[21] Potter, J. and Chitre, M., (1996). ADONIS imaging with a Kalman filter and high-order statistics. In *Proc. Third European Conference on Underwater Acoustics*, pp. 349–354, Heraklion: Crete University Press.

16 The Multisensorial Camera for Industrial Vision Applications

Robert Massen

Massen Machine Vision Systems GmbH, Konstanz, Germany

16.1 Image segmentation with little robustness

One of the most common problems in actual machine vision systems is their lack of *robustness* against small changes in the scene. The poor success of symbolic image processing technologies in industrial applications is mainly due to this lack of robustness; extracted symbols like corners, polygons, etc., tend to fail completely if the iconic preprocessing such as *edge detection* does not work in a proper and reliable way. Traditionally, there have been three attempts to alleviate this uncomfortable situation:

1. Use more intelligent algorithms based on more *a priori knowledge*. This works good on a well-known test scene but no better in many real industrial applications where in general little reliable a priori knowledge is available and many unexpected situations are common;
2. Restrict use to very simple pixel-based algorithms such as blob-analysis, projections, etc. This is the day-to-day approach for most

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

actual industrial vision systems. This reduction to simplicity, however, is also the reason why many important industrial vision problems are actually only poorly solved:

- the inspection of *structured surfaces* such as natural wood is still difficult and of unsatisfactory quality; and
- the inspection of metallic and shiny nonflat and complexly patterned surfaces is often just impossible; and

3. Use several imaging and nonimaging sensors and fuse the information contained in their signals (*sensor fusion*). This is a good approach but of rather high complexity; the different sensors are mostly located at different positions of the scene, and they have different temporal, spatial and radiometric resolutions that make the registration of these signals difficult.

A new concept is discussed that like sensor fusion follows the principle of extracting more information from the scene and that takes full advantage of the cross information and synergies in the signals of imaging sensors that are sensitive to different, preferrably uncorrelated physical features of the imaged objects.

16.2 Sensor fusion and multisensorial camera

Sensor fusion in image processing is usually understood as the combination of several imaging or nonimaging sensors that observe one scene. Here, we take a real-life problem to explain the following ideas, the inspection of timber wood for defects.

To inspect fast moving wood timbers in a production line for the presence of knots, cracks, resin pockets, geometric defects, moisture and hidden metallic contamination (nails, etc.), it is clear that this cannot be achieved by using solely a color camera system. We will possibly combine a color camera with a 3-D line-section system, an x-ray camera and an infrared humidity sensor (Fig. 16.1).

All these noncontact sensors have their own specific spatial resolution (number of pixels, shape of the field of view), temporal resolution (frame rate), and radiometric resolution (number of bits limited by the signal-to-noise ratio). In general, it will be difficult to combine these different sensors in such a way that they observe exactly the same location of the bypassing timber at the same time. If we arrange the sensors along the direction of motion of the timber, we cannot correct for the introduced time delays between the different sensor signals if we do not exactly know the velocity of the timber. In real life, this velocity is neither constant nor precisely measurable. The timber may accelerate and decelerate quite brutally, and it may oscillate in space; all this makes signal registration with pixel accuracy impossible.

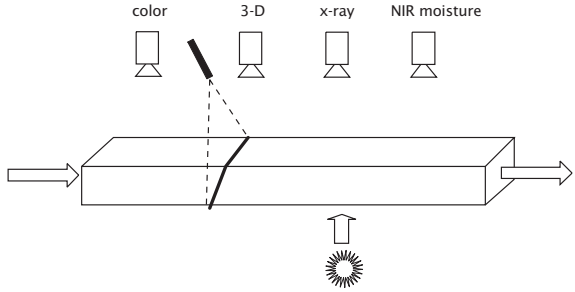


Figure 16.1: *Inspection of timber wood for color and 3-D defects for hidden metallic parts and for local moisture using traditional sensor fusion. The different sensors are not in registration.*

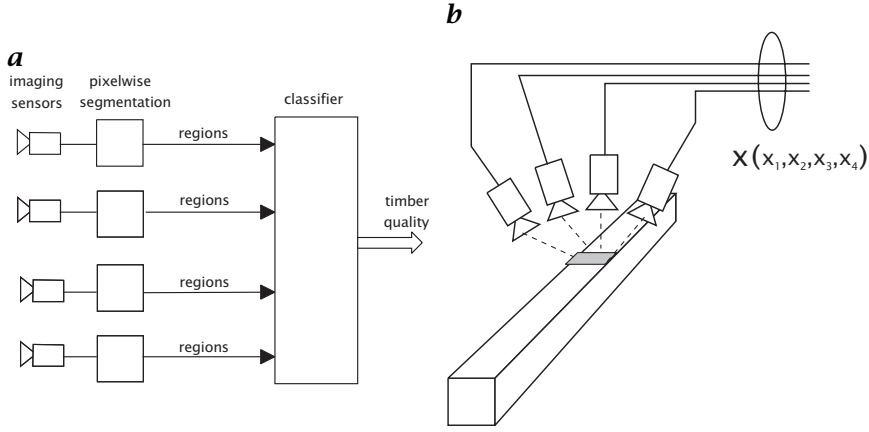


Figure 16.2: ***a*** *The pixel images of the not-registered sensors have to be segmented individually. The fusion of information can only take place at the level of regions where pixelwise registration is not required;* ***b*** *the multisensorial camera combines different imaging sensors that look at the same spot, all sensors being in spatial and temporal registration.*

As a consequence, we have to accept that it will not be possible to have all the signals from the fused sensors in registration. We cannot combine the different sensor signals into one flow of feature vectors, every pixel of the imaged scene being described by a feature vector that has as many components as we have fused imaging sensors. As we are not able to obtain this feature vector, we cannot apply the strong tools of multidimensional pattern classification at the pixel level but, on the contrary, we must process the raw data of every sensor individually. We have to make an individual pixel segmentation in good and defect locations for every sensor separately and without using any information from the other sensors. Only at the later level of symbolic region

processing we will be able to roughly combine the individual segmentations into an overall quality grade (Fig. 16.2a). Thus, we lose much of the cross information that is contained in the individual raw pixel sensor data because we cannot have all these raw signals in pixelwise registration.

This problem can be solved if we are able to design imaging sensors sensing the different interesting physical properties of the timber that

1. have almost the same spatial, temporal and radiometric resolution; and
2. can be arranged in such a way that they image exactly the same location.

We call such a sensor combination a *multisensorial camera* because all the pixels of the different sensors are in registration, both with regard to location and to time (Fig. 16.2b). Such a camera produces a *feature vector* for every pixel, that is, a vectorial signal flow that can be processed pixelwise by multidimensional pattern classifiers. It is shown in the following that the design of such a camera is surprisingly less difficult than it may appear at first glance. We will explain this idea for the case of a wood inspection application.

16.3 A feature vector with every pixel

We would like to inspect fast-moving timber soft wood for the following defects:

1. *dark knots* with good contrast to the background;
2. sound knots that have the same color as the background;
3. regions with blue and red stain;
4. *resin pockets* that differ from the background color;
5. pithy wood pockets having identical color to resin pockets; and
6. 3-D defects such as flat bumps, broken edges, etc., that do not have any contrast to the background

As the timbers travel at a speed of typically 1–6 m/s, we will have to use fast scanning *linescan cameras* [1], mainly:

- A color linescan camera with typically 2048 pixels/line at a scanning frequency of 1500 scans/s for the detection of defects type 1, 3, 4, and 5, all of which can be segmented pixelwise in the color space in real-time, that is, based on our ColourBrain® technology [2].
  The discrimination between blue/red stain regions and dark knots will be difficult even when shape information is extracted. It will not be possible to differentiate between resin pockets and pithy wood capillaries that look identical in color and shape.

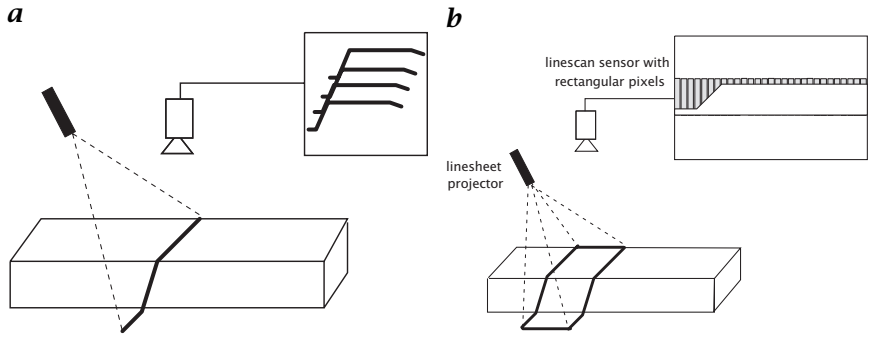


Figure 16.3: ***a*** *Traditional 3-D sensing requires a line projector and a matrix camera. This camera is either slow (50 frames/s) or has low spatial resolution;* ***b*** *the 3-D linescan camera. A sheet of light is imaged onto a line-scan sensor with rectangular pixels. The higher the timber surface, the larger the illuminated area of the corresponding pixel. The sensor signal is directly proportional to the timber 3-D profile.*

- A 3-D line-section camera that measures the 3-D profile of the timber wood. Such a system needs a laser line projector and a fast matrix camera. There is no matrix camera on the market having the resolution and frame rate of the color line camera. Can we design a linescan camera for 3-D triangulation?
- A camera that is sensitive to the surface density and fiber structure of the wood. Sound knots have the same color as the background but they are harder; the local density is much higher. An open pithy wood capillary has the same color and shape as a resin pocket but a different fiber structure (liquid phase for resin, foamy mushroom-type fiber structure for pithy wood, same mushroom structure for red and blue stain). Can we build a density/structure sensitive linescan camera?
- We would like to use all three cameras in pixelwise registration in order to produce a 5-D feature vector per imaged pixel with the mostly uncorrelated vector components (hue, intensity, saturation, height, density). We would like to perform an early pixelwise segmentation by classifying the feature vector of every pixel into previously trained defect classes. How can we perform all this in real-time?

16.4 A real-time three-dimensional linescan camera

One of the major limitations of the widely used *triangulation* based line-section method for measuring the height of a product moving on a conveyer is the need to use a matrix camera to image the projected

line of light (Fig. 16.3a). For a general discussion of triangulation techniques, see Section 18.4.

Matrix cameras are slow, 50 half-frames for TV standard or at most 1000 Hz with special low-resolution charge coupled device (CCD) sensors. Extracting the position of the imaged laser line and converting into z-coordinates relative to a reference plane requires additional processing time. This traditional approach does not lead to a camera with a spatial and temporal resolution identical to the color linescan camera and, thus, cannot be a candidate for a multisensorial camera sensor.

It requires but a little skill [3] to turn a linescan camera into a 3-D camera. We replace the line-of-light projector by a projector that projects a linesheet with a sharp edge onto the moving wood timber; Fig. 16.3b. The sharp white-to-dark edge will be deviated in the same way as the line-of-light by the height of the timber relative to a $h = 0$ reference plane. We image the scene onto a linescan sensor with rectangular pixels such as commonly used in spectroscopy. The magnification of the camera lens is set in such a way as to image the edge of the projected lightsheet just on the lower border of all the pixels if it hits the reference plane defined as level $h = 0$. For $h = h_{\max}$, the edge will hit the upper border of all rectangular pixels illuminating the entire area of every rectangular pixel. The illuminated portion of every pixel surface is, therefore, proportional to the height h of the timber. The CCD line sensor produces an output voltage directly proportional to $h(n)$; n = index of sensor pixel.

Strictly speaking, the luminous intensity I_n falling on the rectangular pixel is proportional to the local height h_n and to the local *reflectivity* ρ_n of the timber surface :

$$I_n = h_n \rho_n \tag{16.1}$$

This means that the sensor video signal is proportional to the local height and to the local reflectivity. There are two ways to get rid of the unwanted dependency on ρ_n: measure ρ_n with another camera and divide the sensor signal by ρ_n or use a source of light for which ρ_n is nearly constant. *Near infrared* light very often produces intensity images that are nearly constant, independent of the contrast in the visible range.

16.5 A real-time linescan scatter camera

We now have solved the problem of getting color and 3-D information with two linescan cameras having the same spatial and temporal resolution. We still need a solution for detecting noncolored *sound knots*, for discriminating between *resin pockets* and *pithy capillaries* and for discriminating between red and blue stain and timber surface spoiled

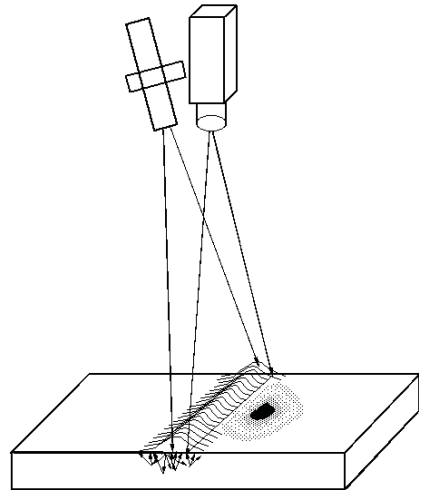


Figure 16.4: *The scatter camera observes the halo around a laser line penetrating the wood and scattering back.*

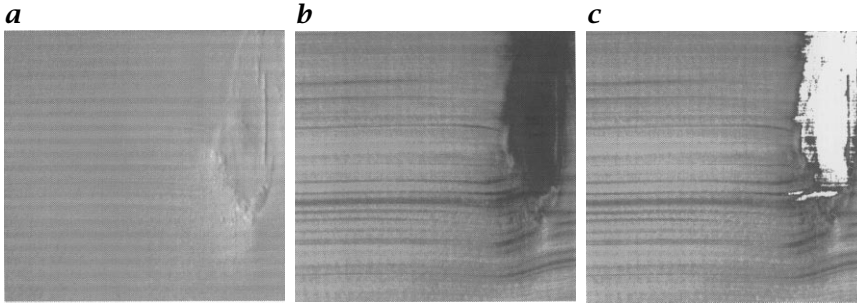


Figure 16.5: *Image of soft wood with a sound (noncolored) knot:* ***a*** *gray-level intensity image;* ***b*** *normalized scatter mage;* ***c*** *binarized scatter image.*

with dirt of similar color (remember that timber is often stored for long time periods under the open sky and that its surface is all but clean).

The sound knot is a region within the wood that is much harder then normal soft wood. As wood is a fibrous material, both the density and the orientation of these fibers will be different for the sound knot and for the normal soft wood.

It has been known for some time that the surface of wood has some degree of translucency; light impinging on the surface is penetrating somewhat into the material, scattered inside of the material and exits the material with an intensity and a spatial orientation that are modulated by the fiber density and fiber orientation [4].

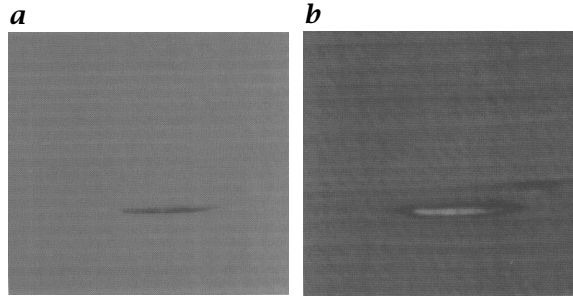


Figure 16.6: *Detection and identification of a resin pocket:* ***a*** *dark looking in the gray-level image. The resin pocket looks identical to a pithy capillary and cannot be discriminated from the latter;* ***b*** *bright looking scatter image. The liquid resin spreads out light and appears as a bright halo.*

We take advantage of this effect by projecting a laser line from an orthogonal position onto the timber and by observing the lateral light halo with a monochrome linescan camera offset from the laser line center by a few pixels (Fig. 16.4). For a sound knot, there will be little translucence and backscattering; the broadening of the projected line will be small, and the light intensity observed by the offset camera is low. Figure 16.5 shows the gray-level image of a sound knot and the corresponding scatter camera image; the much improved detectability of the scatter camera compared to the gray-level camera is obvious. The scatter image has such a high contrast that that the sound knot can be segmented by simple thresholding.

The scatter camera is also very helpful in discriminating between a resin pocket and a pithy capillary. The liquid phase of the resin is a very good light tube. The projected line broadens sharply when hitting the resin. The scatter camera therefore detects the resin pocket as a very bright spot compared to the dark-looking resin pocket image of the color camera (Fig. 16.6).

Red stain and blue stain are organic deteriorations of the wood surface caused by a fungus. This fungus destroys the fiber structure and thus prevents light scattering along the healthy fiber. Stained regions therefore look darker compared to the healthy soft wood. This effect can be used as an additional detector compared to the color camera or for discriminating between stain and spoiled wood surfaces.

There are a number of other defects for which the scatter camera is a very good detector. It is an amazingly simple camera with the benefit of having the same spatial and temporal resolution as the color camera. Therefore, it is a good candidate as a sensor for a multisensorial camera.

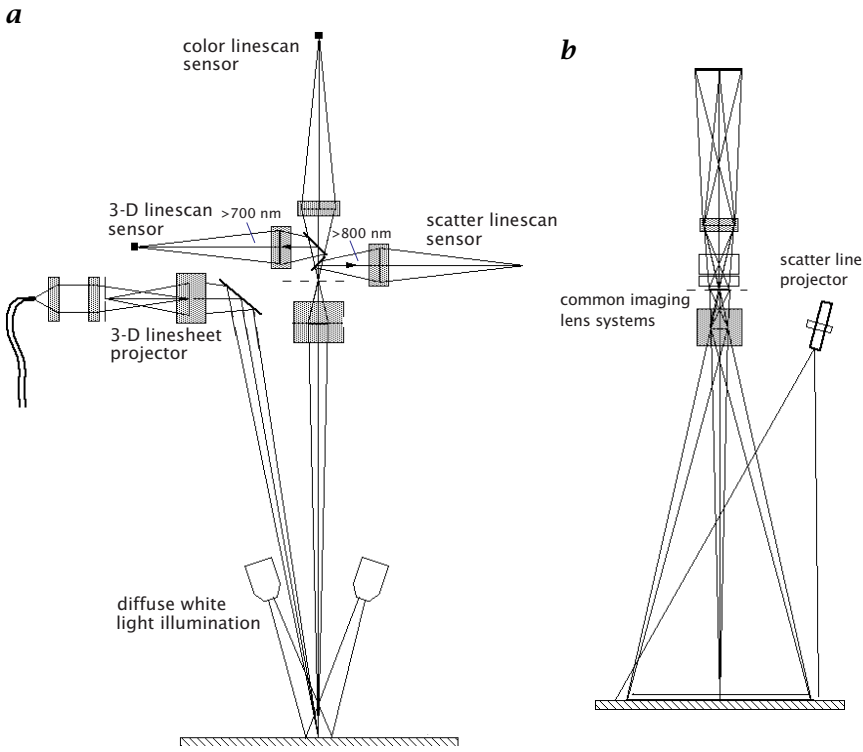


Figure 16.7: *Optical design of the multisensorial camera : **a** side view; **b** front view.*

16.6 The multisensorial color-height-scatter camera

Let us now consider the problem of combining the three linescan cameras into a multisensorial camera that images the same location of the timber with all pixels in registration. First, we use a different illumination for every linescan sensor:

- a diffuse white light illumination with a wavelength from 350 to 650 nm (blue to red);
- a laser line projected from an orthogonal position with a narrowband wavelength around 750 nm; and
- a linesheet from a laterally expanded laser with a narrowband bandwidth around 850 nm and projected under a triangulation angle.

Figure 16.7 sketches the overall arrangement of the linescan sensors and the illuminations. The optical path from each of the illumination systems is guided by *dichroitic mirrors* to the corresponding line sensor. The color sensor thus only receives light from the broadband white

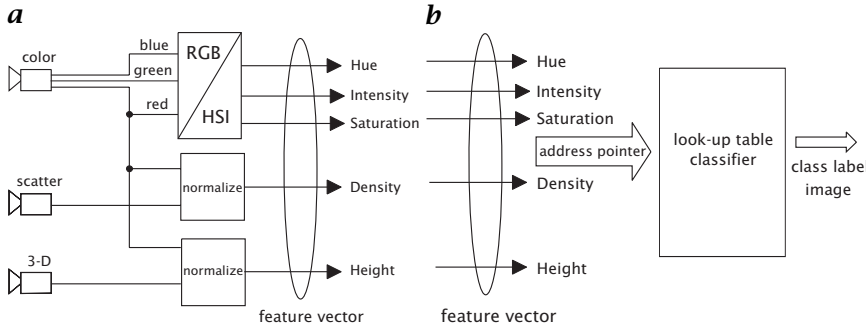


Figure 16.8: ***a*** *Forming a feature vector for every pixel. The scatter signal and the 3-D signal are normalized by dividing with the RED channel that is nearest to the wavelength of the illumination wavelengths of the scatter and the 3-D camera.* ***b*** *Real-time preprocessing of the vectorial signals with a trainable look-up table classifier.*

light source, the scatter sensor only from the orthogonally projected laser line, and the 3-D line sensor only from the lightsheet projector. All linescan sensors use the same imaging lens and are mechanically adjusted such that their pixels are in registration. All are driven by the same clock and synchronized with respect to linescanning frequency and start-of-line signals.

Figure 16.8a shows how the digitized signals from the three different sensors are preprocessed using simple look-up tables. The strongly correlated RGB signals are converted to the uncorrelated *hue*, *intensity*, and *saturation* (HIS) signals for better discrimination in the feature space. The scatter channel is normalized by dividing the signal I_n by the RED signal from the color camera using a *look-up table* divider. The wavelength of the RED channel and of the laser line are quite near to one another so that we can assume under practical conditions that the *reflectivity* ρ_n measured in the RED band of wavelengths and measured in the 750 nm band are not much different. The signals from the 3-D sensor are normalized in the same way.

The digital output of the multisensorial camera now is a 5 Byte signal corresponding to the local intensity, hue, saturation, height and scattering of the wood surface. Thus, the camera produces a 5-D *feature vector* $x = [I, H, S, H, SC]^T$ for every pixel.

We have reached our design goal.

16.7 Compressing the multisensorial camera signals

The data rate from our 5-channel multisensorial camera is five times higher then for a simple b/w camera. We can, however, very effectively compress these data using our concept of trained *look-up table classifiers* [2].

The 5 Byte signal is used as a pointer to a large look-up table for *real-time classification*. By reducing the total number of bits to, say, 25 bits, for example, $I = 6$ bits, $H = 5$ bits, $S = 4$ bits, $H = 4$ bits, $SC = 6$ bits, we do need a 32 Mbyte look-up table that maps every pixel into a defect class label. If we code into 32 different defect classes, we require a 5 bit output for the look-up table (Fig. 16.8b). The multisensorial camera thus produces the pixelwise real-time classification, a 5 Bit /pixel flow that is less than for a conventional scalar gray-scale camera. We have developed proprietary training algorithms for these LUT classifiers (ColourBrain®) that take advantage of all the cross information of the signals from the three linescan sensors. The class label image of a multisensorial camera is strongly compressed and still has not lost much of the original information. All cross information between the five different channels has been exploited during the pixelwise classification.

16.8 The one-chip multisensorial camera

The approach for combining the three different linescan sensors with dichroic mirrors is a rather traditional optical engineering one, which requires a significant amount of fine mechanical design. Is it not possible to combine the three sensors at chip-level?

There are three principles that we can use for this purpose:

- Time-multiplexing of the individual illuminations and imaging onto one (or two) gated sensors;
- Arranging on the same chip the different sensor pixels in an interleaved mosaic together with the required optical filters; or
- Use the fact that the penetration of light into silicon is increasing with wavelength. From 1000 nm onwards silicon is already significantly transparent. For soft x-rays, silicon is totally transparent. This allows the design of a multisensorial camera using stacked sensor chips and illuminations chosen in such a way that one channel penetrates through the first layer, the next through the first two layers, etc. [5]. Figure 16.9 shows as an example a one-chip solution for a camera combining color, 3-D and x-ray.

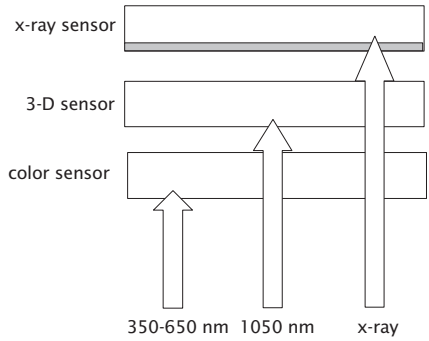


Figure 16.9: *Building a chip-based multisensorial camera by stacking image sensors. The color sensor substrate is partially transparent for the wavelengths of the 3-D light projector. Both the color and the 3-D sensor substrates are transparent to the x-rays. The x-ray sensor is covered with a scintillator.*

16.9 Conclusion

The multisensorial camera is a combination of different imaging sensors having the same spatial and temporal resolution and measuring different physical effects. In contrast to traditional sensor fusion these imaging sensors are arranged in such a way that their pixels are in registration. A multisensorial camera produces a feature vector per pixel and, thus, allows an early and powerful segmentation by real-time pixelwise classification.

We have shown for the case of wood inspection that such a camera can be realized using a rather simple approach. We also have briefly shown that this can be done using either classical optical engineering methods or more advanced imaging chip design methods.

The concept is simple and yet powerful enough to trigger the imagination of the image processing engineer to find different arrangements for particular applications. It also demonstrates that it cannot be unrewarding to go back from time to time to the roots, that is, give more attention to the poor pixel from where all the information comes and which our complex computers and algorithms have to process.

Acknowledgment

Part of this work has been supported by the German Federal Ministry of Research and Technology. We are grateful for this support.

16.10 References

[1] Birkeland, R., (1989). Practical experience with a fully automated line camera scanning system in a window manufacturing company. In *Proc. 3rd Int. Conf. On Scanning Technology in Sawmilling, Oct. 5-6, San Francisco, CA.*

[2] Massen, R., (1990). Color and shape classification with competing paradigms: neural networks versus trainable table classifiers. In *ECO 3rd Int. Conf. Optical Science, Den Haag, March, 1990.*

[3] Massen, R., (1996). Method and arrangement for the optical inspection of products. European Patent Application EP 94931039.5.

[4] Hagmann, O., (1996). *On reflections of wood. Wood quality features modeled by means of multivariate image projections to latent structures in multispectral images.* PhD thesis 198D, Lulea University of Technology, Sweden.

[5] Massen, R., (1996). Multisensorial camera with compact sensor arrangement. German Patent Application DE 19650705.7.

Part IV

Three-Dimensional Imaging

17 Geometric Calibration and Orientation of Digital Imaging Systems

Robert Godding

AICON GmbH, Braunschweig, Germany

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

17.1 Definitions

17.1.1 Camera calibration

Camera calibration in photogrammetric parlance refers to the determination of the parameters of interior orientation of individual cameras. When using digital cameras, it is advisable to analyze the complete imaging system, including camera, transfer units and possibly frame grabbers. The parameters to be found by calibration depend on the type of camera used. Once the imaging system has been calibrated, measurements can be made after the cameras have been duly oriented.

17.1.2 Camera orientation

Camera orientation usually includes determination of the parameters of exterior orientation to define the camera station and camera axis in the higher-order object-coordinate system, frequently called the *world coordinate system*. This requires the determination of three rotational and three translational parameters, that is, a total of six parameters for each camera.

17.1.3 System calibration

In many applications, fixed setups of various sensors are used for measurement. Examples are online measurement systems in which, for example, several cameras, laser pointers, pattern projectors, rotary stages, etc., may be used. If the entire system is considered the measurement tool proper, then the simultaneous calibration and orientation of all the components involved may be defined as *system calibration*.

17.2 Parameters influencing geometrical performance

17.2.1 Interior effects

All components of a digital imaging system leave their marks on the image of an object and thus on the measurement results obtained from processing this image. The following is a brief description of the relevant components (Fig. 17.1).

Optical system. Practically all lenses exhibit typical *radial-symmetrical distortion* that may vary greatly in magnitude. On the one hand, the lenses used in optical measurement systems are nearly distortion-free [1]. On the other hand, wide-angle lenses, above all, frequently exhibit distortion of several 100 μm at the edges of the field. Fisheye lenses are

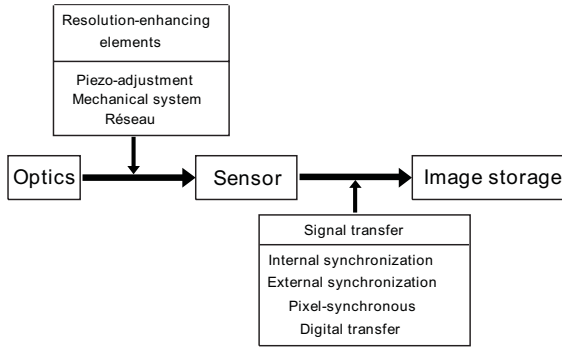


Figure 17.1: *Components of digital imaging systems.*

in a class of their own; they frequently have extreme distortion at the edges. Because radial-symmetrical distortion is a function of design, it cannot be considered an aberration.

By contrast, centering errors often unavoidable in lens making cause aberrations reflected in radial-asymmetrical and tangential distortion components [2]. Additional optical elements in the light path, such as the IR barrier filter and protective filter of the sensor, also leave their mark on the image and have to be considered in the calibration of a system.

Resolution-enhancing elements. The image size and the possible resolution of CCD sensors are limited. Presently on the market are metrology cameras such as Rollei's Q16 MetricCamera with up to 4000×4000 sensor elements [3]. Other, less frequent approaches use techniques designed to attain higher resolution by shifting commercial sensors in parallel to the image plane. Essentially, there are two different techniques. In the case of "microscanning," the interline transfer CCD sensors are shifted by minute amounts by means of piezoadjustment so that the light-sensitive sensor elements fall within the gaps between elements typical of this type of system, where they acquire additional image information [4, 5]. Alternatively, in "macroscanning," the sensors may be shifted by a multiple of their own size, resulting in a larger image format. Individual images are then oriented with respect to the overall image either by a highly precise mechanical system [6, 7] or opto-numerically as in the RolleiMetric Réseau Scanning Camera by measuring a glass-based reference grid in the image plane ("*réseau scanning*") [8].

All resolution-enhancing elements affect the overall accuracy of the imaging system. In scanner systems with purely mechanical correlation of individual images, the accuracy of the stepping mechanism has a

direct effect on the geometry of the high-resolution imagery. In the case of réseau scanning, the accuracy of the réseau is decisive for the attainable image-measuring accuracy [9].

Sensor and signal transfer. Due to their design, *charge-coupled device* (CCD) sensors usually offer high geometrical accuracy [10]. When judging an imaging system, its sensor should be assessed in conjunction with the frame grabber used. Geometrical errors of different magnitude may occur during A/D conversion of the video signal, depending on the type of *synchronization*, above all if *pixel-synchronous* signal transfer from camera to image storage is not guaranteed [9, 11]. However, in the case of pixel-synchronous readout of data, the additional transfer of the pixel clock pulse ensures that each sensor element will precisely match a picture element in the image storage. Very high accuracy has been proved for these types of cameras [1]. However, even with this type of transfer the square shape of individual pixels cannot be taken for granted. As with any kind of synchronization, most sensor-storage combinations make it necessary to make allowance for an affinity factor; in other words, the pixels may have different extension in the direction of lines and columns.

17.2.2 Exterior effects

If several cameras are used in an online metrology system, both the parameters of interior orientation and those of exterior orientation may vary, the former, for example, caused by refocusing and changes of temperature, the latter caused by mechanical effects or fluctuations of temperature. The resulting effects range from scale errors during object measurement all the way up to complex model deformation. This is why all systems of this kind should make it possible to check or redetermine all relevant parameters.

17.3 Model of image formation with the aid of optical systems

Image formation by an optical system can, in principle, be described by the mathematical rules of *central perspective*. According to these rules, an object is imaged in a plane so that the object points P_i and the corresponding image points P_i' are located on straight lines through the perspective center O_j (Fig. 17.2). The following holds under idealized conditions for the formation of a point image in the image plane:

$$\begin{bmatrix} x_{ij} \\ y_{ij} \end{bmatrix} = \frac{-c}{Z_{ij}^*} \begin{bmatrix} X_{ij}^* \\ Y_{ij}^* \end{bmatrix} \tag{17.1}$$

with

$$\begin{bmatrix} X_{ij} \\ Y_{ij} \\ Z_{ij} \end{bmatrix} = D(\omega, \varphi, \kappa)_j \begin{bmatrix} X_i - X_{oj} \\ Y_i - Y_{oj} \\ Z_i - Z_{oj} \end{bmatrix} \tag{17.2}$$

where X_i, Y_i, Z_i are the coordinates of an object point P_i in the object-coordinate system K; X_{oj}, Y_{oj}, Z_{oj} are the coordinates of the perspective center O_j in the object-coordinate system K; X^*_{ij}, Y^*_{ij}, Z^*_{ij} are the coordinates of the object point P_i in the coordinate system K^*_j; x_{ij}, y_{ij} are the coordinates of the image point in the image-coordinate system K_B; and $D(\omega, \varphi, \kappa)_j$ is the rotation matrix between K and K^*_j; and c is the distance between perspective center and image plane, the system K^*_j being parallel to the system K_B with the origin in the perspective center O_j [12].

The foregoing representation splits up the process of image formation in such a manner that in: (1) it is primarily the image-space parameters; and in (2) primarily the object-space parameters, that is, the parameters of exterior orientation, that come to bear.

This ideal concept is not attained in reality where many influences are encountered due to the different components of the imaging system. These can be modeled as departures from rigorous central perspective. The following section describes various approaches to mathematical camera models.

17.4 Camera models

When optical systems are used for measurement, modeling the entire process of image formation is decisive in obtaining accuracy. Basically, the same ideas apply, for example, to projection systems for which models can be set up similarly to imaging systems.

Before we continue, we have to define an *image-coordinate system* K_B in the image plane of the camera. In most electro-optical cameras, this image plane is defined by the sensor plane; only in special designs (e. g., in réseau scanning cameras [8]), is this plane defined differently. While in the majority of analog cameras used for metrology purposes the image-coordinate system is defined by projected fiducial marks or réseau crosses, this definition is not required for digital cameras. Here it is entirely sufficient to place the origin of image-coordinate system in the center of the digital images in the storage (Fig. 17.3). Because the pixel interval in column direction in the storage is equal to the interval of the corresponding sensor elements, the unit “pixel in column direction” may serve as a unit of measure in the image space. All parameters of interior orientation can be directly computed in this unit, without conversion to metric values.

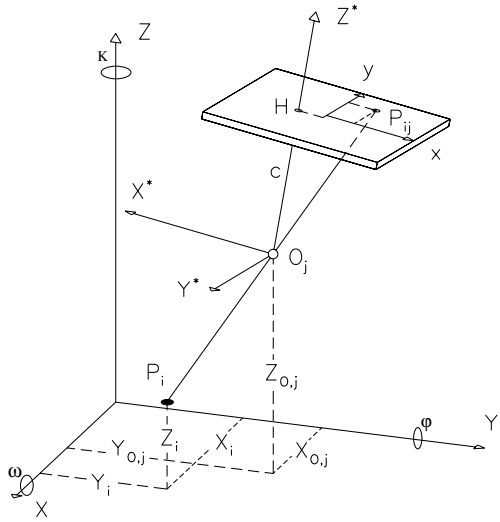


Figure 17.2: *Principle of central perspective [13].*

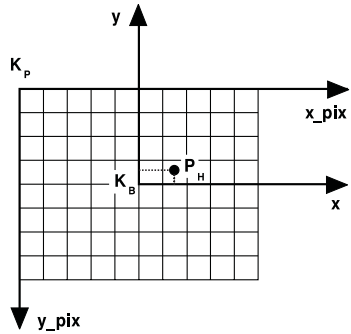


Figure 17.3: *Definition of image-coordinate system.*

17.4.1 Calibrated focal length and principal-point location

The reference axis for the camera model is not the optical axis in its physical sense, but a principal ray, which on the object side is perpendicular to the image plane defined in the foregoing and intersects the latter at the *principal point* $P_H(x_H, y_H)$. The perspective center O_j is located at distance c_K (also known as *calibrated focal length*) perpendicularly in front of the principal point [14].

The original formulation of Eq. (17.1) is thus expanded as follows:

$$\begin{bmatrix} x_{ij} \\ y_{ij} \end{bmatrix} = \frac{-c_k}{Z^*_{ij}} \begin{bmatrix} X^*_{ij} \\ Y^*_{ij} \end{bmatrix} + \begin{bmatrix} x_H \\ y_H \end{bmatrix} \tag{17.3}$$

17.4.2 Distortion and affinity

The following additional correction function can be applied to Eq. (17.3) for radially symmetrical, radial asymmetrical and tangential distortion:

$$\begin{bmatrix} x_{ij} \\ y_{ij} \end{bmatrix} = \frac{-c_k}{Z^*_{ij}} \begin{bmatrix} X^*_{ij} \\ Y^*_{ij} \end{bmatrix} + \begin{bmatrix} x_H \\ y_H \end{bmatrix} + \begin{bmatrix} \mathrm{d}x(V,A) \\ \mathrm{d}y(V,A) \end{bmatrix} \tag{17.4}$$

Here, $\mathrm{d}x$ and $\mathrm{d}y$ may now be defined differently, depending on the type of camera used, and are made up of the following different components:

$$\begin{aligned} \mathrm{d}x &= \mathrm{d}x_{\mathrm{sym}} + \mathrm{d}x_{\mathrm{asy}} + \mathrm{d}x_{\mathrm{aff}} \\ \mathrm{d}y &= \mathrm{d}y_{\mathrm{sym}} + \mathrm{d}y_{\mathrm{asy}} + \mathrm{d}y_{\mathrm{aff}} \end{aligned} \tag{17.5}$$

Radial-symmetrical distortion. The *radial-symmetrical distortion* typical of a lens can generally be expressed with sufficient accuracy by a polynomial of odd powers of the image radius (x_{ij} and y_{ij} are henceforth called x and y for the sake of simplicity):

$$\mathrm{d}r_{\mathrm{sym}} = A_1(r^3 - r_0^2 r) + A_2(r^5 - r_0^4 r) + A_3(r^7 - r_0^6 r) \tag{17.6}$$

where $\mathrm{d}r_{\mathrm{sym}}$ is the radial-symmetrical distortion correction; r is the image radius from $r^2 = x^2 + y^2$; A_1, A_2, A_3 are the polynomial coefficients; and r_0 is the second zero crossing of the distortion curve, so that we obtain

$$\mathrm{d}x_{\mathrm{sym}} = \frac{\mathrm{d}r_{\mathrm{sym}}}{r} x \quad \text{and} \quad \mathrm{d}y_{\mathrm{sym}} = \frac{\mathrm{d}r_{\mathrm{sym}}}{r} y \tag{17.7}$$

A polynomial with two coefficients is generally sufficient to describe radial-symmetrical distortion. Expanding this distortion model, it is possible to describe even lenses with pronounced departure from perspective projection (e. g., fisheye lenses) with sufficient accuracy. In the case of very pronounced distortion it is advisable to introduce an additional point of symmetry $P_S(x_S, y_S)$. Figure 17.4 shows a typical distortion curve.

For numerical stabilization and far-reaching avoidance of correlations between the coefficients of the distortion function and the calibrated focal lengths, a linear component of the distortion curve is split off by specifying a second zero crossing [15].

Lenz [16] proposes a different formulation for determining radial-symmetrical distortion, which includes only one coefficient. We thus obtain the following equation:

$$\mathrm{d}r_{\mathrm{sym}} = r \frac{1 - \sqrt{1 - 4Kr^2}}{1 + \sqrt{1 - 4Kr^2}} \tag{17.8}$$

where K is the distortion coefficient to be determined.

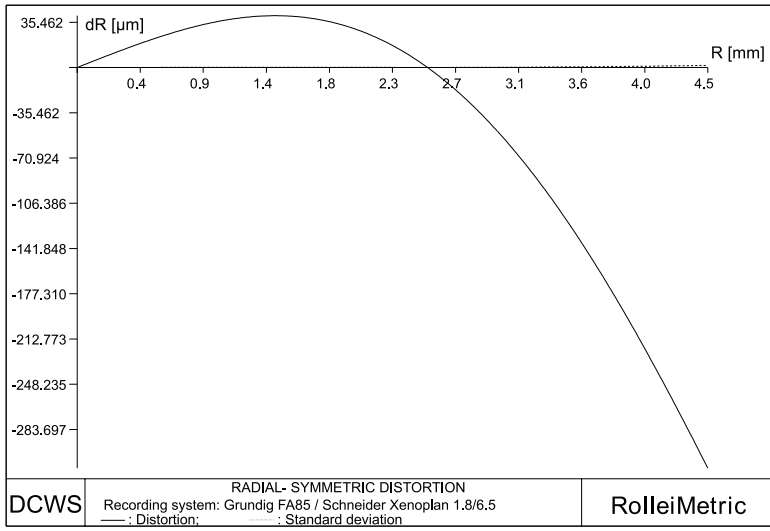


Figure 17.4: *Typical distortion curve of a lens.*

Radially asymmetrical and tangential distortion. To handle *radial-asymmetrical* and *tangential distortion*, various different formulations are possible. Based on Conrady [17], these distortion components may be formulated as follows [2]:

$$\begin{aligned} \mathrm{d}x_{\mathrm{asy}} &= B_1(r^2 + 2x^2) + 2B_2xy \\ \mathrm{d}y_{\mathrm{asy}} &= B_2(r^2 + 2y^2) + 2B_1xy \end{aligned} \tag{17.9}$$

In other words, these effects are always described with the two additional parameters B_1 and B_2.

This formulation is expanded by Brown [18], who adds parameters to describe overall image deformation or the lack of image-plane flatness:

$$\begin{aligned} \mathrm{d}x_{\mathrm{asy}} &= (D_1(x^2 - y^2) + D_2x^2y^2 + D3(x^4 - y^4))x/c_K \\ &+ E_1xy + E_2y^2 + E_3x^2y + E_4xy^2 + E_5x^2y^2 \\ \mathrm{d}y_{\mathrm{asy}} &= (D_1(x^2 - y^2) + D_2x^2y^2 + D3(x^4 - y^4))y/c_K \\ &+ E_6xy + E_7x^2 + E_8x^2y + E_9xy^2 + E_10x^2y^2 \end{aligned} \tag{17.10}$$

In view of the large number of coefficients, however, this formulation implies a certain risk of too many parameters. Moreover, since this model was primarily developed for large-format analog imaging systems, some of the parameters cannot be directly interpreted for applications using digital imaging systems. Equations Eq. (17.7) are generally sufficient to describe asymmetrical effects. Figure 17.5 shows typical effects for radial-symmetrical and tangential distortion.

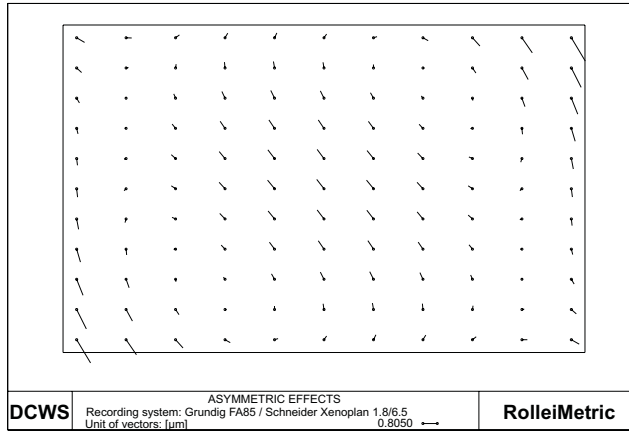


Figure 17.5: *Radially symmetrical and tangential distortion.*

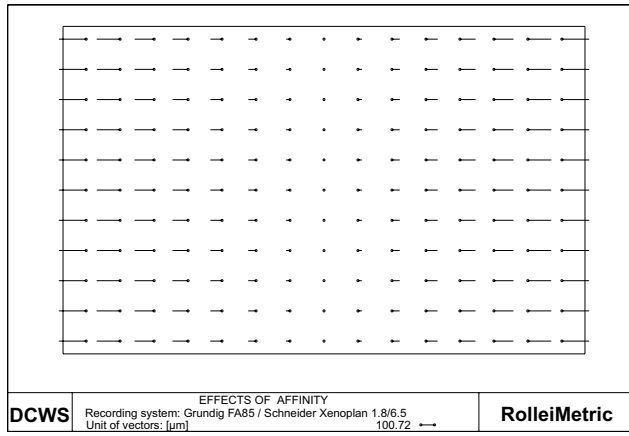


Figure 17.6: *Effects of affinity.*

Affinity and nonorthogonality. The differences in length and width of the pixels in the image storage caused by synchronization can be taken into account by an *affinity factor*. In addition, an affinity direction may be determined, which primarily describes the orthogonality of the axes of the image-coordinate system K_B. An example may be a line scanner that does not move perpendicularly to the line direction. Allowance for these two effects can be made as follows:

$$\mathrm{d}x_{\mathrm{aff}} = C_1 x + C_2 x \quad \text{and} \quad \mathrm{d}y_{\mathrm{aff}} = 0 \tag{17.11}$$

Figure 17.6 gives an example of the effect of affinity.

Additional parameters. The introduction of additional parameters may be of interest for special applications. Fryer [19] and [20] describe formulations that also make allowance for distance-related components of distortion. However, these are primarily effective with medium- and large-image formats and the corresponding lenses and are of only minor importance for the wide field of digital uses.

Gerdes et al. [21] use a different camera model in which an additional two parameters have to be determined for the oblique position of the sensor.

17.5 Calibration and orientation techniques

17.5.1 In the laboratory

Distortion parameters can be determined in the laboratory under clearly defined conditions.

In the goniometer method, a highly precise grid plate is positioned in the image plane of a camera. Then, the goniometer is used to sight the grid intersections from the object side and to determine the corresponding angles. Distortion values can then be obtained by a comparison between nominal and actual values.

In the collimator technique, test patterns are projected onto the image plane by several collimators set up at defined angles to each other. Here also, the parameters of interior orientation can be obtained by a comparison between nominal and actual values, though only for cameras focused at infinity [14].

Apart from this restriction, there are more reasons weighing against the use of the aforementioned laboratory techniques for calibrating digital imaging systems, including the following:

- The equipment layout is high;
- The interior orientation of the cameras used normally is not stable, requiring regular recalibration by the user; and
- Interior orientation including distortion varies at different focus and aperture settings so that calibration under practical conditions appears more appropriate.

17.5.2 Bundle adjustment

All the parameters required for calibration and orientation may be obtained by means of photogrammetric bundle adjustment. In *bundle adjustment*, two so-called observation equations are set up for each point measured in an image, based on Eqs. (17.2) and (17.4). The total of all equations for the image points of all corresponding object points results in a system that makes it possible to determine the unknown

parameters [22]. Because this is a nonlinear system of equations, no linearization is initially necessary. The computation is made iteratively by the method of least squares, the unknowns being determined in such a way that the squares of deviations are minimized at the image coordinates observed. Newer approaches are working with modern algorithms such as balanced parameter estimation [23]. Bundle adjustment thus allows simultaneous determination of the unknown object coordinates, exterior orientation and interior orientation with all relevant system parameters of the imaging system. In addition, standard deviations are computed for all parameters, which give a measure of the quality of the imaging system.

Calibration based exclusively on image information. This method is particularly well suited for calibrating individual imaging systems. It requires a survey of a field of points in a geometrically stable photogrammetric assembly. The points need not include any points with known object coordinates (control points); the coordinates of all points need only be known approximately [22]. It is, however, necessary that the point field be stable for the duration of image acquisition. The scale of the point field likewise has no effect on the determination of the desired image-space parameters. Figure 17.7 shows a point field suitable for calibration.

The accuracy of the system studied can be judged from the residual mismatches of the image coordinates as well as the standard deviation of the unit of weight after adjustment (Fig. 17.8). The effect of synchronization errors, for example, becomes immediately apparent, for instance, by larger residual mismatches of different magnitude in line and column direction.

Figure 17.9 gives a diagrammatic view of the minimum setup for surveying a point array with which the aforementioned system parameters can be determined. The array may be a three-dimensional test field with a sufficient number of properly distributed, circular, retroreflecting targets. This test field is first recorded in three frontal images, with camera and field at an angle of 90° for determining affinity and 180° for determining the location of the principal point. In addition, four convergent images of the test field are used to give the assembly the necessary geometric stability for determination of the object coordinates and to minimize correlation with exterior orientation.

Optimum use of the image format is a precondition for the determination of distortion parameters. However, this requirement need not be satisfied for all individual images. It is sufficient if the image points of all images cover the format uniformly and completely.

If this setup is followed, seven images will be obtained roughly as shown in Fig. 17.10; their outer frame stands for the image format, the inner frame for the image of the square test field, and the arrowhead for

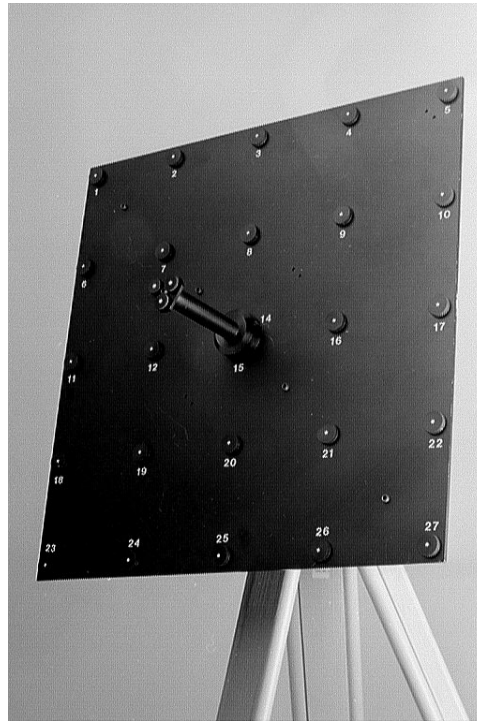


Figure 17.7: *Test array for camera calibration.*

the position of the test field. It is generally preferable to rotate the test field with the aid of a suitable suspension in front of the camera instead of moving the camera for image acquisition. The use of retroreflecting targets and a ring light guarantee proper, high-contrast reproduction of the object points, which is indispensable for precise and reliable measurement. A complete, commercially available software package offering far-reaching automation of the process is described in Godding [1].

Calibration and orientation with the aid of additional object information. Once the imaging system has been calibrated, its orientation can be found by resection in space. The latter may be seen as a special bundle adjustment in which the parameters of interior orientation and the object coordinates are known. This requires a minimum of three control points in space whose object coordinates in the world coordinate system are known and whose image points have been measured with the imaging system to be oriented.

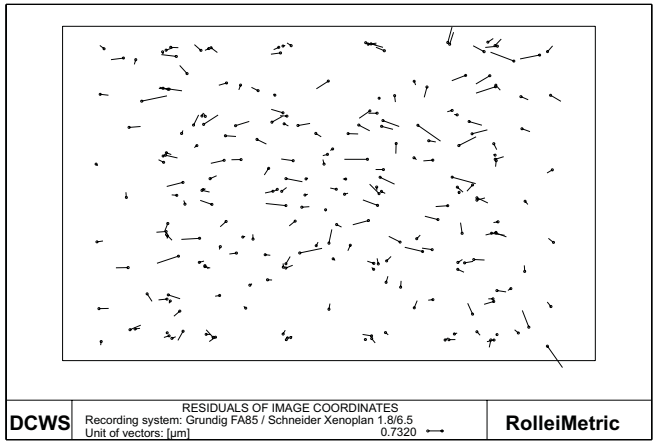


Figure 17.8: *Residual mismatches after bundle adjustment.*

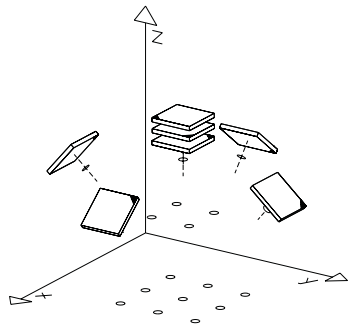


Figure 17.9: *Imaging setup for calibration [1].*

In addition to orientation, calibration of an imaging system is also possible with a single image. However, as a single image does not allow the object coordinates to be determined, suitable information within the object has to be available in the form of a 3-D control-point array [24]. But constructing, maintaining and regularly checking such an array is rather costly, all the more so as it should be mobile so that it may be used for different applications. The control pattern should completely fill the measurement range of the cameras to be calibrated and oriented to ensure good agreement between calibration and measurement volumes.

The expense is considerably less if several images are available. For a two-image assembly and one camera, a spatial array of points that

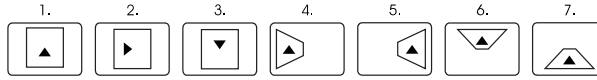


Figure 17.10: *Test field.*

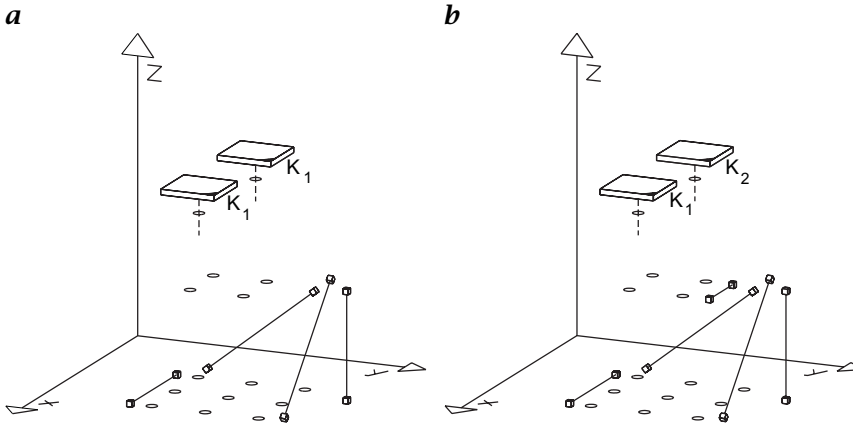


Figure 17.11: *Scale setup for calibrating* ***a*** *one camera and* ***b*** *two cameras.*

need to be known only approximately plus, as additional information, several known distances (scales) distributed in the object space will be sufficient; this is similar to the previous paragraph. In an ideal case, one scale on the camera axis, another one perpendicular to it, and two oblique scales in two perpendicular planes parallel to the camera axis are required (Fig. 17.11a). This will considerably reduce the object-side expense, because the creation and checking of scales is much simpler than that of an extensive three-dimensional array of control points.

A similar setup is possible if the double-image assembly is recorded with several cameras instead of just one. This is, in principle, the case with online measurement systems. An additional scale is then required in the foreground of the object space, bringing the total number of scales to five (Fig. 17.11b).

If at least one of the two cameras can be rolled, the oblique scales can be dispensed with, provided that the rolled image is used for calibration [24].

The setups described in Fig. 17.11 are, of course, applicable to more than two cameras as well. In other words, all the cameras of a measurement system can be calibrated if the aforementioned conditions are created for each of the cameras. At least two cameras have to be calibrated in common, with the scales set up as described. Simultaneous calibration of all cameras is also possible, but then the scale information must also be simultaneously available to all the cameras. If all

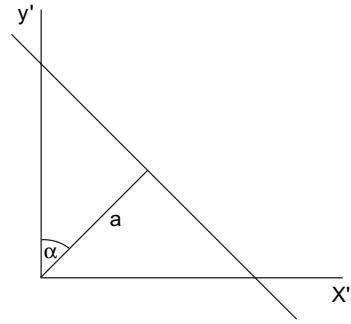


Figure 17.12: *Principle of the plumbline method.*

cameras also are to be calibrated in common, this will have to be done via common points.

System calibration. As we have seen from the previous two paragraphs, joint calibration and orientation of all cameras involved and thus calibration of the entire system are possible if certain conditions are met. With the aid of bundle adjustment, the two problems can, in principle, be solved jointly with a suitable array of control points or a spatial point array of unknown coordinates plus additional scales. The cameras then already are in measurement position during calibration. Possible correlations between the exterior and interior orientations required are thus neutralized because the calibration setup is identical to the measurement setup.

Apart from the imaging systems, other components can be calibrated and oriented within the framework of system calibration. Godding and Luhmann [25] describe a technique in which a suitable procedure in an *online measurement system* allows both the interior and exterior orientation of the cameras involved as well as the orientation of a rotary stage to be determined with the aid of a spatial point array and additional scales. The calibration of a line projector within a measurement system using photogrammetric techniques was, for example, presented by Strutz [26].

17.5.3 Other techniques

Based on the fact that straight lines in the object space have to be reproduced as straight lines in the image, the so-called *plumbline method* serves to determine distortion. The technique is predicated on the fact that the calibrated focal length and principal-point location are known [27].

According to Fig. 17.12, each of the straight-line points imaged are governed by the relationship

$$x' \sin\alpha + y' \cos\alpha = a \tag{17.12}$$

where x' and y' can be expressed as follows:

$$\begin{aligned} x' &= x_{ij} + \mathrm{d}x_{\mathrm{sym}} + \mathrm{d}x_{\mathrm{asy}} \\ y' &= y_{ij} + \mathrm{d}y_{\mathrm{sym}} + \mathrm{d}y_{\mathrm{asy}} \end{aligned} \tag{17.13}$$

where $\mathrm{d}x_{\mathrm{sym}}$, $\mathrm{d}y_{\mathrm{sym}}$, $\mathrm{d}x_{\mathrm{asy}}$, and $\mathrm{d}y_{\mathrm{asy}}$ correspond to the formulations in Eq. (17.7), (17.9), and (17.10). It is an advantage of this method that, assuming suitable selection of the straight lines in the object, a large number of observations is available for determining distortion, and measurement of the straight lines in the image lending itself to automation. A disadvantage of the technique is the fact that simultaneous determination of all relevant parameters of interior orientation is impossible.

Lenz [16] presented a technique in which an imaging system was similarly calibrated and oriented in several steps. The technique requires a plane test field with known coordinates, which generally should not be oriented parallel to the image plane. Modeling radial-symmetrical distortion with only one coefficient (see also Section 17.4.2) and neglecting asymmetrical effects allows the calibration to be based entirely on linear models. Because these do not need to be resolved interactively, the technique is very fast. It is a disadvantage, however, that here also it is impossible to determine all the parameters simultaneously and that, for example, the location of the principal point and pixel affinity have to be determined externally.

Gerdes et al. [21] describe a method in which cameras are permitted to be calibrated and oriented with the aid of parallel straight lines projected onto the image. A cube of known dimensions is required for the purpose as a calibrating medium. Vanishing points and vanishing lines can be computed from the cube edges projected onto the image and used to determine the unknown parameters.

A frequently used formulation for the determination of the parameters of exterior and interior orientation is the method of direct linear transformation (DLT) first proposed by Abdel-Aziz and Karara [28]. This establishes a linear relationship between image and object points. The original imaging equation is converted to a transformation with 11 parameters that initially have no physical importance. By introducing additional conditions between these coefficients it is then possible to derive the parameters of interior and exterior orientation, including the introduction of distortion models [29]. Because the linear formulation of DLT can be solved directly, without approximations for the unknowns, the technique is frequently used to determine approximations for bundle adjustment. The method requires a spatial test field

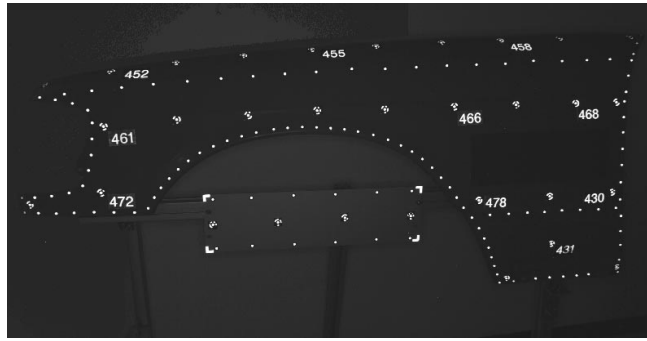


Figure 17.13: *Measurement of a wing.*

with a minimum of six known control points, a sufficient number of additional points being needed to determine distortion. However, if more images are to be used to determine interior orientation or object coordinates, nonlinear models will have to be used here also.

17.6 Photogrammetric applications

17.6.1 Applications with simultaneous calibration

The imaging setup for many photogrammetric applications allows *simultaneous calibration* of cameras. It is an advantage of this solution that no additional effort is required for external calibration of the cameras and that current camera data for the instant of exposure can be determined by bundle adjustment. This procedure, however, is possible only if the evaluation software offers the option of simultaneous calibration. As an example, let us look at measurement of an automobile part (Fig. 17.13).

A total of nine photos were taken with a Rollei Q16 MetricCamera (Fig. 17.14) with a resolution of 4096 × 4096 sensor elements. RolleiMetric Close-Range Digital Workstation software was used for evaluation. This allows fully automatic determination of 3-D coordinates, starting with measurement of image points right up to computation of all unknown parameters. In addition to target sizes and the 3-D coordinates of all measured points in the world coordinate system, these include the camera parameters and all camera stations. For this example the coordinates have an accuracy of approximately 1/100 mm in each of the three coordinate axes. Figure 17.15a, b illustrates another example from the automotive industry. Here, torsion tests were made in the course of deformation measurements. The data were obtained by photogrammetric means. A total of 3000 points all around the vehicle were recorded in a total of 170 images with the aid of a digital cam-

Figure 17.14: *Rollei Q16 MetricCamera.*

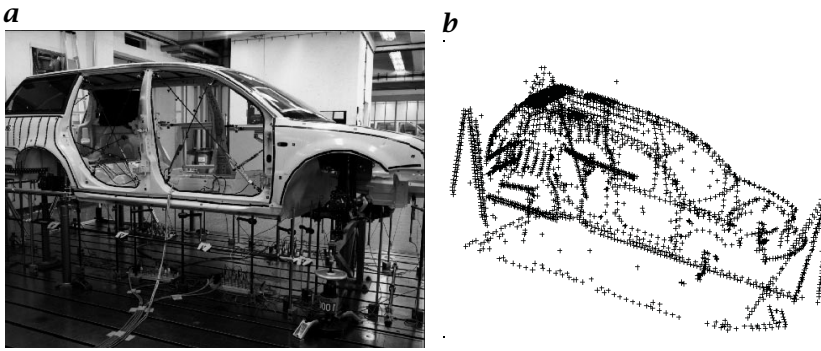


Figure 17.15: ***a*** *Measurement of a car;* ***b*** *3-D view of measured points*

era with a resolution of 3000 × 2000 sensor elements. Here also, the camera was simultaneously calibrated during image acquisition. The points measured were accurate to within about 5/100 mm. Most photogrammetric applications for high-precision 3-D industrial metrology work are based on simultaneous calibration. Numerous other uses can be found in the aviation industry (measuring aircraft components and fixtures), in the aeronautical industry (measuring satellites and antennas), and in civil engineering (measuring finished components). Some of these applications are discussed in Volume 3, Chapter 16.

17.6.2 Applications with precalibrated camera

Robot calibration. At KUKA Robotertechnik of Augsburg industrial robots have been reliably measured, adjusted and calibrated on the assembly line at two specially installed workplaces during the past two years [30]. To measure the required positions and orientations,

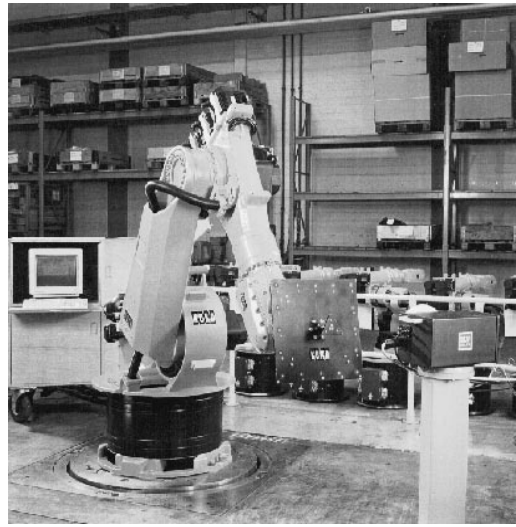


Figure 17.16: *Robot adjustment.*

a photogrammetric metrology system consisting of one or two Rollei-Metric Réseau Scanning Cameras (RSCs) are mounted on a rugged tripod (Fig. 17.16). Using a shiftable standard CCD sensor, these cameras reach a resolution of 4200×4200 picture elements at an image format of 50×50 mm^2 with an accuracy of better than 1 μm in image space. The orientation of the single images in relation to the entire image is done in an optical-numerical way by a réseau measurement. Besides, this principle, which is described in Riechmann [8], allows the focusing of the camera without changing the interior orientation.

The cameras are controlled by a commercial PC with a standard frame-grabber, running under Windows NT. The PC serves for operator prompting, for processing and outputting results and for connection to the robot control. The measurement system is basically independent of the robot control.

The interior orientation of the cameras is determined once in a special calibration measurement. With this known interior orientation, it is possible to determine the orientation of the cameras. Various target plates 450 mm $\times$ 450 mm in size are used, with reflective targets as control points, which are also identified as tools for the robot. A second target plate of 600 mm $\times$ 600 mm with an adapter serves for prior determination of the robot base and external orientation of the camera. To transfer the different coordinate systems, highly precise bores in the target plates are used with special adapters. A mechanical precision measuring machine serves as a higher-order metrology system for measuring the bores.

After orientation the online measurement of the robots is possible. The quality of the system orientation is verified by special measurements. A recalibration of the system normally is necessary only in time periods of some months.

Other applications. Other photogrammetric applications for the 3-D capture of objects can be found, for example, in accident photography and in architecture. In these fields, it is primarily scale drawings or rectified scale photos (orthophotos) that are obtained from the photograms. The cameras employed are generally calibrated for different focus settings using the methods described in the foregoing. An example is the RolleiMetric ChipPack with a resolution of 2000×2000 sensor elements. Special metric lenses, which guarantee reproducible focus setting by mechanical click stops of the focusing ring, keep interior orientation constant for prolonged periods. The data of interior orientation are entered in the software and thus used for plotting and all computations. This guarantees high-precision 3-D plotting with minimum expense in the phase of image acquisition.

17.7 References

[1] Godding, R., (1993). Ein photogrammetrisches Verfahren zur Überprüfung und Kalibrierung digitaler Bildaufnahmesysteme. *Zeitschrift für Photogrammetrie und Fernerkundung*, **2**:82–90.

[2] Brown, D. C., (1966). Decentering distortion of lenses. *Photogrammetric Engineering*, **32**:444–462.

[3] Schafmeister, R., (1997). Erste Erfahrungen mit der neuen Rollei Q16 MetricCamera. In *Publikationen der Deutschen Gesellschaft für Photogrammetrie und Fernerkundung (DGPF)*, Vol. 1, pp. 367–378. Berlin: DGPF.

[4] Lenz, R. and Lenz, U., (1990). Calibration of a color CCD camera with 3000×2300 picture elements. ISPRS Symposium, Com. V. Close-Range Photogrammetry meets Machine Vision, Zürich. *Proc. SPIE*, **1395**:104–111.

[5] Richter, U., (1993). Hardware-Komponenten für die Bildaufnahme mit höchster örtlicher Auflösung. In *Publikationen der Deutschen Gesellschaft für Photogrammetrie und Fernerkundung.*, Vol. 1, pp. 367–378, Berlin: DGPF.

[6] Holdorf, M., (1993). Höchstauflösende digitale Aufnahmesysteme mit Réseau Scanning und Line-Scan-Kameras. In *Symposium Bildverarbeitung '93,*, pp. 45–51, Esslingen: Technische Akademie Esslingen.

[7] Poitz, H., (1993). Die UMK SCAN von Carl Zeiss Jena, ein neues System für die digitale Industrie-Photogrammetrie. In *Tagungsband zur DGPF-Jahrestagung 1992 in Jena, DGPF; Berlin.*

[8] Riechmann, W., (1993). *Hochgenaue photogrammetrische Online-Objekterfassung.* PhD thesis, Technical University of Brunswick.

[9] Bösemann, W., Godding, R., and Riechmann, W., (1990). Photogrammetric investigation of CCD cameras. ISPRS symposium, close-range photogrammetry meets machine vision, Zürich. *Com. V. Close-Proc. SPIE*, **1395**:119–126.

[10] Lenz, R., (1988). Zur Genauigkeit der Videometrie mit CCD-Sensoren. In Bunke, H., Kübler, O., and Stucki, P. (eds.), *Proc. 10. DAGM-Symp. Mustererkennung 1988, Informatik Fachberichte 180*, pp. 179–189, Berlin, DAGM: Springer.

[11] Beyer, H., (1992). Advances in characterization and calibration of digital imaging systems. International archives of photogrammetry and remote sensing. 17th ISPRS Congress, Washington. *Com. V*, **29**:545–555.

[12] Wester-Ebbinghaus, W., (1989). Mehrbild-Photogrammetrie — räumliche Triangulation mit Richtungsbündeln. In *Symposium Bildverarbeitung '89*, pp. 25.1–25.13. Technische Akademie Esslingen.

[13] Dold, J., (1994). Photogrammetrie. Vermessungsverfahren im Maschinen- und Anlagenbau. In Schwarz, W. (ed.), *Schriftenreihe des Deutschen Vereins für Vermessungswesen DVW*, Vol. 13. Stuttgart: Verlag Konrad Wittwer.

[14] Rüger, Pietschner, and Regensburger, (1978). *Photogrammetrie—Verfahren und Geräte*. Berlin: VEB Verlag für Bauwesen.

[15] Wester-Ebbinghaus, W., (1980). Photographisch-numerische Bestimmung der geometrischen Abbildungseigenschaften eines optischen Systems. *Optik*, **3**:253–259.

[16] Lenz, R., (1987). Linsenfehlerkorrigierte Eichung von Halbleiterkameras mit Standardobjektiven für hochgenaue 3D-Messungen in Echtzeit. In Paulus, E. (ed.), *Proc. 9. DAGM-Symp. Mustererkennung 1987, Informatik Fachberichte 149*, pp. 212–216, Berlin, DAGM: Springer.

[17] Conrady, A., (1919). Decentered lens systems. *Royal Astronomical Society, Monthly Notices*, **79**:384–390.

[18] Brown, D., (1976). The bundle adjustment—progress and perspectives. Helsinki 1976. In *International Archives of Photogrammetry*, Vol. 21(3), p. 303041.

[19] Fryer, J., (1989). Camera calibration in non-topographic photogrammetry. In *Handbook of Non-Topographic Photogrammetry,* 2nd edition, pp. 51–69. American Society of Photogrammetry and Remote Sensing.

[20] Fraser, C. and Shortis, M., (1992). Variation of distortion within the photographic field. *Photogrammetric Engineering and Remote Sensing*, **58(6)**: 851–855.

[21] Gerdes, R., Otterbach, R., and Kammüller, R., (1993). Kalibrierung eines digitalen Bildverarbeitungssystems mit CCD-Kamera. *Technisches Messen*, **60(6)**:256–261.

[22] Wester-Ebbinghaus, W., (1985). Bündeltriangulation mit gemeinsamer Ausgleichung photogrammetrischer und geodätischer Beobachtungen. *Zeitschrift für Vermessungswesen*, **3**:101–111.

[23] Fellbaum, M., (1996). PROMP —A new bundle adjustment program using combined parameter estimation. *International Archives of Photogrammetry and Remote Sensing*, **31(B3)**:192–196.

[24] Wester-Ebbinghaus, W., (1985). Verfahren zur Feldkalibrierung von photogrammetrischen Aufnahmekammern im Nahbereich. *DGK-Reihe B*, **275**: 106–114.

[25] Godding, R. and Luhmann, T., (1992). Calibration and accuracy assessment of a multi-sensor online-photogrammetric system. In *International Archives of Photogrammetry and Remote Sensing, Com. V, 17. ISPRS Congress, Washington*, Vol. XXIX, pp. 24–29. Bethesda, MD: American Society for Photogrammetry and Remote Sensing.

[26] Strutz, T., (1993). *Ein genaues aktives optisches Triangulationsverfahren zur Oberflächenvermessung*. PhD thesis, Magdeburg Technical University.

[27] Fryer, J. and Brown, D. C., (1986). Lens distortion for close-range photogrammetry. *Photogrammetric Engineering and Remote Sensing*, **52**:51–58.

[28] Abdel-Aziz, Y. J. and Karara, H. M., (1971). Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. In *Symposium of the American Society of Photogrammetry on Close-Range Photogrammetry*, Virginia: Falls Church.

[29] Bopp, H. and Kraus, H., (1978). Ein Orientierungs- und Kalibrierungsverfahren für nichttopographische Anwendungen der Photogrammetrie. *Allgemeine Vermessungs-Nachrichten (AVN)*, **5**:182–188.

[30] Godding, R., Lehmann, M., and Rawiel, G., (1997). Robot adjustment and 3-D calibration—photogrammetric quality control in daily use. *Optical 3-D Measurement Techniques*, **4**:158–168.

18 Principles of Three-Dimensional Imaging Techniques

Rudolf Schwarte[1,2,3], Horst Heinol[2,3], Bernd Buxbaum[1,3], Thorsten Ringbeck[1,4], Zhanping Xu[3], and Klaus Hartmann[2,4]

[1]Institut für Nachrichtenverarbeitung, Universität-GH Siegen, Germany
[2]Zentrum für Sensorsysteme (ZESS)
[3]S-TEC GmbH, Siegen
[4]aicoss GmbH, Siegen

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

18.1 Introduction

Electronic imaging using charge coupled devices (CCD) cameras and digital image processing found widespread applications in research, industrial production, communications, and consumer goods. Nowadays, 3-D image acquisition and processing appears to be on the verge of a comparably stormy and far-reaching development. Fast and noncontact optical form measurements are of significant importance in industrial production, including inspection on tolerances and completeness, robot vision in automatic assembly, surface inspection, and reverse engineering. They are equally important for surveillance of secured areas, 3-D object recognition and navigation. Three-dimensional optical form measurements deliver the *absolute* 3-D geometry of objects that are largely independent of the object's surface reflectivity, the distance of the objects from the sensor, and illumination conditions. Thus, 3-D optical sensors deliver in real scale the dimensions of an object, which are rotation-, translation-, and illumination-invariant.

No question, there is a much higher effort in 3-D vision compared to 2-D conventional vision. Although in standard projective imaging, for example, by means of CCD cameras, the depth information from 3-D scenes is lost, a human observer does not lose 3-D interpretation, for example, in common TV and movies. Usually, he or she has no difficulty in recognizing the projected 3-D scenes and in interacting with it. It becomes evident that human 3-D perception as a learning system depends largely on *a priori* knowledge, and is capable of exploiting, for example, "structure from motion" and "shape from shading." In an optimal way, human depth perception is supported by two 3-D facilities of the visual system: by stereovision and by autofocus adaptation [1]. In the past, the lack of knowledge about the human visual system and its natural power may have hindered the development and usage of 3-D optical measuring systems. Further obstacles that continue to make it difficult to survey a wealth of different procedures in 3-D acquisition, include specifications that are hard to comprehend or are often inexact, and poor knowledge of their specific features and limits. In this situation, the German working group on "Optical Form Measurements," organized by the DGZfP, Berlin, and the VDI/VDE/GMA, Düsseldorf, has been working for almost a decade on optical depth perception. A result of the working group is the handbook OF1, *Methods of Optical Form Measurements* [2].

It is the goal of this chapter to provide an overview of the techniques for optical form measurements in a well-organized and comparable hierarchical scheme. An insight is given into the basic problems, and new developments are pointed out. Other overviews have previously been given by Breuckmann [3], Engelhardt [4], Jiang and Bunke [5] and Küchel [6]. Also, the followings subjects of 3-D imaging are discussed in other

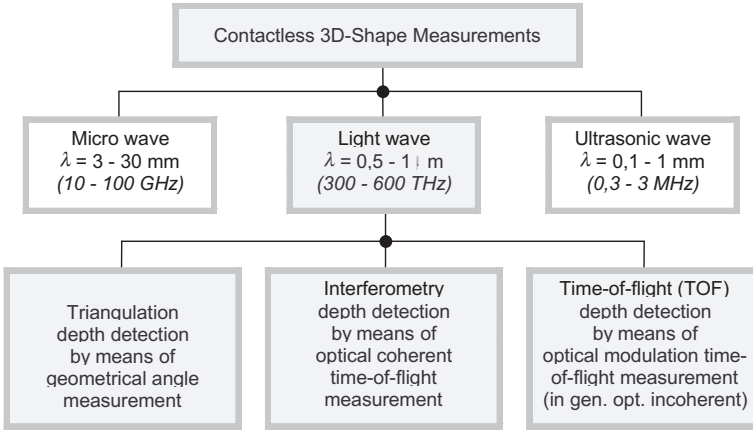


Figure 18.1: *Principles of noncontact 3-D shape measurements.*

chapters of this handbook: camera calibration (Chapter 17), physical limits and optical coherence tomography (Chapter 19), free-form surface measurements combining photogrammetric and structured-light volume techniques (Chapter 20), and confocal microscopy (Chapter 21). Volume 3 includes application reports in Chapters 16, 17, 20, and 21.

18.2 Basic principles

Continuous progress in industrial automation and the demand for increasing product quality are the driving forces for fast, exact, and noncontact 3-D object measurements. Fig. 18.1 shows the three basic types of radiation that can be used for remote measurements.

Microwaves are particularly suitable for large-scale 3-D measurements either by triangulation (e. g., *global positioning system* (GPS), determination of an unknown point of a triangle by three sides) or directly by time-of-flight measurements (e. g., conventional radar and synthetic aperture interferometry) (see [7, 8, 9]). For industrial production automation, these techniques, in general, do not reach the required angular resolution due to diffraction. A circular antenna with a diameter d generates a radiation cone (*Airy pattern*, see Chapter 4) with an angle 2α, where

$$\sin\alpha = 1.22\frac{\lambda}{d} = \frac{w}{2f} \tag{18.1}$$

If we use, for example, an antenna with $d = 122\,\text{mm}$ diameter and an extremely short microwave ($\lambda = 3\,\text{mm}$, $\nu = 100\,\text{GHz}$), the opening angle 2α of the radiation cone is 60 mrad and the minimum spot size or waist

w is already 60 mm at 1 m distance, respectively, at the focal length f. For *ultrasound* we get the same relations for the same wavelength of, for example, $\lambda = 3$ mm ($\nu = 110$ kHz in normal atmosphere). Additional difficulties with ultrasound are the significant sensitivity of the propagation speed of sound from pressure and temperature (with a relative change of about 2.2×10^{-3} per °C without light and only -0.93×10^{-6} per °C with light) and, moreover, the increasing attenuation at higher frequencies and a high reflectivity that is similar to a mirror of technical surfaces.

In contrast to microwave and ultrasound, optical 3-D sensors possess a 10^3 to 10^4 higher lateral and angular resolution due to the shorter wavelength in the range of 300 nm (ultraviolet) to 3 μm (infrared) (Section 2.2.1); (attenuation at gigahertz in air is too high). Depth information can be gained either by triangulation or by time-of-flight measurements. In rare cases it is possible to make use of the quadratic decrease of irradiation with the distance from a point source. This and other such radiometric techniques are not considered in the following.

As shown in Fig. 18.1, optical form measurements are based on three different principles: triangulation; interferometry; and time-of-flight measurements.

Triangulation normally determines an unknown visual point within a triangle by means of a known optical basis and the related side angles pointing to the unknown point. This often-used principle is partitioned in a wealth of partly very different 3-D techniques, as illustrated in Fig. 18.3.

Continuous wave (CW) and pulse time-of-flight techniques measure the time of flight of the envelope of a modulated optical signal (group velocity). These techniques usually apply *incoherent optical signals*. Figure 18.5 shows the hierarchical partitioning of this technique.

Interferometry measures depth also by means of the time-of-flight. In this case, however, the phase of the optical wave itself is used. This requires *coherent* mixing and correlation of the wavefront reflected from the 3-D object with a reference wavefront. Figure 18.9 shows the further subpartitioning of interferometric techniques for 3-D form measurements.

As soon as signals with either spatial or temporal periodicity are used for 3-D image acquisition, all three techniques result in interferograms that can be evaluated partly with the same mathematical algorithms (e. g., phase-shift technique). This is illustrated in the following example.

Example 18.1: Interferograms with 3-D imaging

With *triangulation*, the spatial frequencies (wavenumbers) of a stripe projector interfere with the spatial frequency of a sensor array. With *CW time-of-flight* measurements, the radio-frequency (RF) modulated signal reflected back from the object interferes in time with the RF mixing signal of the detector. Homodyne reception of a 3-D scene with a 2-D mixer results in an RF-modulation interferogram (homodyne/heterodyne means mixing of equal/different frequencies) [10]. Finally, with *interferometry* a mixing and correlation process is generated by coherent temporal superposition of the object and the reference wavefronts and by squaring of the resulting electromagnetic field within the detector's photocharge. The resulting signal in the detector, the quantum-electronically generated photocurrent, is proportional to the optical energy or to the squared field strength (Chapter 7). This process takes places either at each photosensor, at a photographic film, or in the retina of the eye.

18.3 Some criteria and specifications

The depth information measured by a 3-D sensor constitutes a spatial point cloud. It can be given on a regular grid either in Cartesian coordinates $Z(x, y)$, or in polar coordinates $R(\theta, \phi)$. This type of information is called a *depth map* or *depth image*. For many applications, this information is sufficient. Together with the depth map, most 3-D sensors also deliver a signal amplitude. Thus, we obtain a standard intensity or gray-scale 3-D surface image $G(x, y, z)$.

Given the tremendous advances in computer graphics, it is no longer a problem to compute realistic visualizations from 3-D object surfaces even in real time. The true problem remains the fast and precise acquisition of the depth map with a large volume and in a natural environment. Today, we are still far away from such a complete and video rate depth image acquisition.

On the basis of the three fundamental 3-D techniques, triangulation, time-of-flight, and interferometry, more and more powerful measuring systems are developed for quite different application areas. The most critical parameters of such systems are the depth measuring range z and the depth resolution δ_z. Figure 18.2 illustrates the measuring and resolution ranges that are covered by the existing industrial measuring systems. The figure shows the relative resolution δ_z/z as a function of the object distance z. Due to electronic time drifts or mechanical instabilities, the measuring uncertainty σ_z can be (depending on the measuring system) much larger than the resolution δ_z. The increasing use of imaging systems for all three techniques reduces the measuring times significantly.

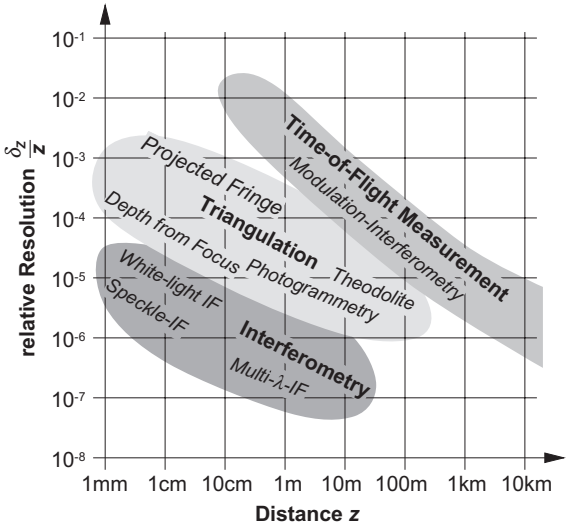


Figure 18.2: *Relative resolution of methods for optical shape measurements.*

The highest absolute resolutions are achieved by interferometry, which achieves values better than $\lambda/100$. Multiwavelength techniques additionally enable *absolute* measurements in the 10 m range.

Triangulation techniques can be used with high resolution from the millimeter range (depth of focus techniques, Chapter 20) to the 100 km range (classical photogrammetry), or even up to light-years distance with the earth orbit diameter as the optical baseline (astronomy).

So-called *active triangulation systems* with a stripe projector work almost like a 3-D camera (see Chapter 16, Chapter 21). Online photogrammetry with digital cameras enables fast 3-D measurements of special targets attached to the 3-D object (see Chapter 16, Chapter 17). A complete surface 3-D reconstruction outside the targets, however, still requires several minutes if at all possible by naturally existing points appropriate for correspondence.

With only 6.7 ps time-of-flight per millimeter, time-of flight depth estimation is an extreme challenge for time measurements. The possible resolution and standard deviation due to electronic time drifts is practically independent of the distance and lies in the millimeter range. Significant improvements are possible if the time-consuming correlation process is transferred as much as possible from electronic components to optical components and by parallel operation. This is realized in a new inherently mixing and correlating photodetector, the Photonic Mixer Device (PMD), which makes possible a high-integral 3-D imaging sensor [11].

Table 18.1: *Criteria and features for 3-D optical form measurements*

No.	Feature
1	Measuring dimensions X, Y, Z or R, θ, ϕ, basic distance, working distance, etc.
2	Relative or absolute distance measurement
3	Flexibility of the measuring dimensions
4	Measuring uncertainty (accuracy) σ_X, σ_Y, σ_Z
5	Resolution (repeatability, precision) δ_X, δ_Y, δ_Z
6	Linearity in the measuring volume
7	Measuring spot (form and size)
8	Measuring time per pixel/voxel
9	Possibilities for calibration
10	Reachable 3-D space (shadowing, boreholes)
11	Type of measurement: depth map and gray-scale image in Cartesian or polar coordinates
12	Depth map $Z(x,y)$ or true 3-D image $g(x,y,z)$
13	Illumination source: spectrum, modulation, power, coherency, polarization, laser class, etc.
14	Detector, correlator
15	Sensitivity of the detector to environmental illumination, multiple reflections, temperature, etc.
16	Surface properties: roughness, specular reflections, etc.
17	Robustness against surface slope and discontinuities
18	Size of measuring system
19	Costs
20	Special features

Table 18.1 summarizes a number of further important specifications and features that should be considered for an optical 3-D sensor.

18.4 Triangulation

Triangulation is the most widely used technique for optical form measurements. Figure 18.3 shows the hierarchy of the most important variants, which, despite the same basic principle, partly appear extremely different. At the highest level, we distinguish the following: (1) focus techniques; (2) active triangulation with structured illumination; (3) passive triangulation techniques on the basis of digital photogrammetry and stereoscopy; (4) theodolite measuring systems; and (5) shape from shading techniques. The rapid progress of optical triangulation

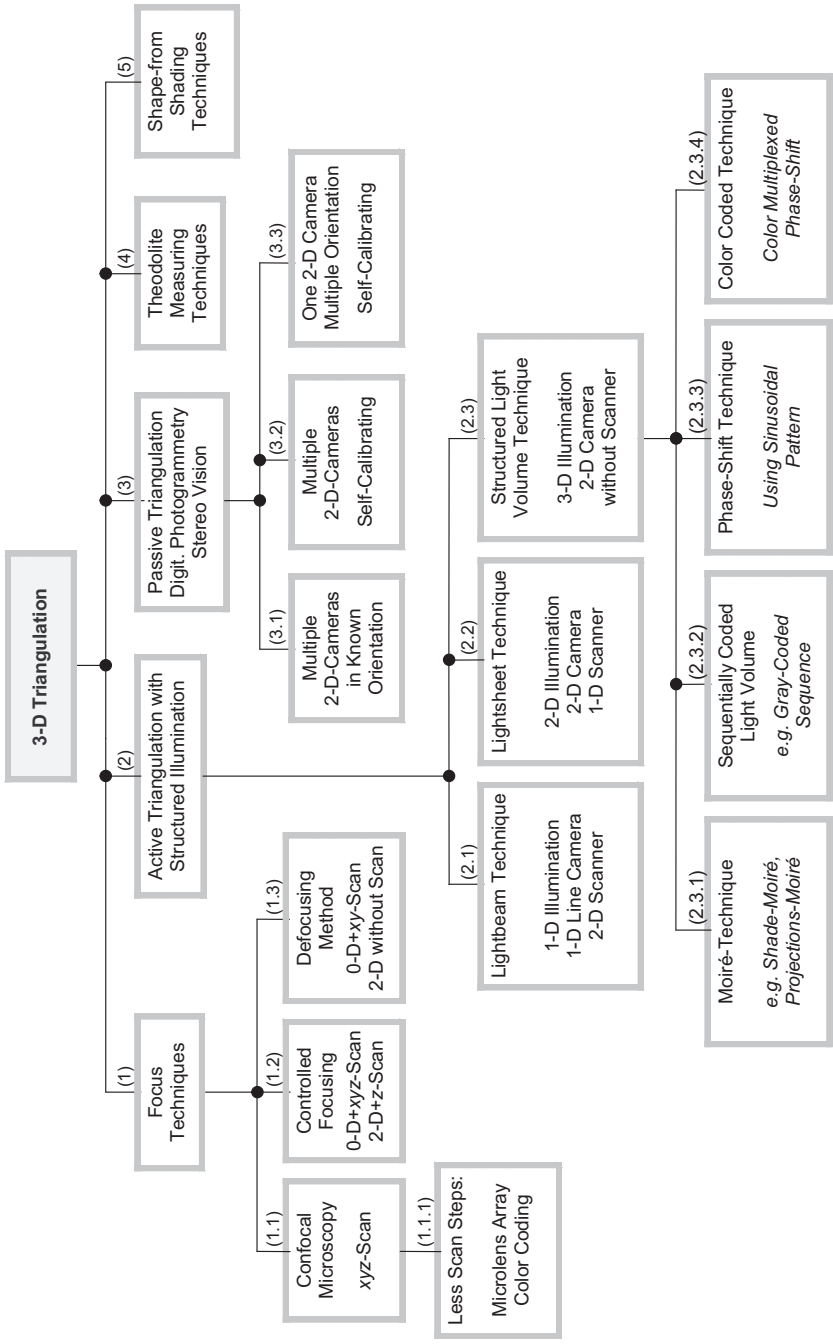


Figure 18.3: *Hierarchy of the most important principles of triangulation techniques.*

(especially active with structured light and passive with digital photogrammetry and with combinations of both) is already a giant step toward the goal of a 3-D triangulation camera and real-time stereovision. In the following subsections, a survey of the five basic variants of triangulation techniques is given.

18.4.1 Focus techniques

The critical parameters of *focus techniques* are the diameter of the diffraction-limited spot or waist w in the focal plane

$$w = 2.44\frac{\lambda f}{d} = 2f \sin\alpha \tag{18.2}$$

and the Rayleigh depth of focus

$$\Delta z_R = \frac{\lambda}{\sin^2\alpha} \tag{18.3}$$

where $\sin\alpha$, f, and d are the numerical aperture, the focal length, and the free diameter of the optical system, respectively.

The technique of *confocal microscopy* (1.1 in Fig. 18.3) utilizes the double spatial filtering at the focal plane by both illumination and detection of the object using a pinhole. The detector "sees" only illuminated points at the focal plane. Because only a single point is measured at a time, the acquisition of a true 3-D image requires scanning in all three spatial directions x, y, and z. Confocal microscopy with a microlens array and a CCD matrix sensor acquires one image at a time and thus needs only a depth scan. This depth scan can be avoided by utilizing the chromatic aberration and performing a spectral analysis (1.1.1 in Fig. 18.3) [12].

Area-extended measurements are also achieved by the systems reported by Engelhardt and Häusler [13] and Engelhardt [14]. A detailed account on 3-D confocal microscopy is given in Chapter 21.

The method of controlled focusing (1.2 in Fig. 18.3) delivers a height profile of a surface $Z(x,y)$ by scanning the xy plane with a fast Z control using, for example, a differential photodiode for high angular resolution [2, 15]. With the defocusing method (1.3 in Fig. 18.3), the distance can either be determined by the diameter or the intensity of the spot. A depth scan can be avoided by spectral analysis provided that the focal length f depends approximately linearly on the wavelength [16].

18.4.2 Active triangulation

With light point or 1-D *laser triangulation* (2.1 in Fig. 18.3), the light source emitting a collimated beam (pencil beam), the detector, and the

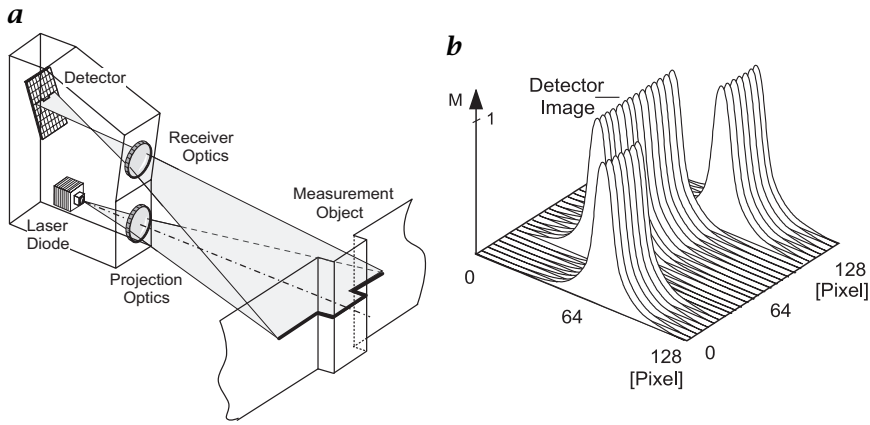


Figure 18.4: *Lightsheet triangulation:* ***a*** *instrument;* ***b*** *detector image.*

illuminated object point form the so-called triangulation triangle. On the side of the sender, the angle to the triangulation basis is fixed while on the side of the detector it is determined either by a CCD line sensor or a position-sensitive photodetector (PSD). From this angle, the depth can be determined. The depth resolution δ_Z for laser illumination is given by

$$\delta_Z = \frac{\lambda}{2\pi \sin\theta \sin\alpha_d} \tag{18.4}$$

and the measuring range Δz (two times the depth of focus [17]) by

$$\Delta z = \frac{2\lambda}{\sin^2 \alpha_d} \tag{18.5}$$

where $\sin\alpha_d$ and θ are the aperture of the detector optics and the triangulation angle, respectively. The acquisition of a depth image with this technique requires an xy scan [2, p. 6] [18, p. 1]. For a more detailed treatment, see Section 19.4.

With the lightsheet technique (2.2 in Fig. 18.3) or 2-D laser triangulation, generally a laser beam is expanded via cylindrical lenses to a light plane. The cross section of the lightsheet and of a 3-D object form a light stripe (the height profile) that is imaged onto a 2-D detector. Thus, only a 1-D scan perpendicular to the light plane is required for 3-D imaging [2, p. 8]; [19, 20].

Figure 18.4 shows schematically such a *lightsheet triangulation* instrument. The height profile generates the charge image on the CCD detector shown in Figure 18.4b. In order to obtain maximum depth resolution, the detector plane, the plane of the image-forming optics (perpendicular to the optical axis), and the plane of the object to be

measured, have a common axis and, thus, meet the *Scheimpflug condition*.

The *light volume triangulation* (2.3 in Fig. 18.3) illuminates the whole 3-D object to be measured with structured light (see Chapter 20). Thus, no scanning is required [21]. With the Moiré technique, (see 2.3.1 in Fig. 18.3), the projected texture is observed through an adapted reference structure. The superposition of these two patterns produces spatial beat frequencies, respectively, and a beat pattern with much lower wavenumbers that can be observed by a detector with a correspondingly lower spatial resolution. In this way, the depth resolution can be increased by one order of magnitude over conventional stripe projector systems [19, p. 16]; [22].

The sequentially coded light volume technique (2.3.2 in Fig. 18.3) illuminates the 3-D object with a sequence of binary patterns with increasing wavenumber in such a way that each pixel can be associated with a code, for example, a 10-digit Gray code, from which the *absolute* distance can be inferred [2, p. 10]; [23]. For detail, see Chapter 20. Another variant of the stripe projection technique, which has also found widespread application, is the phase-shift or projected fringe technique (2.3.3 in Fig. 18.3). A programmable LCD projector illuminates the scene with sinusoidal patterns with different phase positions. In order to evaluate the phase information, at least 3 or 4 (120° or 90° phase shift) independent measurements are required [2, p. 12]. This technique also results in a significant depth resolution. In conjunction with an additional sequential binary coding (so-called *Gray code phase shift technique*), absolute depth can be measured with high resolution.

The color-coded light volume technique (2.3.4 in Fig. 18.3) requires only a single image acquisition as three color channels are acquired simultaneously. The phase and thus the depth can, for example, be computed from red, blue, and green stripe patterns that are phase shifted from each other by 120° [24].

18.4.3 Passive triangulation

Passive triangulation techniques (3 in Fig. 18.3) basically include the different forms of *digital photogrammetry* and (as a special subclass *stereovision*). Passive in this context means that the geometrical arrangement of the illumination is not considered. In the area of industrial inspection, the classical photogrammetric techniques for evaluation of aerial photographs have been optimized for close distances. These techniques have formed the new field of *close-range photogrammetry*. In this handbook, detailed descriptions are given in Chapter 17 and Volume 3, Chapters 16 and 20. For photogrammetric techniques, at least three different views of a point are required to determine its 3-D position. For dynamical processes, often multiple cameras with

known relative positions (3.1 in Fig. 18.3) or self-calibrating methods (3.2 in Fig. 18.3) are used. For static scenes, a single camera that takes images from three or more different unknown views is sufficient (3.3 in Fig. 18.3) [25]. If a 3-D object is taken from different perspectives with a high-resolution digital camera, relative standard deviations in the positions σ_X/X, σ_Y/Y, and σ_Z/Z of better than 10^{-5} come close to time-consuming classical photographic techniques of photogrammetry. High computing power and optimized algorithms make online inspection with about 50 targets and a period of 4 s possible [22]. For further details on techniques of this type and the evaluation of the obtained 3-D point clouds, see Volume 3, Chapter 17. Photogrammetric camera calibration and orientation estimation is dealt with in Chapter 17.

18.4.4 Theodolites

So far, *theodolites* are still the most accurate triangulation systems available with a relative distance error of about 5×10^{-6}. The high accuracy is important due to the long measuring times. A target is focused with at least two theodolites. The horizontal and vertical angles are measured electronically, and the 3-D coordinates of the target are computed from the measured angle and the known positions of the theodolites [2, p. 14]. Theodolites are used for accurate measurements of large-scale objects. In modern systems, sometimes a 1-D laser radar distance is integrated.

18.4.5 Shape from shading

The *shape from shading* techniques infer from the image irradiance and the known position of the camera and the light sources, the surface normal of the surface elements. From this information, the 3-D form can be computed [1, p. 39]. The various types of shape-from-shading techniques including extensions using multiple images with different illuminations or image sequences with moving light sources (*photometric stereo*) are discussed in detail in Volume 2, Chapter 19.

18.5 Time-of-flight (TOF) of modulated light

The distance of an object or the depth z can easily be determined by the echo *time-of-flight* (TOF) τ of a light signal sent by the sensor and reflected back from the object to the sensor via

$$z = c\tau/2 \tag{18.6}$$

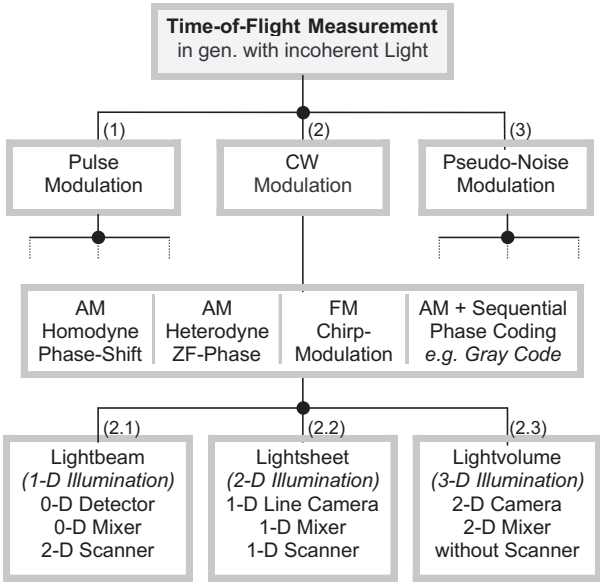


Figure 18.5: *Hierarchy of the most important principles of modulation-based optical depth measurements [26, 27]; [2, p. 26].*

This basic relation is valid for both time-of-flight and interferometric distance measurements. In the first case, the time-of-flight of a modulated optical signal, that is, the *group velocity*, is measured. Generally, this is done by *correlation* with a suitable reference signal. Therefore, the partitioning in Fig. 18.5 distinguishes between the different types of signals: (1) pulse modulation; (2) continuous wave (CW) modulation; (3) and pseudo random modulation. The basic problem of all TOF techniques is the extremely high light speed of 300 m/μs or 300 μm/ps, which requires correspondingly high temporal resolutions for the measuring techniques.

18.5.1 Pulse modulation

With *pulse modulation*, the time of flight is measured directly by correlating a start and stop signal with a parallel running counter. Pulse modulating techniques can distinguish multiple targets. A disadvantage is the temperature-sensitive time delay and the nonlinearity of the transients of pulsed laser diodes in addition to the high demands in bandwidth and dynamics for the amplifiers.

18.5.2 Continuous wave (CW) modulation

This principle of TOF measurement can be understood as a sort of RF Interferometry based on an optical carrier modulation and in that way as "Optical RF Interferometry" (ORFI). All imaginable variations are similar to that of structured illumination interferometry in triangulation as well as to that in rear optical interferometry. The echo-TOF τ of sine wave modulation can be determined either by *heterodyne mixing* (different frequencies are mixed, resulting in the beat frequency and phase difference $\varphi = 2\pi\nu\tau$) or by *homodyne mixing* (same frequencies are mixed, resulting in a baseband signal proportional to $\cos\varphi$). The frequency-modulated chirp modulation is used for higher resolution or to determine the TOF-dependent frequency shift or to expect pulse compression for multitarget detection. The low range of a unique depth determination of only $\Delta z = \lambda_m/2$ can be extended by rectangular 0° to 180° switching of the phase of the rectangular frequency. A definite distance measurement is then possible using several measurements with different switching frequencies according to a Gray code in the same way as with the sequentially coded structured light projection technique, here in time, in Chapter 20 in space or azimuth (Section 18.4.2). Because of the variety of modulation techniques, Fig. 18.5 further partitions only the sinusoidal modulation techniques.

Three-dimensional optical form measurements with TOF techniques is (in contrast to 1-D geodetic distance measurements) not frequently used in industrial applications. This is due to the principal technical problems discussed at the beginning of this section. The block (2.1) in Fig. 18.5, labeled "Lightbeam," describes 1-D TOF instruments that require a 2-D scanning system for the acquisition of depth images. If a modulated lightsheet or plane is used (2.2 in Fig. 18.5), we get 2-D information as a light stripe in space. In this case, a 1-D scanner is sufficient. With a modulated light volume, no scanning is required at all. In this case, the receiver requires a 2-D mixer for CW demodulation. It produces a radio-frequency modulation interferogram in which the depth information is encoded and can be detected by a CCD camera [28]. This field of ORFI is of growing importance [29].

18.5.3 Pseudo-noise modulation

The *pseudo-noise* (PN) modulation of the light signal combines several advantages [30, 31]:

i) quasi-stationary operation of the laser diode and of the receiver;

ii) peak correlation function of PN signals with multitarget detection capability, including a reference for calibration; and

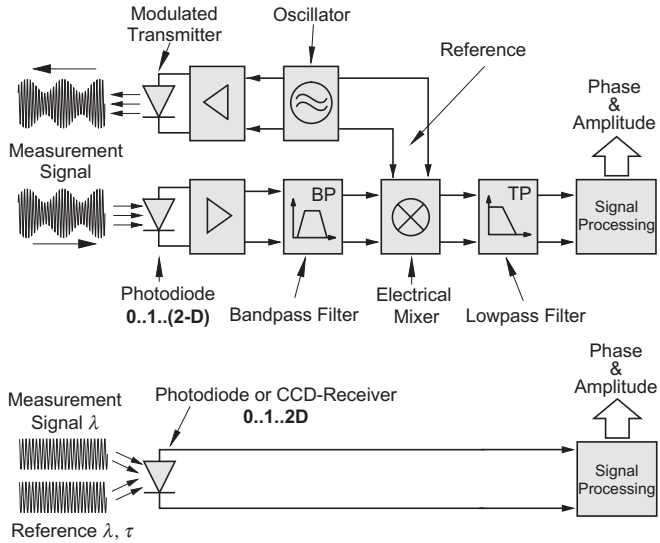


Figure 18.6: *Principle of optical depth measurements by:* ***a*** *incoherent (modulation); and* ***b*** *coherent (interferometry) time-of-flight measurements.*

iii) high distance resolution comparable to CW modulation of the same bitrate, but with a large, unambiguous measuring range due to the PN word length.

18.5.4 Comparison with interferometry

Since the modulation-based and interferometric depth measurements are based on the same TOF principles, the questions arises why modulation-based techniques already encounter significant measuring problems at resolutions in the centimeter range, while interferometric techniques can easily reach resolutions in the nanometer range.

The answer to these questions is illustrated in Fig. 18.6. A conventional TOF measurement according to Fig. 18.6a includes (besides the optical path) a considerable time delay in the high-frequency electronics *before* the signal is mixed and correlated. Especially the entrance amplifier and the electronic mixer give rise to such high errors in the temporal delay that either a continuous time-consuming mechanical calibration or a compensation by a costly second reference channel (not shown in the figure) is required [32]. In practice, the latter reference channel eliminates the unsatisfying time drifts mentioned in the foregoing.

With an interferometric TOF measurement, the mixing and correlation of the signal and reference channel take place directly within the photodetector by coherent field superposition, practically without any

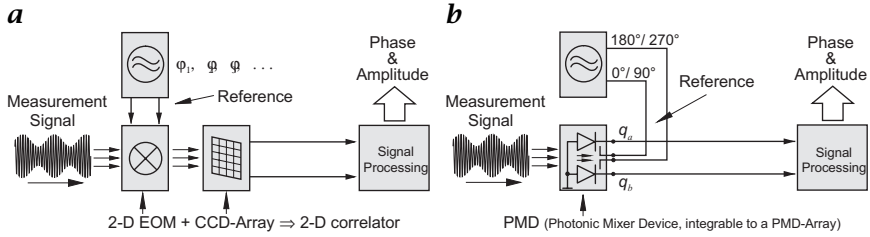


Figure 18.7: *Principle of optical depth measurements by:* ***a*** *3-D image acquisition; and* ***b*** *1-D time-of-flight measurement with the new PMD sensor.*

errors or temporal delay. As a direct consequence, this means that good TOF measurements require the high-frequency mixing process to take place not in the RF electronics but either in an optical component or in the detector itself. In this way, not only can the significant errors due to the time delays in the wideband amplifier, the electronic mixer, and cross-talk be avoided, but also the high costs associated with these components.

This principle has been realized in a new 3-D optical measuring system [28, 33, 34, 35]. Figure 18.7a shows the receiver part of a novel 3-D instrument with a 2-D optical mixer and a 2-D integrating detector. This mixer can be, for example, a Pockels cell that is integrated in the entrance pupil of the optical receiver. The mixing of the wavefront reflected back from the 3-D object with the periodic temporal modulation of received intensity generates a phase-dependent quasi-stationary intensity interferogram. This is integrated simultaneously and evaluated by each pixel of a CCD matrix sensor [10].

Intensity detection and mixing can also be achieved in a new inherent mixing detector in CCD technology [35] or in CMOS technology [36]. This so-called *photonic mixer device* (PMD) shown in Fig. 18.7b illustrates the simplification of a one-point receiver compared with a conventional receiver in Fig. 18.6a. The electrical symbol shown for CMOS technology attempts to describe the principle operation in a simplified way: The photocharge direction generated between the upper and the lower readout diode is controlled by a balanced modulation voltage applied to the upper and lower semitransparent photogate. The difference of the readout charge stands for an analogous correlation function of the optical signal and the modulation signals. A corresponding matrix of such PMDs can directly measure a depth image, as discussed in detail in [27]. This new device offers a high potential for optical sensory systems due to the simple and powerful procedure of multipixel detection, electro-optical mixing, and correlation. Obviously, it enables a new generation of powerful optical 3-D sensors providing real-time 3-D imaging facilities, particularly for applications in industrial automa-

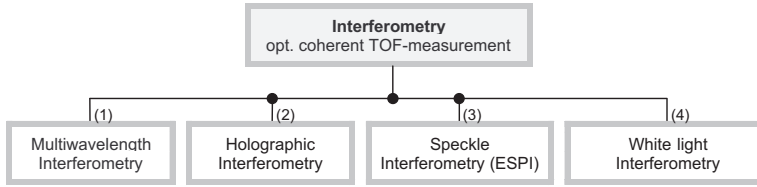


Figure 18.8: *Hierarchy of the most important measuring principles for depth measurements on the basis of optical interferometry.*

tion, security, navigation, robotics, etc. Currently, totally new TOF line cameras and TOF matrix cameras based on PMD technology are under development.

18.6 Optical Interferometry (OF)

Classical *interferometry* [2, p. 30] is a technique in which a coherent wavefront is split into a measuring (or object) and a reference wavefront. These are superimposed (correlated) again in a detector as illustrated in Fig. 18.6. If a 2-D detector is used, an interferogram or correlogram is generated, indicating the phase shift over the detector area. With at least three measurements with different phase positions of the reference, the phase shift between the reference and the signal wavefronts can be determined according to the phase-shift principle. Unfortunately, this technique cannot determine absolute depth. Because of the ambiguity of the signal in multiples of $\lambda/2$, a unique depth determination is only possible in this narrow depth range. With homodyne and heterodyne interferometry, a resolution better than $\lambda/100$ and $\lambda/1000$, respectively, can be reached. The high depth accuracy of interferometric measurements requires a mechanically very stable instrument.

A large number of different interferometric depth measuring systems with different measuring properties are currently available [37, p. 23]; [3, p. 66]. For practical applications in 3-D form measurements, several types of instruments are predominantly used and continuously improved (Fig. 18.8).

18.6.1 Multiwavelength interferometry

Multiwavelength interferometry (1 in Fig. 18.8) offers exceptional features for industrial applications. It is possible to perform *absolute* distance measurements over up to several ten meters with resolutions in the nanometer range under ideal conditions. Measurements can also be made from macroscopically rough surfaces. A basic characteristic

of multiwavelength interferometry is the generation of beat frequencies in the gigahertz and megahertz range by the superposition of two closely spaced wavelengths. The "synthetic" wavelengths of these beat frequencies determine (instead of the wavelength of the light itself) the range in which distances can be measured without ambiguity [3, 38].

18.6.2 Holographic interferometry

Holographic interferometry (2 in Fig. 18.8) enables deformation of 3-D objects caused, for example, by thermal or mechanical stress to be measured in the nanometer range. A hologram of the original object is coherently superimposed by that one under deformation. The resulting interferogram describes the deformation and can be captured or observed online, for example, by a video camera [2, p. 32].

18.6.3 Speckle interferometry

Speckle interferometry (3 in Fig. 18.8) utilizes an otherwise disturbing effect in optical metrology for exact deformation measurements. Speckles are generated when coherent light is reflected from a rough surface. The reflected wavefronts interfere with each other on the detector surface and generate a speckle pattern that is characteristic for the surface roughness elements. If an additional reference beam generates a second speckle pattern, this is coherently superimposed to the first one and produces a speckle interferogram.

Figure 18.9a shows the typical setup of a so-called electronic speckle interferometer (ESPI). After the object is deformed, a second speckle interferogram is captured. If this interferogram is subtracted from the previous interferogram of the original object, a difference interferogram is obtained as shown in Fig. 18.9b. The distance between the stripes corresponds to a height difference of $\lambda/2$. At least three exposures are required to obtain a difference height image $\Delta Z(x, y)$ as shown in Fig. 18.9c [2 S.34][31].

18.6.4 White-light interferometry

White-light interferometry or the *coherency radar* (4 in Fig. 18.8) has only a superficial similarity with focus-based depth measurements as discussed in Section 18.4.1. Because it is based on a time-of-flight measurement, it can take (in contrast to focus techniques) measurements with small apertures and, thus, be used, for example, to measure the depth of narrow boreholes. For a detailed description, see Section 19.5. The setup equals a Michelson interferometer. In one arm of the interferometer the object to be measured is located, and in the other arm a CCD camera is located. The basic idea of the coherency radar is that the

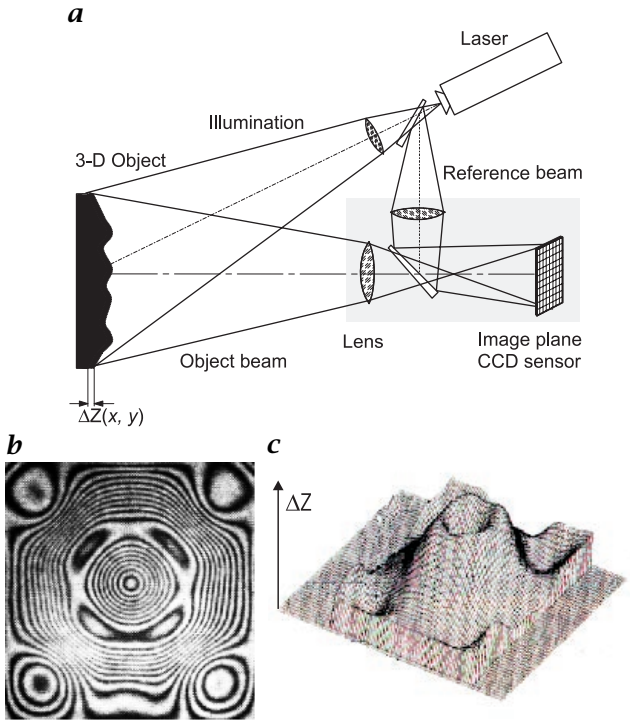


Figure 18.9: *Speckle interferometry:* ***a*** *Schematic illustration of the instrument setup;* ***b*** *Difference interferogram showing a form change;* ***c*** *3-D reconstruction.*

coherence length of broadband light is very short, for example, white light (depending on the nature of the light) only a few microns. Therefore, interference effects (rapid fluctuations of the signal with changing depth) are only observed when the path difference in the two arms approaches zero. In this way the depth can be measured [39, 40]. For a good depth resolution, a short coherence length l_c is required. The coherence length is inversely proportional to the signal bandwidth $\Delta\nu$, respectively, $\Delta\lambda$:

$$l_c = \frac{c}{\Delta\nu} = \frac{\lambda^2}{\Delta\lambda} \tag{18.7}$$

Thus, a light source with a large signal bandwidth is required. In an extension of this technique, a tunable broadband laser can be used. Then, without mechanical shifting, the virtual length of the reference illumination [41] can be obtained.

18.7 Outlook

The measuring technologies for 3-D optical form measurements have been in a phase of rapid development for a number of years. It is expected that this development will continue for some time to come. Better and new components, higher computing power, and faster and more accurate algorithms are on the horizon as well as the fusion of various depth-measuring principles.

Acknowledgments

We would like to thank our co-workers, especially Dipl.-Ing. X. Luan, A. Stadermann, W. Tai, W.-H. Twelsiek, R. Wurmbach, and Z. Zhang, as well as Dr.-Ing. J. Olk, and W. Kleuver for helpful discussions and their support in writing this chapter.

18.8 References

[1] Hauske, G., (1994). *Systhemtheorie der visuellen Wahrnehmung*. Stuttgart: Teubner Verlag.

[2] Breitmeier, U., Daum, W., Häusler, G., Heinrich, G., Küchel, M., Mollath, G., Nadeborn, W., Schlemmer, H., Schulze-Willbrenning, B., Schwarte, R., Seib, M., Sowa, P., and Steinbichler, R., (1995). Verfahren für die optische Formerfassung. In *Handbuch OF 1*, Vol. 38. Deutsche Gesellschaft für zerstörungsfreie Prüfung e.V. DGZfP, 1.HE11/95.

[3] Breuckmann, B., (1993). *Bildverarbeitung und optische Meßtechnik in der industriellen Praxis*. München: Franzis-Verlag.

[4] Engelhardt, K., (1992). *Methoden und Systeme der optischen 3D-Meßtechnik, Ingenieurvermessung '92, Beiträge zum XI. Internationalen Kurs für Ingenieurvermessung, Zürich*. Bonn: Ferd. Dümmler Verlag.

[5] Jiang, X. and Bunke, H., (1997). *Dreidimensionales Computersehen*. Berlin, Heidelberg, New York: Springer-Verlag.

[6] Küchel, M., (1995). Dreidimensionale Meßverfahren. In *Tagungsband Bildverarbeitung '95*, pp. 315–348. Technische Akademie Esslingen.

[7] Hein, A., (1998). *Verarbeitung von SAR-Daten unter besonderer Berücksichtigung interferometrischer Anwendungen*. Phd thesis, INV, Universiät-GH, Siegen.

[8] Fitch, J. P., (1995). *Synthetic Aperture Radar*. Springer.

[9] Leick, A., (1995). *GPS Satellite Surveying*. New York: John Wiley and Sons.

[10] Heinol, H. G., Xu, Z., Schwarte, R., and Ringbeck, T., (1997). Elektrooptische Korrelationseinheit großer Apertur zur schnellen 3D-Objektvermessung: Experimentelle Ergebnisse. In *Tagungsband Optische Formerfassung, DGZfP und VDI-GMA, Langen*.

[11] Schwarte, R., Xu, Z., Heinol, H. G., Olk, J., Klein, R., Buxbaum, B., Fischer, H., and Schulte, J., (1997). A new electrooptical Mixing and Correlating Sen-

sor: Facilities and Applications of this Photonic Mixer Device (PMD). In *SPIE-EOS: Sensors, Sensor Systems, and Sensor Data Processing, München*, Vol. 3100.

[12] Tiziani, H., (1996). Forschung. *Mitteilungen der DFG*, **4**:8–10.

[13] Engelhardt, K. and Häusler, G., (1988). Acquisition of 3-D data by focus sensing. *Appl. Opt.*, **27(22)**:4684.

[14] Engelhardt, K., (1991). Acquisition of 3-D data by focus sensing utilizing the moire effect of CCD cameras. *Appl. Opt.*, **30(11)**:1401.

[15] Breitmeier, U., (1993). Laserprofilometrie-Meßanlage für biomedizinische Fragestellungen. *Biomedizinische Technik*, pp. 35–45.

[16] Jurca, (1997). Firmenunterlagen der Fa. Jurca Optoelektronik, Rodgau.

[17] Dorsch, R. G., Häusler, G., and Herrmann, J. M., (1994). Laser triangulation: fundamental uncertainty in distance measurement. *Applied Optics*, **33(7)**.

[18] Pfeiffer, T. and Sowa, P., (1994). Optoelektronische Meßverfahren sichern die Produktqualität. In *Tagungsband Optisches Messen von Länge und Gestalt, GMA-Bericht 23*. Düsseldorf: VDI.

[19] Klicker, J., (1992). *Ein zweidimensionales Triangulationsmeßsystem mit Online-Meßwertverarbeitung bei hoher Bildrate*. Phd thesis, ZESS, Universität-GH Siegen.

[20] Vitronic, (1997). Firmenunterlagen (Viro-3D) der Fa. Vitronic GmbH, Wiesbaden.

[21] Breuckmann, B., (1996). Grundlagen der bildgebenden optischen 3D-Meßtechnik. In *Tagungsunterlagen Aktuelle Entwicklungen und industrieller Einsatz der Bildverarbeitung*, p. 247. Aachen: MIT Management.

[22] Seib, M. and Höfler, H., (1990). Überblick über die verschiedenen Moiré-Techniken. *Vision & Voice-Magazine*, **4(2)**.

[23] Wolf, H., (1996). Aufrüstung von 2D-Bildverarbeitungssystemen zu 3D-Systemen mit aktiver strukturierter Beleuchtung. In *Tagungsunterlagen Aktuelle Entwicklungen und industrieller Einsatz der Bildverarbeitung*, p. 1. Aachen: MIT Management.

[24] Schubert, E., (1996). *Mehrfachfarbcodierte Triangulationsverfahren zur topometrischen Erfassung und Vermessung von 3D-Objekten*. PhD thesis, ZESS, Universität-GH Siegen.

[25] Grün, A., (1995). High Accuracy Object Reconstruction With Least Squares Matching. In *Tagungsband Bildverarbeitung 95 an der Technischen Akademie Esslingen*, pp. 277–296.

[26] Schwarte, R., Hartmann, K., Klein, R., and Olk, J., (1994). *Neue Konzepte für die industrielle 3D-Objektvermessung nach dem Laufzeitverfahren*. PhD thesis, Düsseldorf.

[27] Schwarte, R., Heinol, H. G., Xu, Z., Olk, J., and Tai, W., (1997). Schnelle und einfache 3D-Formerfassung mit einem neuartigen Korrelations-Photodetektor-Array. In *Tagungsband Optische Formerfassung, DGZfP und VDI-GMA, Langen, 1997*.

[28] Heinol, H. G., Schwarte, R., Xu, Z., Neuhaus, H., and Lange, R., (1996). First Experimental Results of a New 3D-Vision System Based on RF-Modulation

Interferometry. In *Kongreßband OPTO96-Optische Sensorik Meßtechnik Elektronik*. Leipzig, Germany: AMA Fachverband für Sensorik.

[29] Yu, Z., (1998). *Investigation of a 3D-Imaging System based on ORF-Modulation*. PhD thesis, INV, Universität-GH, Siegen.

[30] Klein, R., (1993). *Ein laseroptisches Entfernungsmeßverfahren mit frequenzvaribler Pseudo-Noise-Modulation*. Dissertation, INV, Universität-GH Siegen.

[31] Schwarte, R., Heinol, H. G., Xu, Z., Li, J., and Buxbaum, B., (1997). Pseudo/Noise (PN)-Laser Radar without Scanner for Extremly Fast 3D-Imaging and -Navigation. In *MIOP97-Microwaves and Optronics, Sindelfingen*.

[32] Olk, J., (1997). *Untersuchung von Laufzeitentfernungsmeßsystemen unter besonderer Berücksichtigung des Referenzproblems*. PhD thesis, INV, Universität-GH Siegen.

[33] Schwarte, R., Hartmann, K., Klein, R., and Olk, J., (1994). Neue Konzepte für die industrielle 3D-Objektvermessung nach dem Laufzeitverfahren. In *Tagungsband Optisches Messen von Länge und Gestalt*. Düsseldorf: VDI-GMA and DGZfP.

[34] Schwarte, R., H.Heinol, and Xu., Z., (1995). A new fast, precise and flexible 3D-camera concept using RF-modulated and incoherent illumination. In *SENSOR95 Kongreßband, AMA Fachverband für Sensorik, Nürnberg*.

[35] Schwarte, R., Xu, Z., and Heinol, H. G., (1996). Large aperture optical modulators/demodulators for 3D-cameras. In *Kongreßband OPTO96-Optische Sensorik Meßtechnik Elektronik, AMA Fachverband für Sensorik, Leipzig*.

[36] Schwarte, R., Xu, Z., Heinol, H. G., Olk, J., and Buxbaum, B., (1997). New optical four-quadrant phase-detector integrated into a photogate array for small and precise 3D-cameras. In *SPIE-Multimedia Processing and Applications: Three-Dimensional Image Capture, San Jose*, Vol. 3023.

[37] Gruen, A. and Kahmen, H., (1993). *Optical 3D-Measurement Techniques II*. Karlsruhe: Wichmann-Verlag.

[38] Dändliker, R., Hug, K., Politch, J., and Zimmermann, E., (1995). High accuracy distance measurements with multiple-wavelength interferometry. *Optical Engineering*, **34(8)**:2407.

[39] Dresel, T., Häusler, G., and Venzke, H., (1992). Three-dimensional sensing of rough surfaces by coherence radar. *Applied Optics*, **31(7)**.

[40] de Groot, P. and Deck, L., (1995). Surface profiling by analysis of white-light interferograms in the spatial frequency domain. *Journal of Modern Optics*, **42(2)**:389–401.

[41] Tiziani, H., (1997). Flächenhafte absolute Speckle-Topometrie durch Wellenlängenvariation. Vortrag auf dem DFG-Kolloquium Automatische Sichtprüfung, Stuttgart, 18.2.97.

19 Three-Dimensional Sensors—Potentials and Limitations

Gerd Häusler

Lehrstuhl für Optik, Physikalisches Institut
Universität Erlangen-Nürnberg, Erlangen, Germany

19.1 Introduction

Most of the problems of *industrial inspection*, *reverse engineering*, and *virtual reality* require data about the geometrical shape of objects in 3-D space. Such 3-D data offer advantages over 2-D data: shape data are invariant against alteration of the illumination, soiling, and object motion. Unfortunately, those data are much more difficult to acquire than video data about the 2-D local reflectivity of objects. We will discuss the physics of *3-D sensing*, and will address the following subjects:

- different type of illumination (coherent or incoherent, structured or unstructured);
- interaction of light with matter (coherent or incoherent, at *rough surfaces* or at *smooth surfaces*); and

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

- the consequences of Heisenberg's *uncertainty relation*.

From the knowledge of the underlying physical principles that define the limitations of measuring uncertainty, one can design optimal sensors that work just at those limits, as well as judge available sensors. We will show that the vast number of known 3-D sensors are based on only three different principles. The three principles are different in terms of how the measuring uncertainty scales with the object distance. We will further learn that with only two or three different sensors a great majority of problems from automatic inspection or virtual reality can be solved.

We will not explain many sensors in detail (those explanations can be found in the references); instead, we will discuss the potentials and limitations of the major sensor principles for the physicist, as well as for the benefit of the user of 3-D sensors:

- *laser triangulation*;
- *phase measuring triangulation*; and
- *white-light interferometry* on rough surfaces.

As mentioned, it turns out that with those sensors 3-D data of objects of different kind or material can be acquired. The measuring uncertainty ranges from about one nanometer to a few millimeters, depending on the principle and the measuring range.

We will give examples of the potentials of each sensor by using measured objects and discussing the physical and technological drawbacks. We will specifically address the interests of potential users of those sensors concerning the applicability to real problems.

19.2 Why three-dimensional sensors?

Those who investigate or apply computer vision have the long range goal of machines that can "see" in a technical or, less often, natural environment, without the aid of an expert. Our practical definition for "seeing" evolved from many years of problem-solving for industrial inspection: for those problems, "seeing" an object means:

- segmentation from other objects;
- localization in 6 degrees of freedom; and
- feature detection, and template matching.

If this can be done robust against:

- shift, rotation, scale variation, perspective distortion, and
- hidden parts, shading, and soiling,

Table 19.1: *Visual system vs technical system*

	visual system	technical system
optical hardware	−−	++
post processing	++	−

we can solve a majority of the problems that relate to industrial inspection, metrology, reverse engineering, and virtual reality. The German VDI/VDE Standard 2628, VDI/VDE Handbuch Meßtechnik [1], [2] gives an interesting overview on how to define problems in automatic optical inspection.

It is obvious that for many real world problems we cannot achieve the required robustness by just more intelligent image processing algorithms. Our approach is that the intelligent algorithms have to be supported by intelligent sensors. To explain this, let us look at Table 19.1, where we compare the visual system and a technical system in terms of its automatic inspection ability. It is important that we separate the system into the "preprocessing optical hardware" and the "post-processing computer hardware."

The visual optical system is very poor in terms of *space-bandwidth product*, the lateral resolution is high only in a field smaller than the angular diameter of the moon. We can see only radiance and color, but not phase or polarization. Thus we judged the visual optical hardware with "−−". We want to emphasize that a video image has about the same quality as a visual image (video cameras are built so as to satisfy the eye).

Modern optical technology is very different: the space-bandwidth product can be extremely large, we not only see radiance but can take advantage of phase, polarization, and coherence. For example, with a differential phase contrast microscope we can resolve chromosomic structures and with 3-D sensors we can resolve 10,000 different depth steps. We judged this category by a "++" (Table 19.1).

Looking at the post-processing power, the judgment is the other way round: the brain is still unbeatable; for the tasks defined in our definition of "seeing": we gave a "++", while computers are mildly judged with a "−", due to their serial processing structure.

Now we come to the suggestive point: if we combine a video camera with a computer, we put in series the tools in the main diagonal of our table (three minus signs). That is the wrong diagonal. Unfortunately, we cannot put in series the elements of the other diagonal (four plus signs), for "automatic" inspection. But we can take advantage of the high standard of optical sensors to support the weak computer post-

processing (right-hand column). And we will demonstrate that optical 3-D sensors can help to satisfy the forementioned requirements.

Three-dimensional sensors acquire data about the geometrical "shape" of the object in 3-D space. (Sometimes, for example in tomography, we are interested in "volume data" [see 3]). The "shape"—compared to the 2-D video image—is invariant against rotation and shift of the object, against alteration of the illumination, and against soiling. And the shape is that which a majority of applications need: measuring and inspection; localization and recognition; virtual reality. Using 3-D sensors, we can solve many problems from this list simply by using standard computers. This is possible mainly by relieving algorithms from the need to consider image appearance variations.

If that is so, then why do we not have optical 3-D sensors everywhere? There are two answers: nature introduces severe physical limits (still under investigation) to acquire remote 3-D information. And our anthropomorphic thinking convinces us to consider a video camera and computer as a "technical eye."

To understand the physical limits of 3-D information acquisition gives deep insight in the propagation of information in space, which is interesting by itself. Moreover, the knowledge of those limits enables us to build better sensors (and judge the sensors of the competitors). Eventually, we might even overcome those limits. How could this be possible? We ask the reader for patience and curiosity.

To summarize this section: a video camera and a computer are not like the eye and the brain. An optical 3-D sensor can relieve the computer to consider the variations of image appearance because the geometrical shape is invariant. Moreover, knowing geometrical shape is important for many applications. In order to build, select, or apply good optical 3-D sensors, we have to understand the potentials as well as the limitations of 3-D data acquisition. This will be explained in the next sections.

19.3 Some important questions about three-dimensional sensing

Range sensors can measure the distance of stars or measure the thickness of atomic layers on a crystal. What are the principles that give us access to 20 or more orders of magnitude? There is a confusingly large number of different sensors. Some of them are user friendly, as described in Häusler [2] and in Reference [4]. A scientific review was given by Besl [5]. Chapter 18 also gives a systematic overview and Chapter 20 discusses an industrial surface measuring system for reverse engineering.

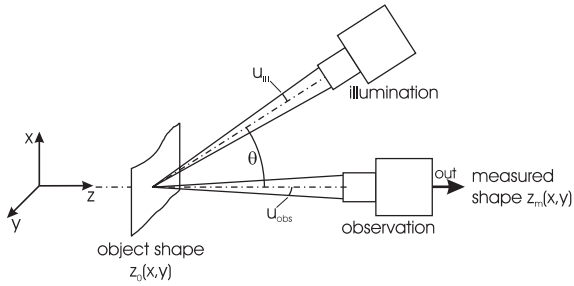


Figure 19.1: *Principle of triangulation.*

Can we put the large number of sensors into only a few categories and then judge their potentials and limitations by basic principles? We will discuss these questions in Sections 19.4 and 19.5. They summarize the results achieved by our group during the last several years.

19.4 Triangulation on optically rough surfaces

We will start with a very common method: *triangulation*. We want to acquire the shape of an object surface $z_o(x, y)$, as shown in Fig. 19.1. We will evaluate this shape by measuring the distance of one or many object pixels. To be distinguished from the real object shape, the measured shape will be called $z_m(x, y)$.

We will discuss here only active triangulation sensors: that is, there is a source that illuminates the object. Light interacts with the surface of the object. Light is reflected or scattered towards the sensor. The sensor has an aperture to gather the light, and an optical system to image each surface pixel onto a spatially resolving detector (array of photodetectors).

The *illumination* can be structured or diffuse. It can be spatially coherent or (partially) incoherent. It can be temporally coherent or broad-band. It can be polarized or unpolarized. Active triangulation needs structured illumination. Either a small light spot is projected onto the object (we call this a "point sensor" because it measures the distance of just one single point). Or we project a narrow line ("line sensor"; this method is known as *light sectioning* [see 6]). Or we project a grating (phase measuring triangulation [see 7, 8].

The object surface can be optically smooth like a mirror, or it can be optically rough like a ground glass. It is important to note that the attribute smooth or rough depends on the lateral resolution of the observation optics: If we resolve the lateral structure of a ground glass, for example, by a high aperture microscope, the surface is smooth for our purpose. "Smooth" means for us that the elementary waves that

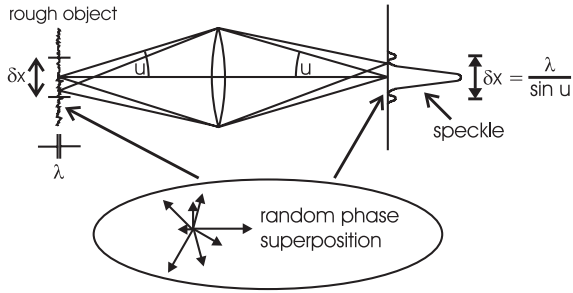


Figure 19.2: *For rough surfaces, the accumulation of complex elementary waves within a diffraction limited resolution cell of the sensor leads to the classical random walk problem, because the statistical phase differences are greater than $\pm\pi$. As a result, the lateral location of the spot image will suffer from an uncertainty. This lateral uncertainty leads to a fundamental distance uncertainty δz_m in triangulation systems.*

are collected from the object to form a diffraction limited image spot contribute only with minor phase variations of less than $\pm\lambda/4$, as shown in Fig. 19.2. If there are larger phase variations within the elementary waves then we have diffuse reflection, or scattering.

We will see diffuse reflection in connection with spatial coherence and introduce a statistical uncertainty of the measured shape $z_{\mathrm{m}}(x, y)$. The reason is that the wave scattered from an illuminated object "point" is no longer a spherical wave, but suffers from statistical phase variations. These phase errors lead to speckles in the image plane of the sensor [see 9].

For simplicity, let us consider a point sensor. It works by projecting a small light spot onto the object via a certain illumination direction, and observing this spot via a different angle of observation. The angle between these directions will be called "triangulation angle." A change of the distance will be encoded as a lateral shift of the spot image on the observing target. From lateral shift, we can evaluate the distance of the object point, or if we do that for many points, we acquire the shape of the object.

19.4.1 Physical limit of measuring uncertainty for triangulation at rough surfaces

The weakness of point triangulation is obvious: it is not robust against shape variation of the spot image. And just such a variation is introduced by speckle, as shown in Fig. 19.3.

As the shape of the spot image depends on the unknown *microtopology* of the surface, there will be a principal random localization error,

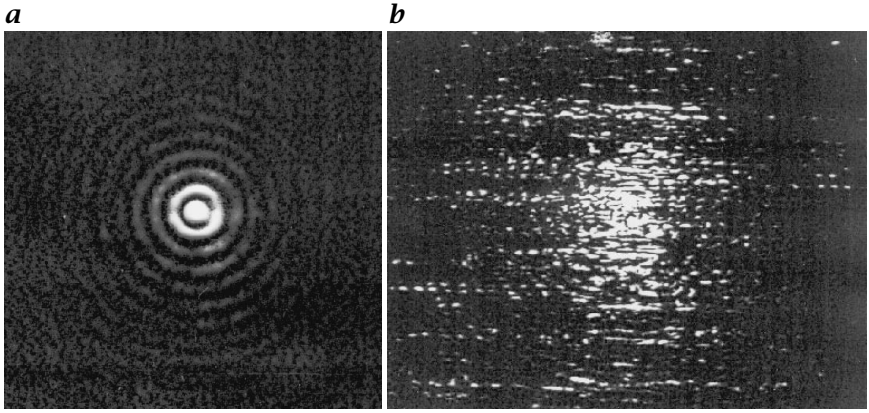


Figure 19.3: ***a*** *Spot image after reflection at a smooth surface;* ***b*** *spot image after reflection at a rough surface. The localization of the spot image is possible only with some uncertainty, introduced by the surface microtopology to which we have no access.*

theoretically and experimentally determined in Dorsch et al. [10]. Its standard deviation δz_{m} will be given by

$$\delta z_{\mathrm{m}} = \frac{c\lambda}{2\pi \sin u_{\mathrm{obs}} \sin \Theta} \tag{19.1}$$

With Fig. 19.1, θ is the angle of triangulation, $\sin u_{\mathrm{obs}}$ is the aperture of observation, λ is the wavelength of light, and c is the *speckle* contrast. The speckle contrast is unity for laser illumination. We have to emphasize that it is not the monochromaticity that causes speckle. It is the spatial coherence. And strong spatial coherence is always present, if the aperture of the illumination u_{ill} is smaller than the aperture of observation. With a small light source we can achieve high contrast speckles, even if the source emits white light! Hence, Eq. (19.1) is valid for phase measuring triangulation as well; we just have to use the correct speckle contrast, which is smaller than unity for properly designed PMT systems [see 11].

Equation (19.1) introduces a physical lower limit of the measuring uncertainty of triangulation sensors. To explain this let us measure a macroscopically planar ground glass with a surface roughness of 1 μm. We will use a sensor with an aperture of observation of 1/100, which is realistic for a macroscopic object, an angle of triangulation of 20°, and a wavelength of 0.8 μ, from laser illumination. Then we will find a standard deviation of the measured distance of about 37 μ, which is much larger than the surface roughness. Such a large statistical error is not acceptable, for many applications.

Can we somehow overcome the limitation? Before we start thinking about such a task we should be aware that we fight a deep physical principle: it turns out that the measuring uncertainty δz_m can be calculated by *Heisenberg's principle*. From the uncertainty δp_z of the photon impulse p_z in z-direction, introduced by the aperture of observation, we get an uncertainty δz_m of the measured distance z_m. This is a common physical fact. It surprises however, that we have an uncertainty as if we measure with only one single photon—although in fact we measure with billions of photons [see 12]. Obviously we cannot profit from statistical averaging over many photons. The reason is that each photon is in the same quantum mechanical state because of spatial coherence. This is discouraging because we probably cannot overcome quantum mechanics.

19.4.2 Triangulation beyond the coherent limit

Nature is difficult to overcome. The problem here is that we have to destroy spatial coherence! For a point sensor this can be done only at the object surface. Fig. 19.4 displays the result of an experiment that proves the importance of spatial coherence for distance uncertainty.

A different method of destroying spatial coherence is to heat up the surface and make it thermally radiant. This happens in laser material processing. We make use of the thermal radiation from the laser induced plasma, to measure the material wear on line, with very low aperture, through the laser beam, with an uncertainty of less than 5 μm [see 11].

As the two preceding possibilities are not generally applicable, the question arises as to whether we can reduce spatial coherence by illumination with a large source. This can be done principally for phase measuring triangulation. However, for practical reasons, the size of the illumination aperture can not be much larger than that of the observation aperture. Hence, there will always be a residual speckle contrast of $c = 0.1$ or more. Introducing this into Eq. (19.1), we will get a reduced measuring uncertainty [see 11]. It should be mentioned that triangulation is the inherent principle of so-called focus sensing as well. This principle is, for example used in the CD player or in the confocal scanning microscope [see 13]. Because illumination and observation are implemented through the same aperture u, in these systems Eq. (19.1) degenerates to

$$\delta z_m = \frac{c\lambda}{2\pi \sin u_{obs}^2} \tag{19.2}$$

which is, neglecting the factor $c/(2\pi)$, the classical Rayleigh depth of focus.

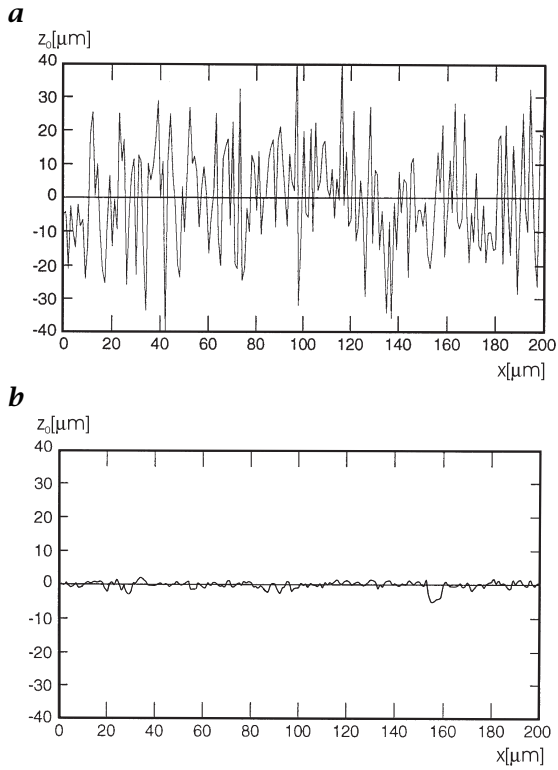


Figure 19.4: ***a*** *Observed image spot from a rough surface, measured with spatially coherent triangulation (laser illumination).* ***b*** *The same object measured in fluorescent light: the surfaces were covered with a very thin fluorescent film. Because* fluorescence *is perfectly incoherent, the noise is dramatically reduced. This experiment proves the big role of spatial coherence as a limiting factor in triangulation.*

19.4.3 More triangulation drawbacks

We have seen that triangulation systems of different kinds cannot overcome the coherent noise limit given by Eq. (19.1). This needs additional remarks: Triangulation usually does not even reach the physical limit on real technical surfaces, because the microtopology of the milling or turning process causes errors much larger than that of nice ground surfaces. In Fig. 19.5 we see the measuring uncertainty for laser triangulation and different angles of triangulation: the theoretical value, the experimental value achieved on an "ideal" surface (gypsum), and on a technical surface (scraped and hardened).

The reason is again the sensitivity of triangulation against shape alterations of the spot image. For real triangulation sensors that can

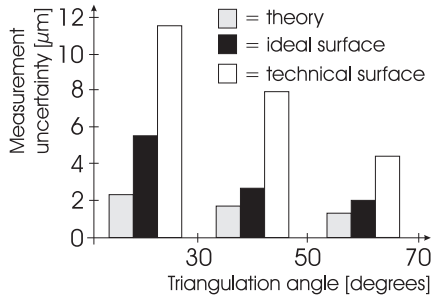


Figure 19.5: *Measuring uncertainties for laser triangulation with different angles of triangulation. Gray bar: theoretical value according to Eq. (19.1). Black bar: ideal surface (gypsum). White bar: technical surface (scraped and hardened).*

measure macroscopic objects, it turns out that, in practice, we cannot get a better uncertainty than about 5 μm—the micrometer regime was until now the domain of coordinate measuring machines (CMM), which utilize a mechanical probe (and which does not suffer from coherent noise). Although CMM technology is slow and the object is touched, it is still the technology of choice to measure large objects with about 1 μm uncertainty.

A further drawback is that in triangulation, illumination and observation are not coaxial. Hence, we cannot avoid shading: some parts of the object are either not illuminated or cannot be seen by the observation system.

For many applications, 3-D laser scanners are still used. The system we developed (Häusler and Heckel [6]) is still used for scanning sculptures [see Chapter 17].

19.4.4 A "real-time" three-dimensional camera with phase measuring triangulation

In spite of the drawbacks discussed here, the reduced speckle noise, together with its technical and conceptual robustness, made phase measuring triangulation technically and commercially successful during the last years. Some of the commercially available systems are described in Reference [4]. We will describe in this section one more system developed by our group. We call it "real-time 3-D camera," because it has the potential to supply 3-D data within one video cycle (40 ms, in CCIR standard), or even faster. The system is described in Häusler et al. [14].

There are two basic tricks that make the system accurate and fast. A major difficulty for PMT systems is to project a perfect sinusoidal pattern onto the object. Usually, Ronchi gratings are projected and the

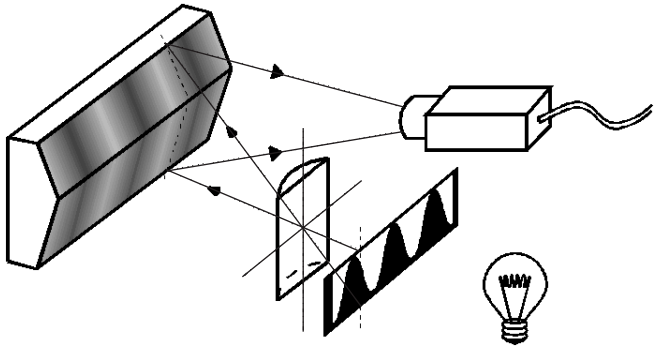


Figure 19.6: *Principle of astigmatic projection for phase measuring triangulation.*

sinus will be "somehow" created by defocusing, thus creating higher harmonics on the measured shape. However, it is possible to project a perfect sinusoidal pattern even with a binary mask, by using an astigmatic projection lens system [see 11, 15] and Fig. 19.6.

High accuracy needs a precise *phase shift* between the (usually 4 or 5) exposures. This requires expensive mechanical stages or an LCD projection. The system developed in Erlangen does not use mechanical shift, but an addressable ferroelectric liquid crystal display (FLC). As illustrated in Fig. 19.7, with only 4 electrodes we can generate 4 perfect sinusoidal patterns, each with the proper phase shift. The patterns can be switched within microseconds, due to the fast FLC. The "3-D real-time camera" that is realized by these tricks, is shown in Fig. 19.8. Phase measuring triangulation systems can easily be scaled, to measure within a human mouth, or even an entire human body. Our system is well adapted to measuring humans (see Fig. 19.9) or of being applied in a production line because it is quite fast.

It should be noted that in the case of uncooperative objects, if the SNR on the camera is low, sinusoidal patterns are not optimal. For these applications, binary patterns are appropriate, as described by Malz [8].

19.5 White-light interferometry on rough surfaces

Although triangulation systems are simple and often useful, their measuring uncertainty. A precise optical method, interferometry, is nonetheless unable to measure rough objects. Other, more successful 3-D sensing principles are needed to solve this problem.

With classical interferometry we can measure the height variations of a mirror in the range of nanometers, or less. This is possible as smooth objects do not introduce speckle noise. This suggests the ques-

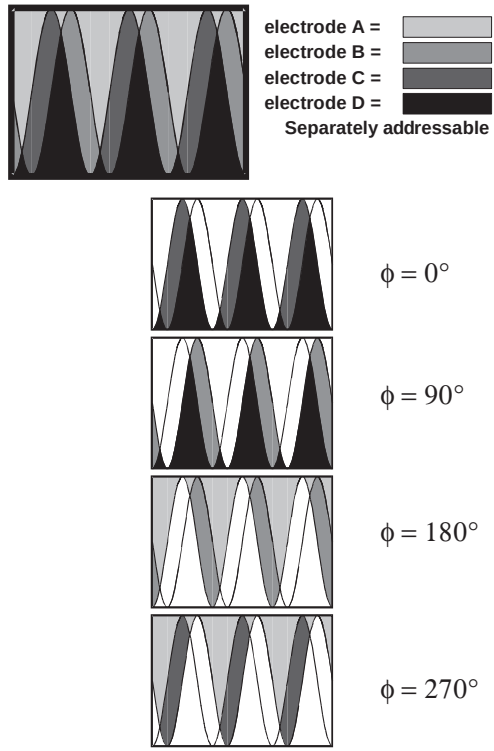


Figure 19.7: *With only four electrodes interlaced like a mosaic, we can generate four sinusoidal patterns with the proper phase shift of 90°.*

tion: What happens if we try to measure rough surfaces interferometrically?

Until recently, rough surface interferometry was not possible because the speckles in the image plane of the interferometer display an arbitrary phase, the phase within each speckle independent from the phase in other speckles [see 9]. Therefore, we cannot see fringes if we replace one mirror in a Michelson interferometer by the rough object. And it is useless to evaluate the phase of the interference contrast within each speckle. There is no useful information within that phase.

However, there are two major ways to measure rough surfaces with an uncertainty in the 1 μm regime [see 16].

In the first, the phase is *constant* within one speckle, allowing us to generate *interference contrast* in each speckle separately if *only* speckles are generated. This can be accomplished by using a sufficiently small *aperture of illumination* (as explained in preceding material)—even in the case of a white, extended light source.

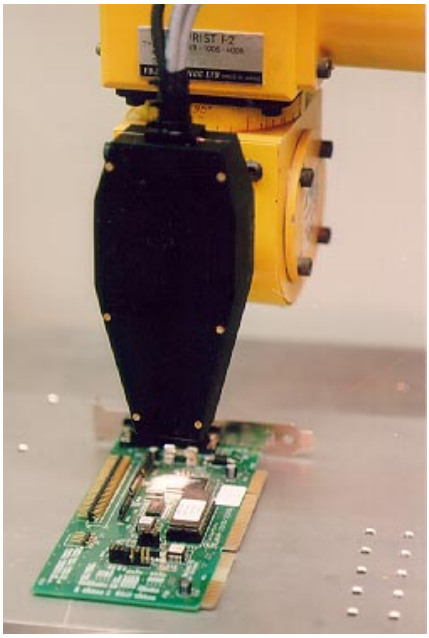


Figure 19.8: *"Real-time 3-D camera" for inspection of electronic boards.*

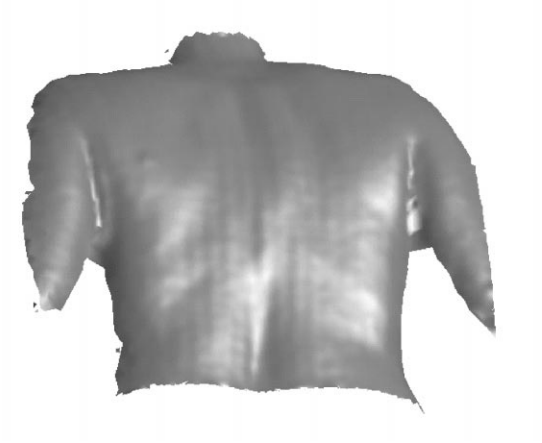


Figure 19.9: *Human body, measured by a "real-time 3-D camera."*

In the second, *broad band illumination* is used to exploit the limited *coherence length* of the light. It turns out that interference contrast can be observed only within those speckles that satisfy the equal path length condition: The path length in the object arm of the interferome-

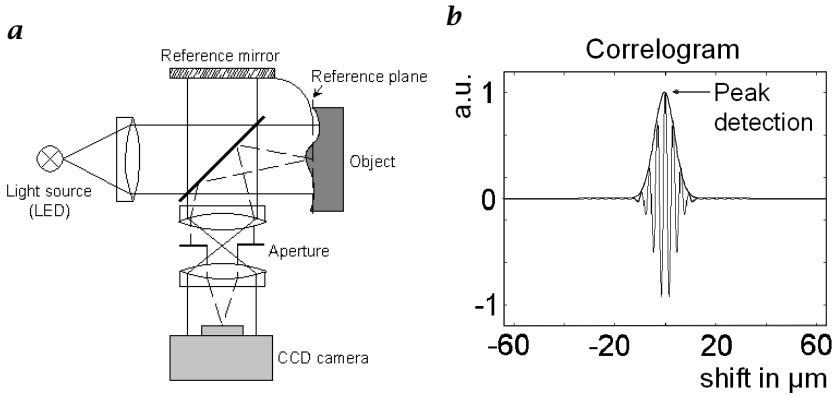


Figure 19.10: ***a*** *Principle of the "coherence radar."* ***b*** *The correlogram shows the (temporal) interference pattern in one single speckle while scanning the object along the z-axis.*

ter has to be approximately the same as that in the reference arm. For a certain object position on the z-axis, we will see interference contrast at one certain contour line of equal distance (or "height"). To acquire the shape of the object, we have to scan the distance (along the z-axis) (see Fig. 19.10).

(It should be noted that Michelson has already used white-light interferometry to measure the length of the (polished!) meter standard. New approaches enable us to measure optically rough surfaces as well.)

While scanning through the z-axis, each pixel of our observation system displays a modulated periodic time signal, which is called "correlogram." It is displayed in Fig. 19.10b. The length of this correlogram signal is about coherence length, and the time of occurrence, or the position $z_\mathrm{m}(x, y)$ of the scanning device at that time, is individual for each pixel: The correlogram has its maximum modulation, if the equal path length condition is satisfied. We store z_m for each pixel separately and find the shape of the surface.

White-light interferometry on rough surfaces, as it is realized in the coherence radar, is extremely powerful. There are unique features that will be summarized and illustrated by measuring examples:

- The coherence radar is a coaxial method: illumination and observation can be on the same axis. No shading occurs.
- The coherence radar is inherently telecentric, independently from the size of the object. All depths are imaged with the same scale.
- The distance measuring uncertainty on rough surfaces is not given by the apparatus or limited by the observation aperture. It is given only by the *roughness of the surface* itself.

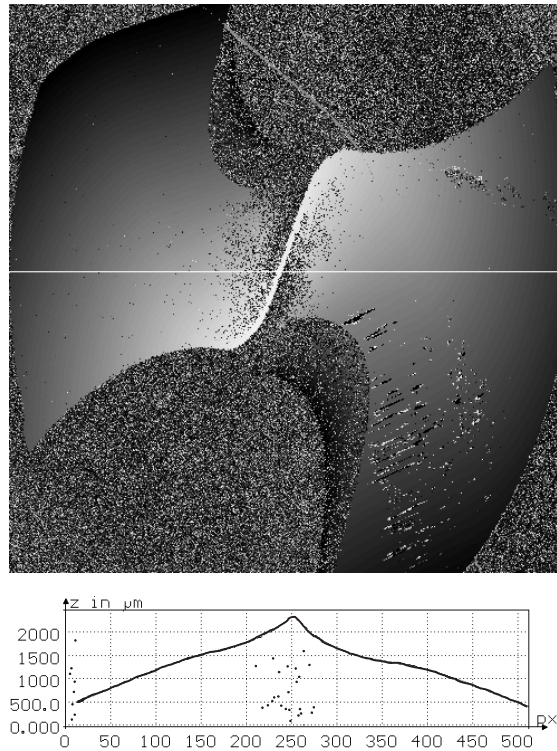


Figure 19.11: *Depth map of a drill, seen from the top, and cross section.*

This needs further explanation: Let us consider a macroscopically flat ground glass, where we do not resolve the lateral structure. The standard deviation of the ground glass surface is δz_0. The data $z_m(x, y)$ acquired by the coherence radar will display some statistical variation. Theory and experiments [see 17] show that the arithmetic value of the magnitude of these variations is equal to δz_0. This is surprising: We can measure the surface roughness although we cannot laterally resolve the microtopology. This remarkable feature is completely different from all other systems for roughness measurements [see 18].

Because the measuring uncertainty is independent of the aperture, it is independent of distance from objects (standoff), as well. Hence, we can measure distant objects with the same longitudinal accuracy as close objects. In particular, we can measure within deep boreholes without loss of accuracy.

There are modifications of the coherence radar principle that allow the measurement of objects much larger than the interferometer beam-splitter. This is possible by replacing the reference mirror by a groundglass, and utilize divergent illumination in both arms.

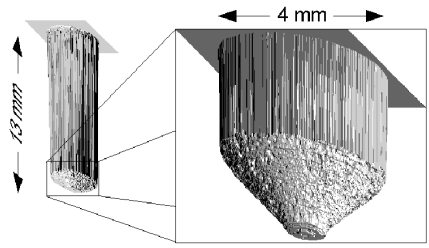


Figure 19.12: *As the measuring uncertainty of the coherence radar does not depend on the observation aperture, we can measure within deep boreholes, with about 1 µm accuracy.*

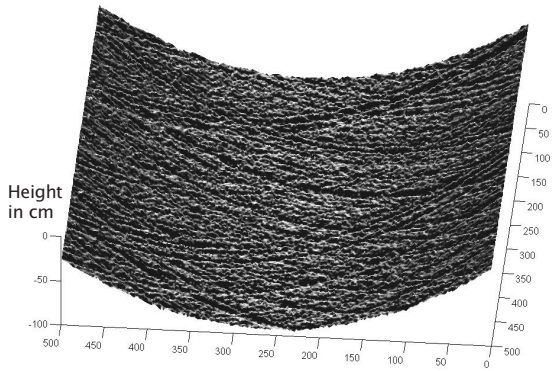


Figure 19.13: *Honed cylinder surface.*

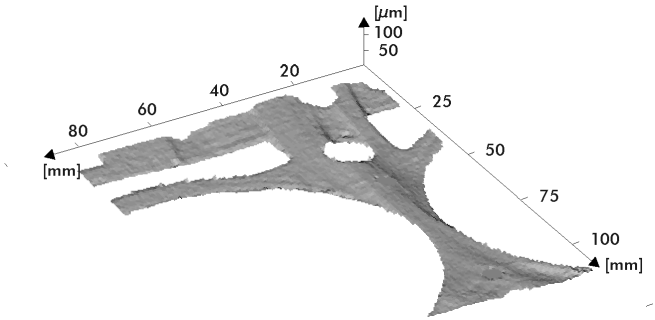


Figure 19.14: *Motorblock, measured by large field coherence radar. With the same method we can measure strongly curved polished surfaces [19].*

We can perform the shape comparison of an object against a master, by putting both object and master into the arms of the interferometer.

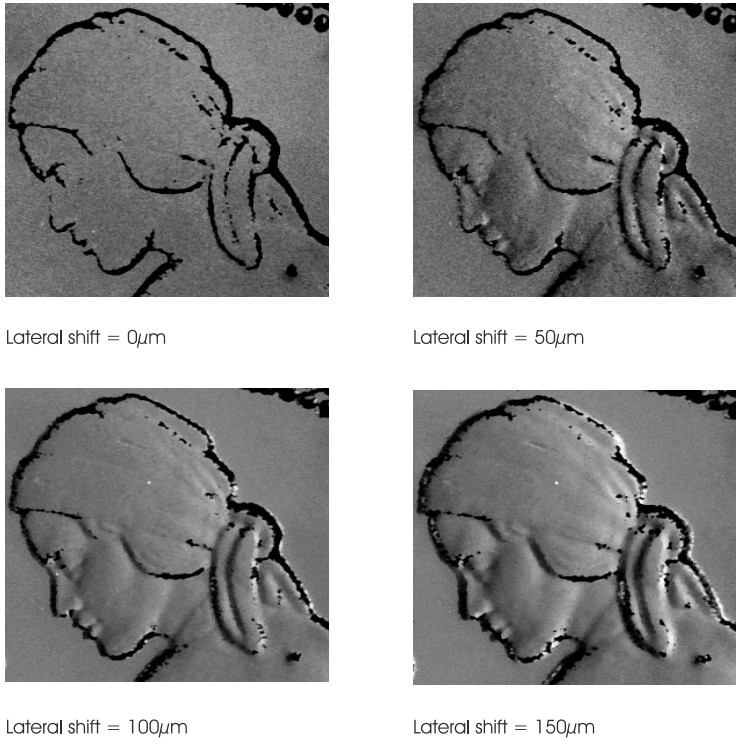


Figure 19.15: *Shape difference of two coins (of different age).*

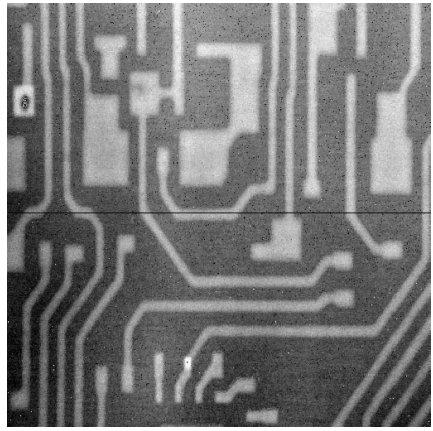


Figure 19.16: *Depth map of a hybrid circuit (metal on ceramics), measured by coherence radar. The metal lines have a height of 8 μm.*

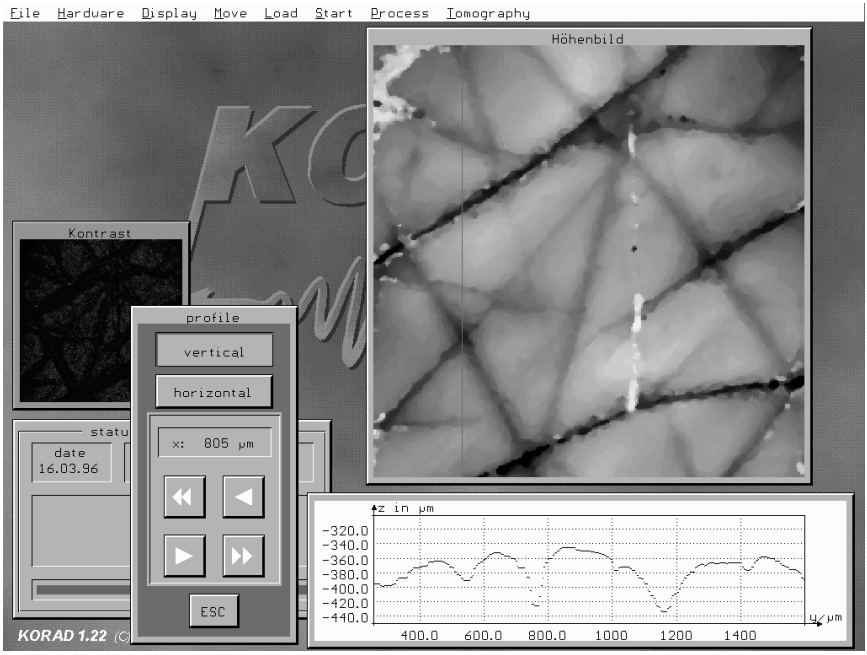


Figure 19.17: *In vivo measurement of human skin with coherence radar.*

The coherence radar supplies only the difference of the shapes, provided master and object are well adjusted.

We can even measure very small deformations in the nanometer regime, provided there is no decorrelation of the speckle patterns before and after deformation. In the case of very large deformation, speckles will be decorrelated. But we still can determine such a deformation absolutely, with a standard deviation slightly larger than the surface roughness. These features cannot be delivered by standard speckle interferometry or ESPI [see 18].

One more feature that cannot be achieved by triangulation is the ability of the coherence radar to measure translucent objects such as ceramics, paint or even skin. The reason is that we measure essentially the time of flight (with the reference wave as a clock). Thus, we can distinguish light scattered from the surface from light scattered from the bulk of the object.

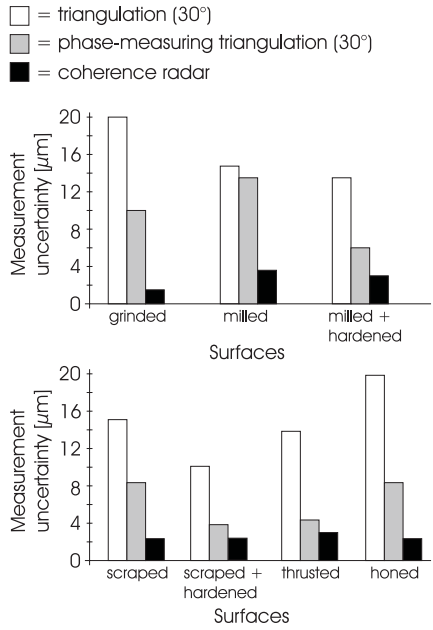


Figure 19.18: *Measuring uncertainty, achieved for different sensors and different surface structure.*

19.6 Summary of the potentials and limitations of optical three-dimensional sensors

Up to this, we have discussed triangulation and white-light interferometry in terms of potentials and physical limitations and the consequences for application. Of course, it is not only the physical limitations that are important for real problems. There are other conditions such as measuring time, size of the sensor, *standoff*, costs, availability, sensitivity against temperature, etc. Although these features cannot be discussed here, this section will summarize the results of Häusler [20] with the goal of providing the reader ideas as to how to select the right sensor for a given problem.

Triangulation and white-light interferometry at rough surfaces are based on completely different physical principles. This can be expressed by the fact that the measuring uncertainty δz_{m} scales differently with the standoff distance z_{off}. For triangulation, δz_{m} scales with z_{off}^2. We call this "type I." For the coherence radar, δz does not scale at all with the standoff.

It remains to mention that *classical interferometry* at smooth surfaces has a third type (type III) of scaling behavior [see 20]. It features

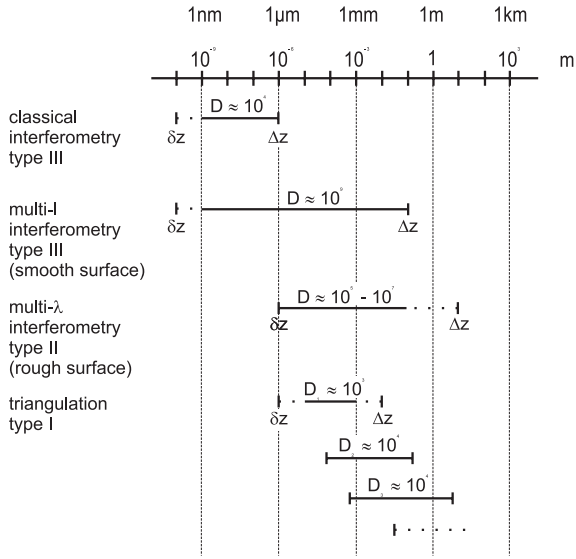


Figure 19.19: *The diagram displays the achievable measuring uncertainty δz for different sensor principles (triangulation (type I), coherence radar (type II), and classical white-light interferometry on smooth surfaces (type III)). Δz is the* measuring range *and D is the* dynamic range *($D = \frac{\Delta z}{\delta z}$). It should be noted that the gap between interferometry on smooth surfaces and triangulation, between 1 µm and 5 µm, can be closed by white-light interferometry.*

optical averaging over the microtopology: δz_m is proportional to the standoff:

- **type I:** $\delta z_\mathrm{m} \sim z_\mathrm{off}^2$;
- **type II:** δz_m independent from z_off (coherence radar); and
- **type III:** $\delta z_\mathrm{m} \sim z_\mathrm{off}^{-1}$ (classical interferometry).

We can compress the results in the diagrams of both Fig. 19.18 and Fig. 19.19

19.7 Conclusion

Designing good optical sensors requires an understanding of physical limits. Properly designed, optical 3-D sensors supply accurate data about the geometrical shape of objects that are as accurate as is possible vis-a-vis physics. The *dynamic range* allows researchers to distinguish 1000-10,000 different depths. Two main sensor principles, active triangulation and white-light interferometry at rough surfaces (“coherence radar”) can measure a majority of objects with different surface

structures. Once acquired, the geometrical shape complements intelligent algorithms very well in solving inspection problems (see, for example, Chapter 17), because the algorithms do not have to care about the variable appearance of objects, as is the case in 2-D image processing.

Acknowledgment

This chapter condensed the results of many years of research and of many people, to whom I owe thank you for their valuable discussion, experiments, software support, and new ideas. To those not mentioned in the bibliography I apologize.

I mention with gratitude that nearly all projects from our group were funded by the DFG, the BMBF, and the Bayrische Forschungsstiftung, as well as by many companies.

19.8 References

[1] GMA (ed.), (1985). *VDI/VDE Handbuch Meßtechnik II*, Vol. VDI/VDE 2628. VDI/VDE Gesellschaft Meß- und Automatisierungstechnik (GMA).

[2] Häusler, G., (1997). Möglichkeiten und Grenzen optischer 3D-Sensoren in der industriellen Praxis. In Koch, Rupprecht, Toedter, and Häusler (eds.), *Optische Meßtechnik an diffus reflektierenden Medien*. Renningen-Malmsheim, Germany: Expert-Verlag.

[3] Häusler, G. and Lindner, M., (1998). Coherence radar and spectral radar – new tools for dermatological diagnosis. *J. of Biomed. Optics*, **3**:21–31.

[4] Deutsche Gesellschaft für zerstörungsfreie Materialprüfung e. V. (ed.), (1995). *Handbuch OF1: Verfahren für die optische Formerfassung*. Deutsche Gesellschaft für zerstörungsfreie Materialprüfung e. V.

[5] Besl, P. J., (1988). Active, optical range imaging sensors. *Machine Vision and Application*, **1**:127–152.

[6] Häusler, G. and Heckel, W., (1988). Light sectioning with large depth and high resolution. *Applied Optics*, **27**:5165–5169.

[7] Halioua, M., Liu, H., and Srinivasan, V., (1984). Automated phase-measuring profilometry of 3-D diffuse objects. *Applied Optics*, **23**(18): 3105–3108.

[8] Malz, R. W., (1992). *Codierte Lichtstrukturen für 3-D-Meßtechnik und Inspektion*, Vol. 14 of *Berichte aus dem Institut für Technische Optik*. University of Stuttgart.

[9] Goodman, J. W., (1984). Statistical properties of laser speckle patterns. In Dainty, J. C. (ed.), *Laser Speckle and Related Phenomena*, p. 46 ff. Berlin: Springer-Verlag.

[10] Dorsch, R., Herrmann, J., and Häusler, G., (1994). Laser triangulation: fundamental uncertainty of distance measurement. *Applied Optics*, **33** (7):1306–1314.

[11] Häusler, G., Kreipl, S., Lampalzer, R., Schielzeth, A., and Spellenberg, B., (1997). New range sensors at the physical limit of measuring uncertainty. In *Proc. of the EOS, Topical Meeting on Optoelectronic Distance Measurements and Applications, Nantes, July 8–10*. Orsay, France: European Optical Society.

[12] Häusler, G. and Leuchs, G., (1997). Physikalische Grenzen der optischen Formerfassung mit Licht. *Physikalische Blätter*, **53**:417–421.

[13] Wilson, T. and Sheppard, C., (1994). *Theory and Practice of Scanning Microscopy*. London: Academic Press.

[14] Häusler, G., Hernanz, M. B., Lampalzer, R., and Schönfeld, H., (1997). 3D real time camera. In Jüptner, W. and Osten, W. (eds.), *Fringe '97*, 3rd International Workshop on Automatic Processing of Fringe Pattern. Berlin: Akademie Verlag.

[15] Gruber, M. and Häusler, G., (1992). Simple, robust and accurate phase-measuring triangulation. *Optik*, **89**(3):118–122.

[16] Dresel, T., Häusler, G., and Venzke, H., (1992). 3D sensing of rough surfaces by coherence radar. *Applied Optics*, **33**:919–925.

[17] Ettl, P., (1995). Studien zur hochgenauen Objektvermessung mit dem Kohärenzradar. Master's thesis, Friedrich Alexander University of Erlangen-Nünberg.

[18] Häusler, G., Ettl, P., Schmidt, B., Schenk, M., and Laszlo, I., (1998). Roughness parameters and surface deformation measured by coherence radar. In *Proc. SPIE*, Vol. 3407. Bellingham, Wash.: The International Society for Optical Engineering. Accepted.

[19] Häusler, G., Ammon, G., Andretzky, P., Blossey, S., Bohn, G., Ettl, P., Habermeier, H. P., Harand, B., Laszlo, I., and Schmidt, B., (1997). New modifications of the coherence radar. In Jüptner, W. and Osten, W. (eds.), *Fringe '97*, 3rd International Workshop on Automatic Processing of Fringe Pattern. Berlin: Akademie Verlag.

[20] Häusler, G., (1997). About the scaling behavior of optical range sensors. In Jüptner, W. and Osten, W. (eds.), *Fringe '97*, 3rd International Workshop on Automatic Processing of Fringe Pattern.

20 Three-Dimensional Sensors for High-Performance Surface Measurement in Reverse Engineering

Reinhard W. Malz

Forschungszentrum Daimler-Chrysler AG, Ulm, Germany

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

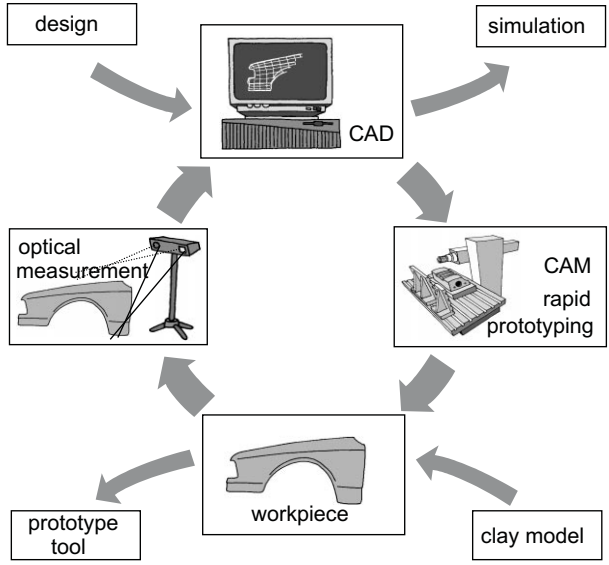


Figure 20.1: *Optical 3-D measurement in reverse engineering and product development processes.*

20.1 Introduction

20.1.1 Requirements for reverse engineering sensors

Digitizing 3-D objects is an important task in almost every stage of industrial product or tool life cycle. Optical measurement of *free-form surfaces* can be much faster than mechanical measurement. We expect that new techniques will speed up reverse engineering and quality control processes considerably (Fig. 20.1).

The following characteristics and properties of optical measuring systems are desirable for *reverse engineering* applications:

1. high spatial sampling rate, that is, acquisition of dense coordinate point clouds;
2. high signal dynamic, that is, working on unprepared object surfaces with different color and quality;
3. high speed, that is, fast acquisition of several millions of coordinate points in several minutes (not hours or more);
4. low costs compared to mechanical or mechano-optical coordinate measurement machines;
5. ease of use, for example, like a still or video camera on a tripod;

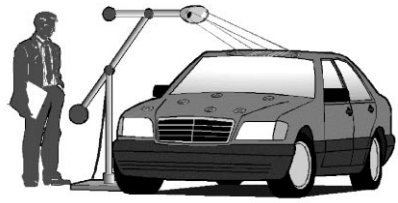


Figure 20.2: *Concept of a mobile "free-flying" 3-D sensor that can easily be used at different places in the factory.*

6. fast and easy sensor calibration without micrometer positioning equipment;
7. automatic sensor orientation, that is, transformation of coordinate point clouds measured from several views into one object coordinate system; and
8. semiautomated tools for fast data analysis and post-processing of the measured point clouds.

A system with these characteristics could be used as indicated in Fig. 20.2. The basic idea is to have a sensor that can easily be moved step by step around (or along) an object, while a computer processes the information from the sensor to generate a dense point cloud of coordinates representing the surface of the object. The user can see immediately the measured surface patches on the screen for planning the further measurement.

20.1.2 Optical sensor principles

Three-dimensional measuring principles have been developed and used in different scales (Fig. 20.3, Chapter 18).

Optomechanical scanning principles. Point measuring sensor principles (laser triangulation, time-of-flight, laser heterodyne interferometers) can be used in scanning mode for surface measurements (Fig. 20.4). As a major advantage compared to area-based sensors, parameter optimization is possible for every measured point. Dynamic control of lens focus, aperture, and signal amplification can, in principle, be used to overcome the physical limitations of fixed focus sensors, which need small apertures for a large depth of focus (see Chapter 19, Häusler [1]). For example, a 3-D laser sensor with dynamic focus, dynamic aperture, and modulated laser power has been developed for research purposes [2].

On the other hand, high-resolution digital matrix cameras and programmable light projectors will become less expensive in the future

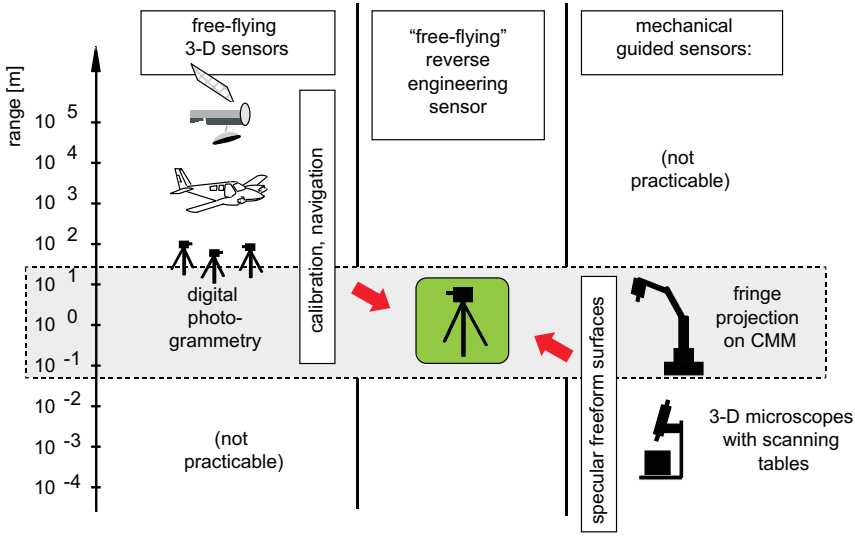


Figure 20.3: *Three-dimensional sensor applications in different scales.*

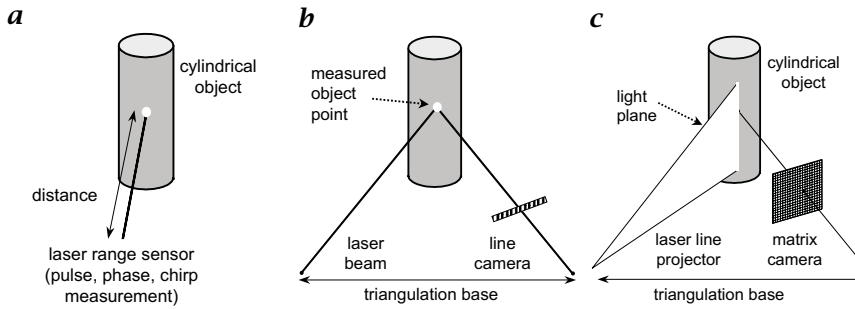


Figure 20.4: *Triangulation-based point and line sensors, which need precise scanning devices for surface measurement:* ***a*** *time-of-flight sensor;* ***b*** *point triangulation sensor;* ***c*** *light sectioning with a single light plane.*

compared to scanning devices with acceptable lateral accuracy, stability, and linearity. This is an important advantage of matrix camera-based sensors.

Photogrammetry. There are matrix camera based sensor principles that need no mechanical positioning. For example, the acquisition of 3-D data from multiple views with reconstructed camera positions has been realized in many passive remote sensing applications using natural textures and cooperative features [3]. In close-range applications, digital *photogrammetry* can measure some (10–1000) retroreflecting

target points with high precision (e. g., 1 : 100,000). However, it is not possible to obtain the dense point clouds required for realistic free-form surface descriptions and accurate CAD model reconstruction.

Active optical sensors. On the other hand, there are active optical sensor methods in medium- and small-scale applications that use well-defined artificial illumination in combination with high accuracy mechanical positioning systems (mechanically guided sensors in Fig. 20.3). For example, 3-D sensors consisting of a matrix camera and a coded light projector can measure a large number of coordinate points (approximately the number of camera pixels) from unprepared surfaces with good resolution (approximately one tenth of a pixel). However, to digitize complex 3-D objects it is necessary to move the sensor or the object while maintaining relative orientation information. In addition, the object area, which is measurable from a single view, can not, in practice be larger than approximately 1 m^2 because of limited projector light energy.

Sensor fusion for higher performance. The topological complexity and photometrical variety of measuring objects are a big challenge for optical sensor techniques. Because no sensor principle can solve all problems, a combination of several methods is necessary to create a multipurpose reverse engineering measuring system.

The following sections in this chapter describe camera-based principles for 3-D point and surface measurement, designed for industrial reverse engineering applications where typical sizes are centimeters to several meters. Their fundamental constraints and practical assumptions will be discussed. As an example, the "Ganymed" systems developed at Daimler-Chrysler will be explained. These sensors are based on several sensor principles, namely photogrammetry and active optical techniques with calibrated light projectors. The particular combination of principles used in the new sensor and its integration into a tripod mounted optical 3-D sensor will be shown. Finally, some applications and measurement results will illustrate its use.

20.2 Close-range photogrammetry

20.2.1 Target-based measurement

Principle. *Close-range photogrammetry* has developed into a highly reliable and precise measurement technique and is an efficient, economical and convenient tool for the measurement of point coordinates (Chapter 17; Volume 3, Chapter 16). The basic principle of photogrammetry can be described in five steps:

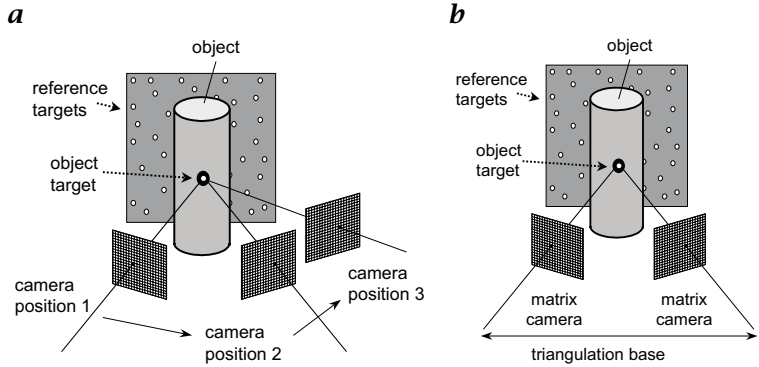


Figure 20.5: ***a*** *To measure static scenes and objects a single (still video) camera can be used for sequential image acquisition.* ***b*** *Static and dynamic scenes can be measured in snapshots with two or more synchronized cameras.*

1. Install targets and a scale bar as an absolute reference;
2. Take some pictures of the object from several views;
3. Find, label, and measure corresponding points in the images;
4. Combine all camera model and observation equations; and
5. Solve the numerical problem with photogrammetric (bundle adjustment) software, which calculates all unknown parameters for camera(s), positions and object points simultaneously.

This powerful and flexible concept can also be used with low-cost off-the-shelf cameras. The lenses and cameras have to be stable only for the short period of image acquisition (Fig. 20.5). As it will be shown later, these functionalities are important for mobile "free-flying" 3-D sensors.

In practical applications, multiple constraints can be used to speed up and stabilize the measuring process:

1. a stable camera lens and its model parameters from prior lens calibration;
2. a fixed lens-camera system and inner orientation parameters from prior camera calibration;
3. a stable multicamera setup and outer orientation parameters; and
4. known target coordinates in object space from prior measurements.

If one or more of these parameters are known and fixed, the remaining unknown parameters (e. g., the object coordinates) can be calculated based on less images or with higher reliability. This is an important feature for robust and self-checking systems.

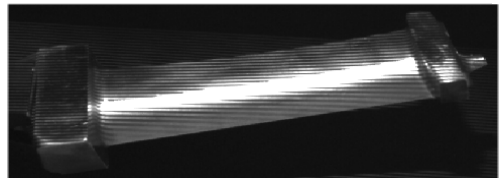


Figure 20.6: *Turbine blade (titanium) with specular reflections.*

Target points. A fundamental concept in photogrammetry is the intersection of rays in object space. The quality of intersection "points" determines the quality of measurement. In fact, these points can be represented by any spatial physical features or optical phenomena, provided that there are models which fit precisely independent from viewing angle. Temporal invariance is also required, if sequential imaging is used.

Good physical representations of points can be characteristic coordinates of geometrical primitives (like spheres, circles, squares, crosses etc.) or radiometric primitives (as edges between areas of different diffuse scattering materials). It is important that the borders between regions in the image are borders between different diffuse materials only.

In most close-range applications, the assumption of spatial and temporal invariance of features in several images is hardly to fulfill. Parts in industrial production processes are often made from a single material. Instead of color edges or other diffuse textures, technical surfaces have highly non-isotropic optical properties caused by machining processes. Shadows, jump edges and intensity values change when the light sources, objects and cameras are moving (Fig. 20.6). If the feature models are too complex, their parameters cannot be derived from a few images. Those surfaces cannot be measured directly with passive photogrammetric techniques (Fig. 20.6).

Artificial targets. To overcome these problems, well-designed targets are used to establish stable intersection points in the scene or on the object. With *retro-reflecting targets* or diffuse white flat or spherical targets of good optical quality (homogeneous and symmetrical intensity distribution) and sufficient size (5–10 pixels diameter) standard deviations of 1/20–1/50 pixels in the image plane can be achieved [4].

As will be shown later in Section 20.5, this point measurement principle is an important component of the "free-flying" sensor principle. The advantages of close-range photogrammetry with targets are

- high accuracy of target measurement, and
- short measuring time.

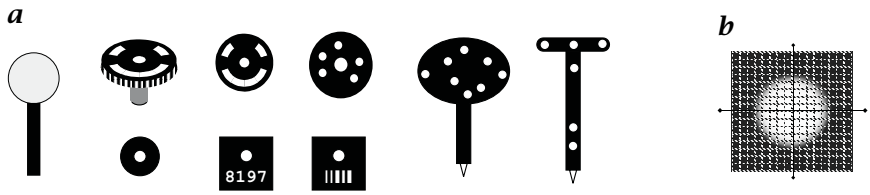


Figure 20.7: ***a*** *Different target types for point coordinate measurement with subpixel precision.* ***b*** *Subpixeling with circular targets needs extended image regions of interest (e. g., 16 × 16 pixels for the central blob).*

However, some disadvantages have to be accepted:

- the object has to be prepared and cleaned, which is a time consuming task before and after the measurement,
- the measured coordinate points are target coordinates, but not coordinates of the object itself,
- interesting object features like edges, corners or holes are discontinuities and cannot be prepared with standard targets,
- high densities of coordinates, which are necessary for the description of free-form surfaces cannot be achieved with targets, because there is always a need for extended image regions of interest for each point.

For good results, it is important to have a high quality signal chain. Simple errors on targets, illumination, optics, sensor chips, sensor electronics reduce the accuracy substantially. For example, mechanical and optical errors on targets can be:

- variable thickness of attached retro-reflective targets,
- enclosed particles and bubbles,
- dirty target surface and frayed target edges,
- virtual dislocations from inhomogeneous illumination.

Digitizing probes with passive and active targets. For interactive coordinate measurement, handheld probes with passive or active (laser, LED) targets are used. For each coordinate, the tip of the probe has to be placed at the intended object point. Then, a manually triggered synchronized image acquisition process with cameras from several views follows.

Cameras and probe have to be calibrated before measurement. The cameras can either be fixed and oriented once or moveable and oriented with every shot relative to a known reference target field (Fig. 20.5). As opposed to off-line photogrammetry, where all targets have to be fixed

before image acquisition, the selection of points with a digitizing probe is done during the measurement and points can be added if necessary.

On the other hand, the precision is somewhat lower because of poor redundancy in the data of independent measurements. The sampling rate is about 1 point per second, the variance approximately 1/20–1/2 pixel in the image plane.

20.2.2 Surface reconstruction with photogrammetry

Texture-based matching. Obviously, target-based measurement is good for a limited number of selected points. But how can we measure surfaces and produce point clouds with thousands of points? Higher spatial sampling rates can be achieved using textures on the object surface. To define and find homologue points from different views, these textures should be dense, high frequent and aperiodic to get unique and narrow correlation peaks for different scales.

Texture analysis needs correlation windows of sufficient sizes (typically 10–25 pixels diameter) to get stable and unique results with high precision. This reduces the lateral resolution and the available number of independent coordinate points.

The reader may try the following little experiment to understand the problems of texture analysis: Take three sheets of paper and cut a hole in the middle of each paper, one with 1 mm diameter, the second and third with 10 and 100 mm, respectively. Use the first mask to cover this text, look what you see through the hole, then move the paper randomly and try to find the same hole position again—nearly impossible. Then use the second mask. If you see a word, you can easily find the word again—but is it really the same word?

With the third mask the repositioning result seems to be stable and unique. But what happens when you lift, rotate and tilt the mask? You will run into problems again and get noisy results. But even hierarchical search using sentences and words to find identical letters and points in two images is not good enough for the measurement on rough or *specular surfaces*, where the textures look different from every view.

Natural and painted textures. Remote sensing applications need and use natural textures on the surface. Parts in industrial production processes, however, are often made from one material with low texture contrast. Such surfaces cannot be measured directly with passive photogrammetric techniques. Painted or printed diffuse textures would be optimal, but this kind of object manipulation would not be acceptable in most applications.

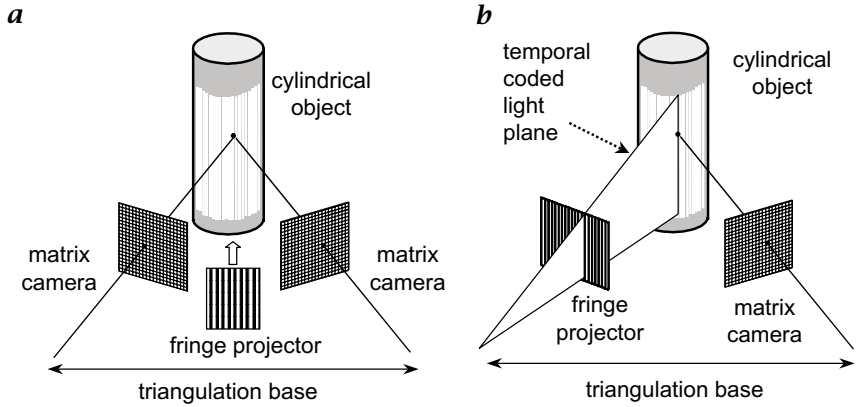


Figure 20.8: ***a*** *Stereo camera pair with texture projector.* ***b*** *Active-optical sensors with a calibrated camera-projector pair perform highest sampling rate and lateral resolution. Every camera pixel can, in principle, produce a separate coordinate.*

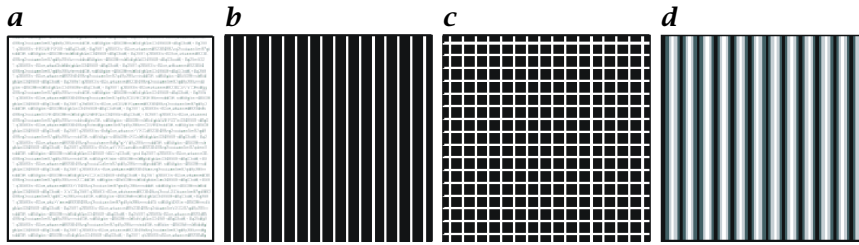


Figure 20.9: *Projection patterns, commonly used in two-camera photogrammetry:* ***a*** *pseudo noise;* ***b*** *line grid;* ***c*** *crossed lines;* ***d*** *sine grid.*

Projected textures. For materials which scatter light from the surface (and not from subsurface regions), light projection can be used to produce textures on the surface (Fig. 20.8a).

Figure 20.9 shows some examples for patterns, which are used to create textures. For example, the projection of pseudo noise (Fig. 20.9) is often used in digital photogrammetry with two or more views. It works quite well on smooth and non-specular surfaces. The lateral continuity of the surface is important, because the image processing needs neighboring pixel values to find the center of the spot, the center line(s) or the absolute phase of the sine grid [5].

Problems with specular surfaces. Technical surfaces have often highly non-isotropic optical properties. On *specular surfaces* as shown in (Fig. 20.10) or on transparent or volume scattering materials texture

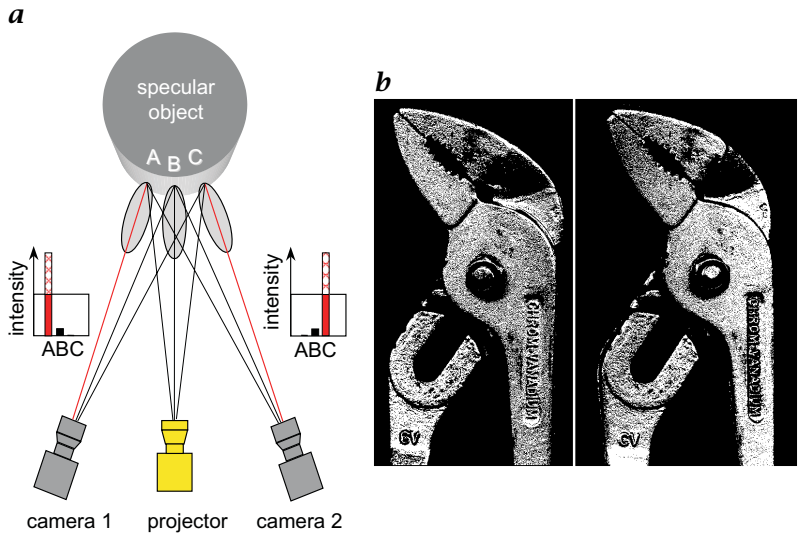


Figure 20.10: ***a*** *Problems with specular surfaces arise from extreme intensity variation. Different intensity patterns (ABC, left ≠ ABC, right) cannot be used for precise image correlation.* ***b*** *Each of the two images seems to have sufficient texture for image matching. But the textures are caused by specular effects and shadows and an image matching is nearly impossible.*

based matching doesn't work. Different camera views produce different image patterns (Figure 20.10). The results are not acceptable.

20.3 Sequential light processing and information theory

20.3.1 Calibrated light projection

All single pattern based measuring principles use the intensity distribution in small image areas. This lateral image processing assumes lateral continuity of light remission and topology on the object surface. Also, the lateral resolution of 3-D information will be reduced. This is not acceptable in industrial applications with non-diffuse and non-smooth object surfaces.

At least, only sensors which are based on local encoding/decoding principles instead of image correlation, can get results under such critical circumstances.

Local intensity transfer function. A fundamental task which has to be solved for active-optical 3-D measurement is the estimation of the projector "image" coordinates from the light as seen by a single camera pixel. However, the intensity, p, of a projector light beam scattered

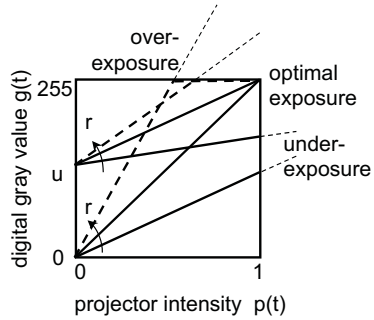


Figure 20.11: *A simple model with two parameters r and u for some typical intensity transfer functions.*

from the object surface into a camera pixel cannot be derived from a single gray value, g, in the digitized image.

If the camera characteristic is linear and the analog-digital converter clips at a maximum value $g_{\max}$, the digitized value g can be described by a simple model as

$$g = \min(u + rp, g_{\max}) \tag{20.1}$$

where u is an unknown offset value from environmental light and r is the unknown local intensity transfer factor mainly caused by local surface properties (Fig. 20.11).

We can estimate the local parameters u and r, if we accept that sensor and object must be in a fixed relative position, while the projector produces sequential patterns and the camera digitizes the image sequence (Fig. 20.12). This sequential concept is applicable in almost every industrial reverse engineering task. (However, it would become a problem in robotics, where a fast on-line feedback for pose-control is desirable.)

The common idea of all subsequent discussed principles is the implicit or explicit estimation of the parameters u, r and p from at least three gray values measured at different projector light patterns.

Normalization of the local intensity transfer function. The simplest principle (which is good for explanation) is shown in Fig. 20.12. A sequential projection of $p_0 = 0$ (black), $p_1 = 1$ (white) and $p_w(\zeta_p) = \zeta_p/\zeta_0$ (linear wedge from black to white) produces three different intensities g_0, g_1 and g_2 at every image coordinate. From the linear equation system

$$\begin{aligned} g_0 &= u + 0 \cdot r \\ g_1 &= u + 1 \cdot r \\ g_2 &= u + p_w \cdot r \end{aligned} \tag{20.2}$$

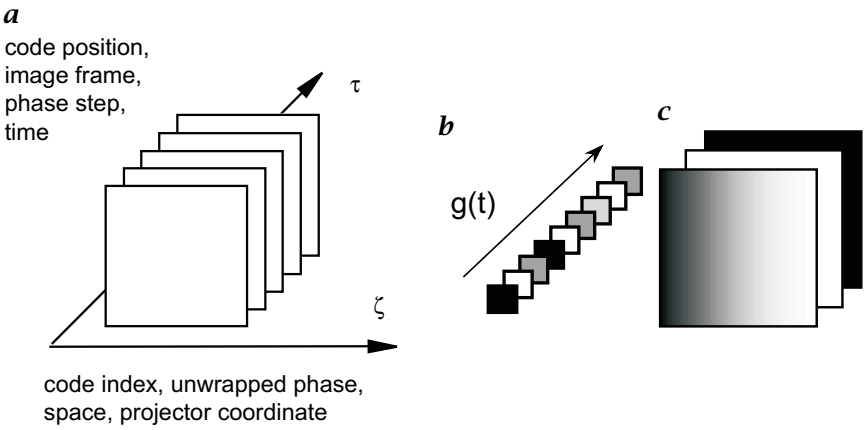


Figure 20.12: ***a*** *Temporal light encoding principle which results in local intensity vectors $g(t)$;* ***b*** *at every pixel position;* ***c*** *simple pattern sequence for calibrating the intensity transfer function (right).*

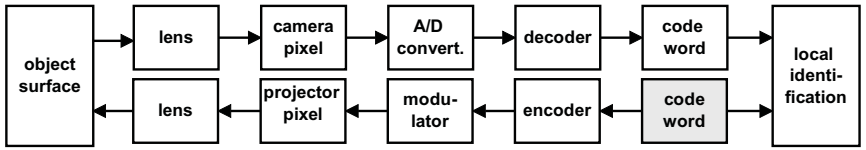


Figure 20.13: *Signal chain in the active-optical sensor.*

we can derive the decoder function and estimate the projector coordinate as

$$\zeta_p \approx \zeta_d = \zeta_0 \cdot (g_2 - g_0)/(g_1 - g_0) \tag{20.3}$$

Each camera pixel can produce a range value independent from neighbouring pixels, environmental light and object reflectivity. The range resolution of this simple concept is limited by the temporal noise in the image signal. This limit is still far below the limit given by spatial speckle noise described in Häusler [1, 6] and Chapter 19. Before we discuss several encoding principles which improve the range resolution a brief introduction to information theory is given.

20.3.2 Information theory

A basic task of *coding theory* is to maximize the information flow in a noisy communication channel. For optical 3-D measurement with projectors and matrix cameras, the projected temporal light patterns are transmitters of encoded projector coordinates. Each camera pixel is a

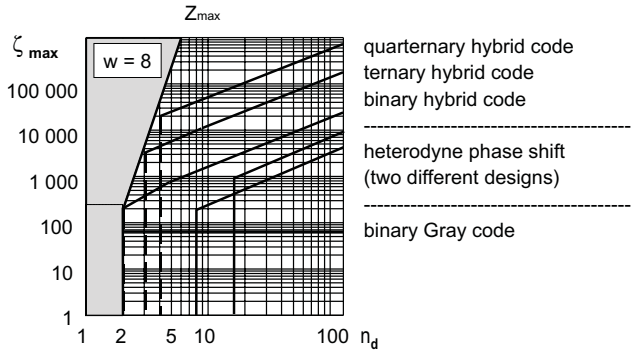


Figure 20.14: *Maximum resolution Z_{max} of projector coordinate ζ_p at a given number of separable gray levels for different coding principles with a code word length of w = 8 (gray values, images).*

receiver, which has to decode the sequential intensity signal (Fig. 20.13, [7]). The signal quality received at a pixel position is determined by typical channel parameters like gain, offset, signal-to-noise-ratio, nonlinearity and cross-talk from neighboring pixels. A phenomenological parameter which describes the sum effect in a linear channel is the number, n_d, of separable gray levels.

If we use such a communication channel model to describe the optical transmission of w projector intensities digitized in w images, the limits of information flow is described by the entropy, H_i, [7]:

$$H_i = \frac{1}{w} lb \left\{ {n_d}^w + 2 \cdot (n_d - 1)^w + (n_d - 2)^w \right\} \quad \text{bit per value} \qquad (20.4)$$

In a long sequence of values ($w \gg 3$) this entropy tends to the constant entropy, H_o, of a stationary channel [8]:

$$H_0 = \frac{1}{w} lb \left\{ {n_d}^w \right\} = lb\ n_d \quad \text{bit per value} \qquad (20.5)$$

For every given number of signal values, w, a theoretical limit, Z_{max}, can be calculated. It describes the maximum resolution of projector coordinate, ζ_p for a given number of separable gray values in the image sequence. In practice, however, the resolution of sequential light principles is below Z_{max}.

Figure 20.14 shows calculations of resolutions for different projection coding techniques. These functions always start at the minimal value, nd, that the coding principle was designed for (for example $n_d = 2$ for binary Gray code). If the channel is worse, the decoding (or the calculation of absolute phase) will fail. If the channel is better, the resolution increases. A pure digital code produces either a correct number or a large error.

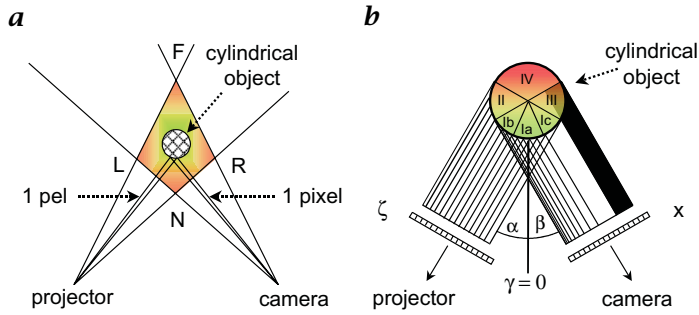


Figure 20.15: ***a*** *Variable spatial discretization.* ***b*** *Variable spatial frequency transfer function.*

The above discussion is for a single pixel position in an image sequence. However, in almost every application the conditions will vary from pixel to pixel. Because a decision for specific sensor parameters (code, camera shutter time, video signal amplification, ...) affects the whole sensor system, an optimization strategy should be based on knowledge about system, object and measurement task.

20.3.3 Signal limitations and basic errors

The spatial resolution of an active optical sensor is limited by several factors. There are fundamental physical limitations from spatial noise which is caused by speckle effects [1], [6], (Chapter 19) and technological factors like the number of camera pixels in x and y) and the number of separable projector stripes or pels. In addition, these effects are not constant. From Fig. 20.15, we see that in close-range applications, object-camera and object-projector distances can vary greatly, and this affects the lateral image and the longitudinal range resolution in a more geometrical sense. Throughout the measurement space there is a variety of voxel sizes and shapes. Voxels are more square near N, more rhomboid near F, and more rectangular near L and R. Diffraction and defocusing, as well as the variation of surface orientation relative to camera and projector axes lead to additional problems. The apertures of camera and projector lenses are rather poor compromises: they should be opened for higher contrast and signal-to-noise ratios at limited projector intensities. On the other hand they should be reduced for a wide depth of focus reaching from N to F.

Another critical effect is the variable frequency transfer from projector to camera. Figure 20.15b shows the variation in the projector-to-camera frequency transfer function for a cylindrical object. Only the regions Ia and b fulfill the Nyquist criteria (e. g., the sampling frequency must be at least twice the upper limit of the object spectrum). The in-

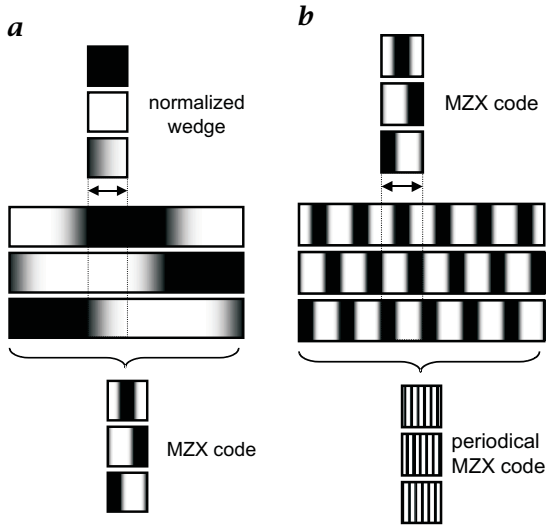


Figure 20.16: ***a*** *The permutation of a normalized wedge produces the MZX code with a six-times higher spatial resolution compared to a wedge.* ***b*** *The periodical repetition shows a pattern that is close to the phase shift principle with sine functions (0°, 120°, and 240°)*

creasing spatial frequency seen by the camera in region I_c leads to a strong undersampling and cross-talk between neighboring pixels and finally results in decoding errors. Regions II, III and IV are not measurable with this sensor position and have to be measured from different views.

In addition to optical resolution effects, the variation of surface orientation relative to camera and projector causes extreme intensity variations on non-diffuse surfaces. Even on a perfect lambertian surface, the camera sees lower intensities in region I_a. Finally, there remains a small measurable region, I_b, on the cylinder. In the center of the sensor's workspace, we find the best conditions for measurement.

20.3.4 Surface measurement based on sequential light

All effects mentioned above reduce the capacity (that is, the number of separable gray levels) of the local communication channel and have to be taken into account for a optimal code design.

MZX code. First, how can we improve the resolution of projector coordinate ζ_p in a three-image sequence? The first step of improvement is using all six permutations of the three patterns black, white and wedge (Fig. 20.16a). The result has the following properties:

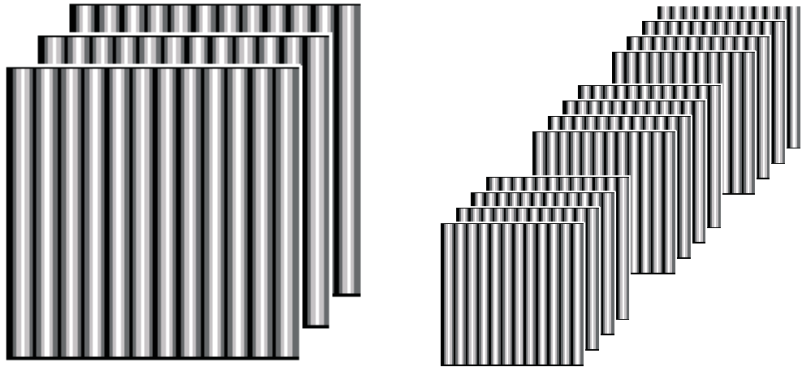


Figure 20.17: *Projected patterns with 0, 120, 240 degrees phase shift) and 3 groups of slightly different frequencies, each with phase steps of 0, 90, 180 and 270 degrees.*

1. it uses all combinations of three projector levels which fulfil the constraint that at least one value must be white and one must be black.
2. the Hamming distance is 1 for spatial continuity.
3. the spatial gradient of decoder output is constant and maximal.

This code was named MZX for Maximum level, Zero level, Crossover [9]). It will be used later for the construction of hybrid codes.

Phase shifting with a single-frequency sine pattern. A great variety of interferometrical *phase-shifting* techniques has been developed since the 1970s. Phase-calculating and phase-unwrapping algorithms can also be used in triangulation-based sensors where periodic patterns are projected [10, 11, 12, 13, 14].

The advantage of a set of phase shifted patterns compared to a single pattern is the same as described for MZX code: from three gray values that are measured at the same pixel position, a local phase can be evaluated that is independent from the lateral distribution of gray values.

This local phase value, which is always in the range $(0, 2\pi)$ can be seen as an absolute phase φ modulo 2π, where φ corresponds to the projector coordinate ζ_p. If the object surface is continuous, the absolute phase can be calculated by an incremental phase unwrapping algorithm, which allows no phase increments between neighboring pixels larger than $\pi/2$.

Table 20.1: *Basic projector patterns for phase shift*

Function	Projector principle (example)	Quality of phase output (3-D)
Rectangular	Programmable LCD projectors	Non-continuous, systematic errors
Trapezoid	Bucket integrating [7]	Best signal-to-noise ratio[1]
Sine	Interferometry, fixed mask projectors	Insensitive to varying MTF[2]
Triangular	Crossed rectangular grids [14]	Lowest signal-to-noise ratio[1]

[1] At full modulation without signal clipping (e.g., 0...255).
[2] The optoelectronic system's modulation transfer function.

All phase shifting principles (Table 20.1) normalize the local intensity transfer function and measure the local phase independent from the channel parameters u and r as shown in Fig. 20.12.

Phase shifting with two or more frequencies. To produce absolute and local phase information $\varphi(x, y)$ at non-continuous surfaces, multi-frequency (heterodyne) principles have been used in interferometry [13].

Independent phase shift measurements at slightly different light frequencies or wavelength (Fig. 20.17) lead to a singular absolute phase value. This principle has also been used for triangulation based sensors [14].

Gray codes. Binary *Gray codes* (Fig. 20.18) [15, 16] as well as multi-frequency phase-shift techniques with periodical and continuous patterns [14] have been widely used to acquire dense (that is, in principle for each camera pixel) and unique 3-D point data from objects in short range. To binarize the digitized images, it is necessary to know the local threshold (which may be different for each pixel). There are several ways of using additional images to calculate this threshold:

1. project unstructured gray with 50% intensity and use the acquired image as threshold, or
2. project unstructured white and black and use the averaged images as threshold, or
3. project both normal and inverse patterns, and use the sign (1,0) of the difference as bit.

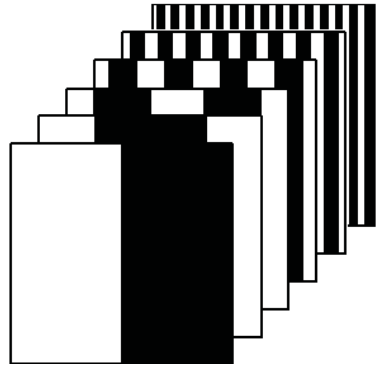


Figure 20.18: *Binary Gray code (additional images for threshold generation are not shown).*

20.3.5 Hybrid codes

As developed by Malz [9], *hybrid codes* combine the advantages of digital and analogue principles and yield results close to the theoretical limits of resolution, $Z_{\max}$, which can be achieved with temporally encoded light structures.

The trapezoid light distribution of the MZX subcode is continuous in space and intensity. Hybrid codes can be used with variable numbers of images ($w \geq 3$) and also with variable digital code bases (binary, ternary, quarternary Gray codes). It has the highest resolution compared to all other temporal principles (under equal conditions, namely the number of images used, and the lowest acceptable number of separable gray levels. See also Fig. 20.19).

20.3.6 Light stripe projectors

An important factor in the signal chain is the programmable light projector. The decoder result can only be linear and noiseless, if the spatial projector modulation is exact. Hybrid codes need analog projecting devices for best results. At least, the decoder function has to be strictly monotone with no steps.

Some technical light projectors, however, are not able to produce continuous sine or MZX-modulation. For example, a rectangular projection pattern used for a phase-shifting with 90° produces a step-by-step decoder function. This causes systematic errors of the detector signal in the range of $\pm\pi/4$ (Fig. 20.20).

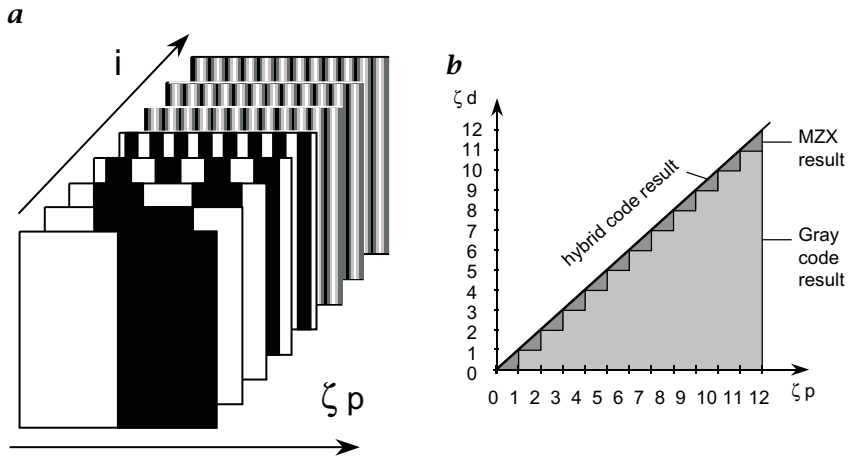


Figure 20.19: ***a*** *Hybrid code (binary Gray code, MZX code);* ***b*** *decoder function.*

20.4 Advanced self-calibration of three-dimensional sensors

High accuracy based on physical long-term stability has its price and its weight (e. g., conventional coordinate measuring machines). If the system calibration is under continuous software control, high accuracy can also be achieved with lightweight constructions, which have to be short-term stable for hours only. This is a big advantage for mobile systems.

20.4.1 Camera calibration

The calibration of matrix cameras is a standard procedure today and will not be described here (for details see Chapter 17 and [17]). Camera calibration can also be used for all sensor systems based on two or more fixed cameras in combination with targets or digitizing pens as shown in Fig. 20.5 or in combination with uncalibrated texture or stripe projectors as shown in Fig. 20.10 or in [18].

20.4.2 Projector calibration

To exploid the fundamental advantages of the one-camera-one-projector pair described in Section 20.3.4 it is necessary to calibrate the projector also as a second measure for the triangulation geometry. The calibration of the stripe projector itself, however, is more complicated for several reasons. Projectors have, in principle, a lot more error sources compared to cameras. So the standard camera models will not describe

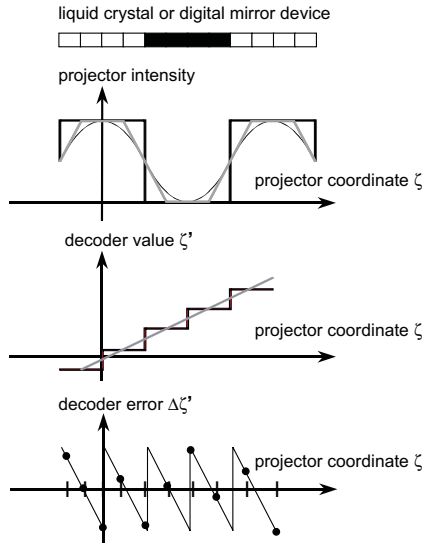


Figure 20.20: *Binary light modulators produce systematic errors in the detector signal.*

and calibrate all specific projector problems which have been partially discussed in Section 20.3.6.

An intermediate result of a sensor based on camera and coded light projector is an image matrix filled with local decoder values ζ_d at (almost) every pixel positions x,y. From these $x - y - \zeta_d$-triples the XYZ-coordinates in the object space have to be calculated.

Polynomial approximation of three-dimensional sensor raw data. Approximation of 3-D sensor data to a three-dimensional mechanical reference can be one solution of the 3-D calibration problem. For example, the measured 3-D target positions can be fitted to the well known coordinates of a moved plate (Fig. 20.21) by polynomial functions.

This calibration algorithm does not require an understanding of how the 3-D-sensor works. The only assumption is that the sensor's mapping function $X_{out} = F(x, y, \zeta_d)$ is unique and smooth. This means that the higher order frequency terms of the deviations from the orthogonal reference coordinate system are negligible, and the mapping function F can be sampled and approximated by a rather coarse three-dimensional grid of distinct points.

This calibration concept has certain limitations:

1. it needs accurate, long-term-stable 3-D reference systems (e.g. a flat calibration plate with targets and a mechanical positioning system with high accuracy),

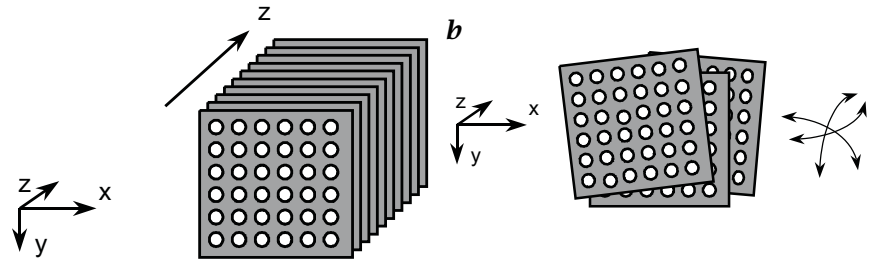


Figure 20.21: ***a*** *Moving a 2-D grid in normal direction represents of a 3-D grid. Errors in the positioning device effects the calibration.* ***b*** *For photogrammetric calibration the positioning of the calibration plate can be simple, because there is no need for precision.*

2. data reconstruction requires grid interpolation, which is a time consuming numerical operation, and
3. sub-pixel precision is sufficient for 3-D measurement, but is not suitable for photogrammetrical applications.

Model-based projector calibration. Photogrammetric algorithms for camera calibration are based on physical camera models. To find the model parameters, the image coordinates of targets (output) are measured with very high sub-pixel precision and compared to the actual object coordinates (input).

Light projectors, which are programmable in two dimensions, can be seen as inverse cameras. To calibrate such projectors, the object coordinates (output) of projected targets (e. g., crossed stripes) are measured with two or more cameras and compared to the known projector coordinates (input) (stripe numbers in x and y).

Active-optical sensors with stripes in one direction and with one camera only cannot be calibrated with photogrammetric standard calibration tools. Therefore a new self-calibration technique was developed. In terms of photogrammetry, the sensor is modeled as a stereo camera pair where one camera is an inverse camera with “long” pixels (stripes). This new calibration method combines several advantages:

- the calibration delivers all lens and orientation parameters of camera(s) and projector,
- all calibration tasks can be done in a few minutes,
- the calibration plate (or sensor) may be positioned by hand,
- a calibration check can be done during measurement,

- the sensor system is suitable for photogrammetric tasks such as autonomous sensor orientation, and
- the measured 3-D coordinates can be calculated and transformed very fast into a given coordinate system.

This calibration concept which has been developed for the "free-flying" sensor uses a hierarchical concept to speed up and stabilize the calibration process. If possible, it makes use of

1. a stable camera lens and its model parameters from prior lens calibration,
2. a fixed lens-camera system and their inner orientation parameters from prior camera calibration,
3. a stable multi-camera setup and their outer orientation parameters,
4. known target coordinates in object space from prior measurements.

If one or more of these parameters are known and fixed, the remaining unknown parameters (e. g., the object coordinates) can be calculated based on less images or with higher reliability. This is an important feature for robust systems with integrated on-line self checking tools and easy-to-use calibration functionalities.

20.5 Hybrid navigation of three-dimensional sensors

Generally, 3-D objects have to be measured from several views. In every view, the sensor measures point clouds given in the *sensor coordinate system* (SCS). To collect all measured data in one predefined world or *object coordinate system* (OCS) it is necessary to know the spatial transformation matrix from SCS to OCS in every view.

Several principles and terms are used in this context. We use the term "navigation" here in a more general sense to integrate different concepts.

Mechanical positioning systems with high accuracy. In principle, object and sensor can be "coupled" to each other by a mechanical positioning system. It controls all of the six degrees of freedom and therefore it defines the transformation matrix. However, there are some disadvantages to this method:

- high precision mechanics with five or more axes are very expensive, and will remain so in the future,
- the path of mechanical connectors and interfaces is complex and its transformation and error propagation is difficult to calibrate,

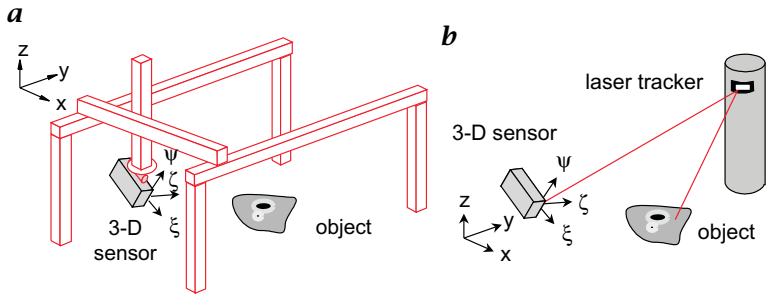


Figure 20.22: ***a*** *Mechanical system used to control the six parameters of sensor orientation.* ***b*** *External navigation system to measure the sensor orientation.*

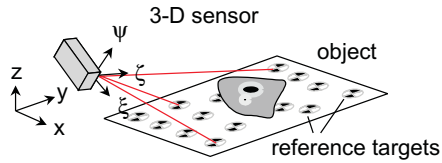


Figure 20.23: *Autonomous sensor orientation, based on natural features or additional targets.*

- mechanical systems define the mechanical coordinate system, not the optical coordinate system, and therefore a hand-eye calibration is needed, and
- the principle is very sensitive to angular errors.

External high-accuracy navigation instruments. Laser interferometers could be used to measure the absolute positions of sensor and work piece. However, there are similar problems as in mechanical positioning systems. In addition, external navigation instruments with sufficient accuracy in all six degrees of freedom are very expensive.

Self-navigation of free-flying sensors. As opposed to the mechanical or optical guided concepts, the sensor itself can measure its orientation relative to the object. This method uses predefined optical targets on or near the object with known coordinates in the OCS (Fig. 20.23).

The self-navigation principle uses photogrammetric methods for target detection, target labeling and a first estimation of the orientation based on one image. The following precise 3-D navigation uses target models and the measured point cloud. The navigation errors can be minimized very efficiently, because the sensor measures navigation targets and object surface at the same time with the same measurement principle in the identical SCS.

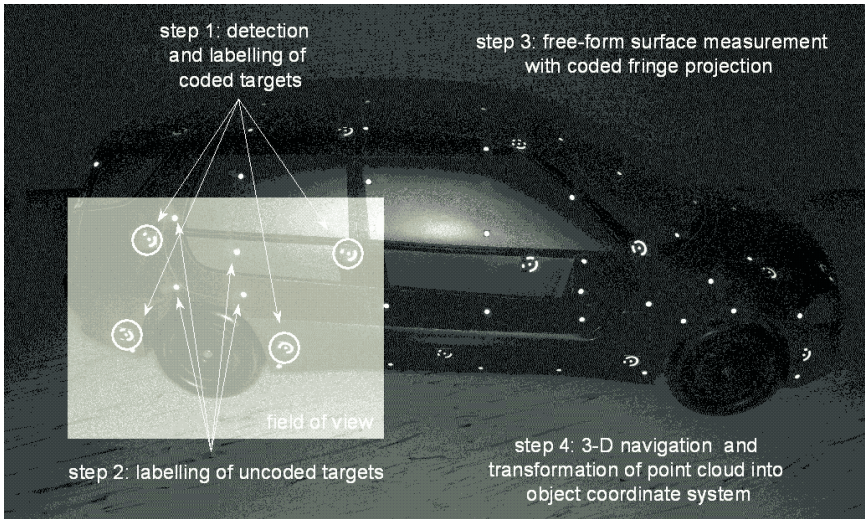


Figure 20.24: *Principle of autonomous sensor navigation based on predefined targets.*

Feature-based registration of point clouds. Some objects have "geometric contrast" from edges and other discontinuities which can result in features. Features extracted in several point clouds can be used to fit these point clouds to each other.

This "data navigation" method should be used with care, because point clouds of features are often incomplete and contain different systematic errors in different views.

Hybrid navigation. A generalized navigation concept as proposed in [19] uses spatial information from different sources and integrates all methods discussed above. 2-D images and optical 3-D data from predefined targets or object features, mechanical positioning parameters or CAD model features can be used to complete the navigation matrix.

In some applications coordinate measuring machines, milling machines, robots or simple linear or rotation axes are available for sensor or object movements. In most cases, however, some degrees of freedom have good reproducability and precision, other axes are weak or unknown. On the other hand, optical navigation information may be available in some positions and in others not.

To calculate the complete transformation matrix, partial optical and mechanical navigation can be combined. Simple translation or rotation axes can be very helpful for the hybrid measurement concept. For example, a rotation table with a reference target field can "freeze" all rotational degrees of freedom into one known axis. Objects placed on

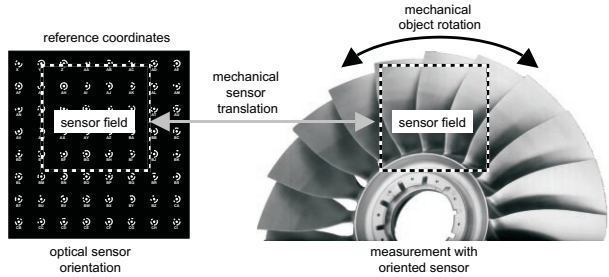


Figure 20.25: *Hybrid navigation example, which combines optical navigation, linear and rotational motion.*

this table may hide most of the targets, but, in principle, one target seen during the object measurement is sufficient for a complete navigation.

A more complex navigation principle is shown in Fig. 20.25. It was designed for applications where optical navigation is impossible in the measuring position. The sensor is mounted on a linear axis, the object is mounted on a rotation axis. In the left position the sensor can derive all six degrees of freedom from optical navigation. Then the known matrices from mechanical translation and rotation are added for a full navigation in the right measuring position, where no targets can be seen.

20.6 Mobile measuring system "Ganymed"

"*Ganymed*" is a mobile measuring system for reverse engineering purposes developed at Daimler-Chrysler AG. The measuring principle is based on one CCD camera and one code projector (Fig. 20.8b). This allows measurement of isolated coordinates based on single camera pixels. No image correlation is necessary and no assumption of lateral continuity of the object's surface. Optimized hybrid codes allow the measurement of specular metal surfaces without preparation. Typical sensor components are a 1×1 k pixel digital camera and a high resolution stripe projector for digital and analog patterns. Camera and projector are mounted on a tripod and are typically configured for a measuring distance of 1 m and a measuring volume of $0.5\ \text{m} \times 0.5\ \text{m} \times 0.5$ m. The projector is supplied from a powerful external light source via fiber optic cable.

The sensor can be positioned by hand (Fig. 20.26). Point clouds can simply be acquired by pressing a button. The system transforms the measured points (up to 1 million per view) into the object coordinate system. For interactive path planning and detecting missing data the actual state of point cloud is displayed on the system monitor. To mea-

Figure 20.26: *"Free-flying" sensor Ganymed in use.*

sure larger objects the navigation targets can be sticked to the object's surface as shown in Fig. 20.24 and Fig. 20.26. Smaller objects can easily be measured using a latticed box prepared with navigation targets.

20.6.1 Measuring system "Oblisk" in a milling machine

For in-process measurement of bladed disks (Blisk) in a milling machine a modified "Ganymed" system with a special sensor head and a smaller field of view has been developed. The measured point clouds from all blades are combined by the hybrid navigation principle as shown in Fig. 20.25. Based on the dense point cloud a virtual touch probe function has been realized. The operator can easily define and generate control points (Fig. 20.27) or profile sections (Fig. 20.28).

20.6.2 Digitizing rapid prototyping objects

As a further example, Fig. 20.29a shows a metal part designed for testing the complete reverse engineering / rapid prototyping cycle (Fig. 20.1). Twelve different views (one is shown in Fig. 20.29b) were acquired to obtain a complete point cloud (Fig. 20.29c) that was suitable for generating a CAD model and rapid prototyping data (Fig. 20.29d). Based on these data several parts were built with different rapid prototyping principles such as stereo lithography or laser sintering and measured again for process control.

20.6.3 Virtual target concept

Due to the analog code projection, the "Ganymed" sensor can measure with the full lateral sampling rate of the camera matrix without any

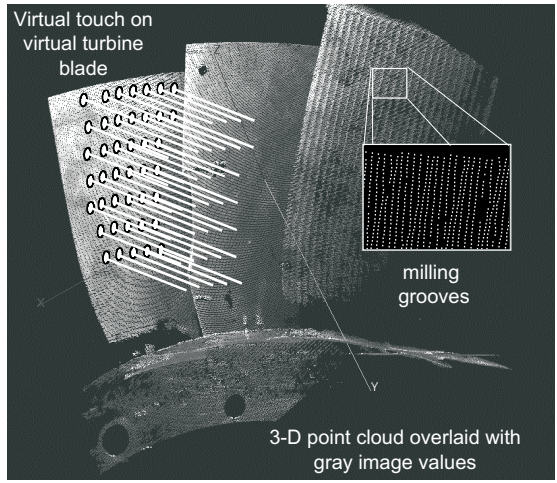


Figure 20.27: *Virtual touch probe on a virtual turbine blade.*

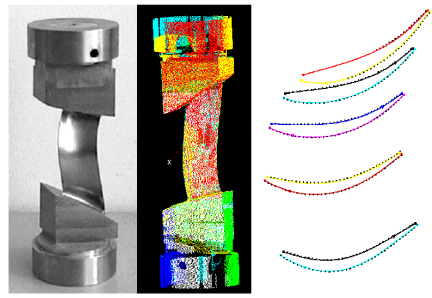


Figure 20.28: *Titanium turbine blade (left), measured point cloud and five profile sections (right) derived from the point cloud.*

lateral interpolation. The sensor produces spatial coordinates for every camera pixel, for example, up to 1 million coordinate points with a $1\text{k} \times 1\text{k}$ camera in a minute. Even fine objects like thin wires with diameters less than the pixel size in object size can be detected (Fig. 20.30).

As shown in Fig. 20.31, in principle, the subpixel resolution of an object coordinate point corresponds to the number of pixels that can be used for the calculation of this coordinate point. Furthermore, the number of pixels that can be used for a local fit depends on the object's surface quality (Fig. 20.32). The best subpixel resolution can be achieved with extended flat targets of good optical quality. On the other hand, an increasing number of pixels per calculated coordinate point reduces the lateral sampling rate of coordinate points.

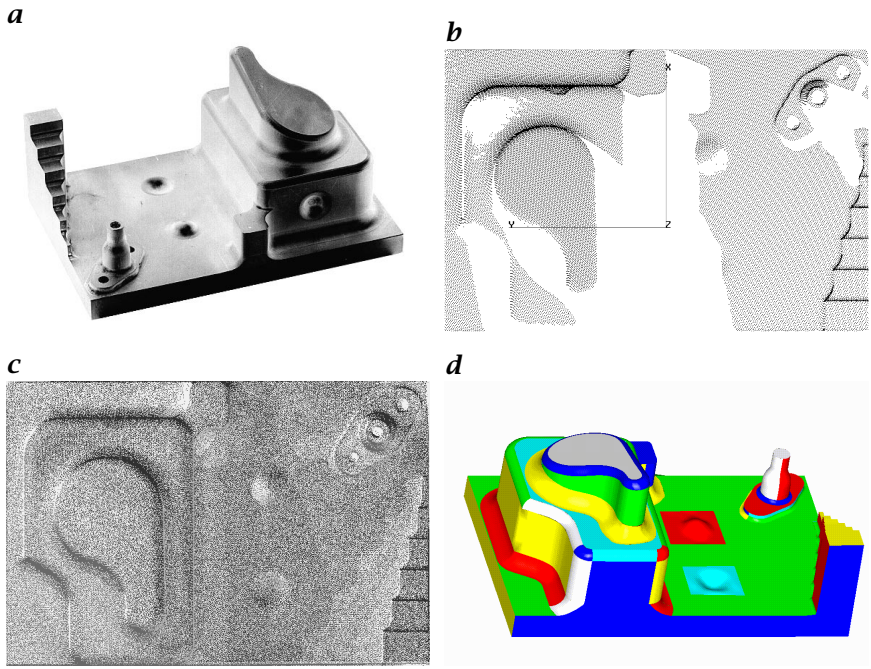


Figure 20.29: ***a*** *Metal test part (approx. 150-mm length) and* ***b*** *one measured point cloud with some nonmeasurable areas.* ***c*** *Superimposed point cloud from 12 measurements and* ***d*** *CAD model, derived from the complete point cloud.*

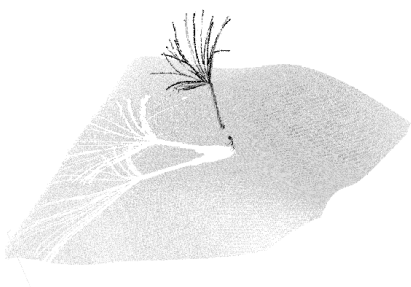


Figure 20.30: *Measured point cloud of thin wires with diameters less than pixel size.*

While sensors based on image correlation with targets, projected textures or coded light projection always must use several neighboring pixels for calculating an object point, the active triangulation with calibrated projector can use lateral information to improve the point quality. Depending on application and object properties the user can

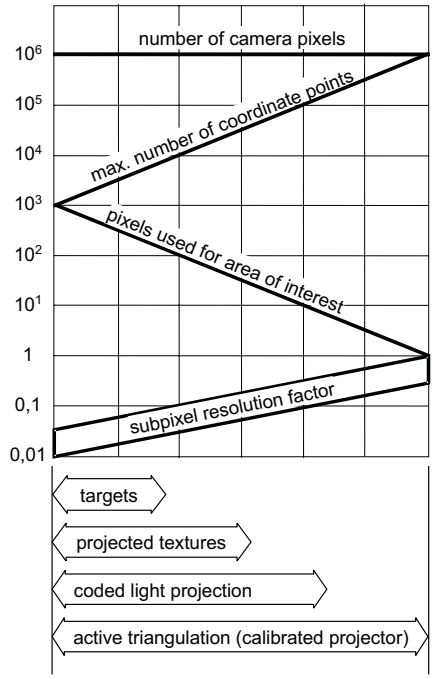


Figure 20.31: *Sensors based on calibrated projectors span the whole range from pixelwise dense surface measurement to low-noise measurement of targets with variable size.*

define "virtual targets" of variable size on the dense point cloud (instead of physical targets on the object). With this concept the system can operate in the whole range from pixelwise dense surface measurement to low-noise measurement of real or virtual targets with variable size.

20.7 Conclusions

The "free-flying" sensor system described in this chapter integrates the advantages of active optical and photogrammetrical techniques (Table 20.2). Like photogrammetry-based systems it does not require expensive, high-precision positioning systems for calibration, navigation or measurement. Optical sensor navigation without or in combination with different mechanical positioning hardware is possible. The self-calibration principles for camera, projector and system operates with a simple hand-positioned calibration plate. The active-optical sensor performs pixelwise dense surface measurement or low-noise measurement of virtual targets with variable size. The sensor can be used for

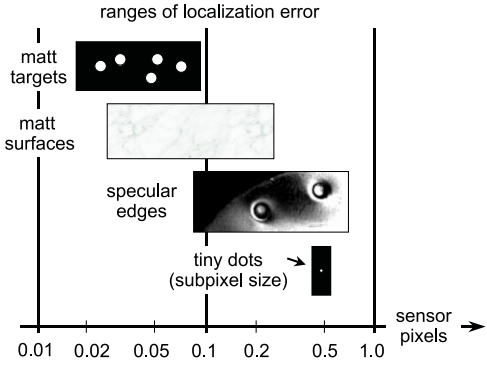


Figure 20.32: *Ranges of measurement errors on different materials.*

Table 20.2: *The new sensor principle integrates photogrammetric, active-optical and newly developed features (see also Fig. 20.3)*

Principles Properties	Multiple Cameras	Ganymed Sensor	Stripe Sensor
Large textured areas at daylight	•	—	—
Camera calibration	•	•	—
Camera orientation	•	•	—
Subpixeling on targets	•	•	•
Vector target point measurement	•	•	•
Nontextured surfaces	—	•	•
Specular surfaces	—	•	•
High signal dynamics	—	•	•
Point clouds of pixel density	—	•	•
Stripe projector calibration with one camera	—	•	—
Sensor orientation with camera and projector	—	•	—

a broad spectrum of surfaces. High robustness and signal dynamic based on hybrid codes allows the measurement of specular surfaces in industrial environment.

The flexible system concept developed at Daimler-Chrysler Research integrates and uses all these features. It can be configured as a mobile optical coordinate measurement system for reverse engineering tasks or quality control applications. It is also useful for automated in-process measurement systems, for example, in a milling machine.

20.8 References

[1] Häusler, G., (1990). About fundamental limits of three-dimensional sensing. *SPIE Optics in Complex Systems*, **1319**:352–353.

[2] Malz, H. J., R.; Tiziani, (1991). Schnelle 3-D Kamera mit adaptierbaren, hochauflösenden Markierungscodes. In Waidelich, W. (ed.), *Proc. 10th International Congress Laser '91, Munich*, pp. 153–157. Springer.

[3] Konecny, G., (1996). Hochauflösende Fernerkundungssensoren für kartographische Anwendungen in Entwicklungsländern. *ZPF Zeitschrift für Photogrammetrie und Fernerkundung*, **2**.

[4] Luhmann, T., (1995). Punktmessung in digitalen Bildern mit Subpixel-Genauigkeit. In Ahlers, R. (ed.), *Proc. Bildverarbeitung '95 - Forschen, Entwickeln, Anwenden*, Esslingen: Technische Akademie.

[5] Takeda, M. and et al., (1982). Fourier-transform method of fringe pattern analysis for computer-based topography and interferometry. *Jour. Opt. Soc. of America*, **72**.

[6] Häusler, G., (1997). About the scaling behaviour of optical range sensors. In Jüptner, W. and Osten, W. (eds.), *Proc. 3rd Intl. Workshop on Automatic Processing of Fringe Patterns*, pp. 147–155, Berlin: Akademie Verlag.

[7] Malz, R., (1992). *Codierte Lichtstrukturen für 3D-Meßtechnik und Inspektion*. Thesis, Institute for Technical Optics, University of Stuttgart, Stuttgart, Germany.

[8] Shannon, C. E. and Weaver, W., (1949). *Mathematical Theory of Communication*. University of Illinois Press.

[9] Malz, R., (1989). Adaptive light encoding for 3-D sensing with maximum measurement efficiency. In *Proc. 11th DAGM-Symposium, Hamburg, Informatik-Fachberichte*, Vol. 219. Springer.

[10] Bruning, J. e. a., (1974). Digital wavefront measuring for testing optical surfaces and lenses. *Applied Optics*, **13**.

[11] Creath, K., (1988). Phase-measurement interferometry techniques. In Wolf, E. (ed.), *Progress in Optics*. Amsterdam: North-Holland.

[12] Küchel, M. F., (1997). Some progress in phase measurement techniques. In Jüptner, W. and Osten, W. (eds.), *Proc. 3rd Intl. Workshop on Automatic Processing of Fringe Patterns*, pp. 27–44, Berlin: Akademie Verlag.

[13] Takeda, M., (1997). The philosophy of fringes—analogies and dualities in fringe generation and analysis. In Jüptner, W. and Osten, W. (eds.), *Proc. 3rd Intl. Workshop on Automatic Processing of Fringe Patterns*, pp. 18–26, Berlin: Akademie Verlag.

[14] Zumbrunn, R., (1987). Automatic fast shape determination of diffuse reflecting objects at close range, by means of structured light and digital phase measurement. In *Proc. ISPRS Intercommission Conference on Fast Proc. of Photogrammetric Data, Interlaken, Switzerland*.

[15] Wahl, F. M., (1986). A coded light approach for depth map acquisition. In *Proc. DAGM-Symposium Mustererkennung 1986. Informatik-Fachberichte*, Vol. 125. Springer.

[16] Yamamoto, H., Sato, K., and Inokuchi, S., (1986). Range imaging systems based on binary image accumulation. *IEEE*, pp. 233–235.

[17] Wester-Ebbinghaus, W., (1989). Trends in non-topographic photogrammetry systems. In Karara, H. M. (ed.), *Non-Topographic Photogrammetry (2nd edition)*, pp. 377–387. American Society for Photogrammetry and Remote Sensing.

[18] Bergmann, D., Galanoulis, K., and Winter, D., (1997). Advanced 3D-fringe-projection system. In Jüptner, W. and Osten, W. (eds.), *Proc. 3rd Intl. Workshop on Automatic Processing of Fringe Patterns*, pp. 432–442, Berlin: Akademie Verlag.

[19] Malz, R. W., (1993). Anforderungen an photogrammetrische On-Line-Meßsysteme. In *ISPRS/DGPF-Workshop "Industriephotogrammetrie." Braunschweig, 15.-17.3.1993 (unpublished)*.

21 Three-Dimensional Light Microscopy

Ernst H. K. Stelzer

European Molecular Biology Laboratory (EMBL),
Heidelberg, Germany

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

21.1 Three-dimensional microscopy

Real objects are not flat but have a thickness and a surface height variation, that is, a topology. Although conventional light microscopes can be used to study different layers in an object and to measure differences in height, for some reason thickness has not been taken account in a rigorous manner. Instead, the theory of light microscopy is usually taught with two-dimensional specimens in mind. This applies especially to the well-known limit to the *resolution* of *transmission light microscopes*

$$\Delta x = 1.22 \frac{\lambda}{N.A.} = 1.22 \frac{\lambda}{n} \frac{1}{\sin \alpha} \tag{21.1}$$

that describes the distance Δx of two point objects in terms of the illumination wavelength λ and the numerical aperture $N.A.$ (i. e., the opening half-angle α of the objective lens) embedded in a medium with a refractive index n. Equation (21.1) is valid only if the two objects are in a common plane.

A flat object is by definition an object in which all features are in focus in a single plane. In a thick object it is not possible to focus on all features simultaneously. One must also distinguish between objects of different opacity. Opaque objects have a well-defined surface, while translucent objects can provide a stack of images in which each layer contributes to the three-dimensional image of the object.

The most important developments in the past few years have been *confocal microscopy* and special software for deconvolving stacks of images. The confocal microscope has made three-dimensional microscopy a permanent and lasting handy tool in many modern laboratories. The large number of users has in turn resulted in an improved understanding of three-dimensional image formation in light microscopy.

This chapter emphasizes the importance of *telecentricity* to the whole concept of three-dimensional microscopy, introduces the most important aspects of three-dimensional image formation, explains the rationale and limits of deconvolution, points out the basic idea of three-dimensional image reconstruction, and outlines problems that provide fundamental limits to the ways in which three-dimensional microscopes are used today. Finally developments that might result in higher resolution are presented.

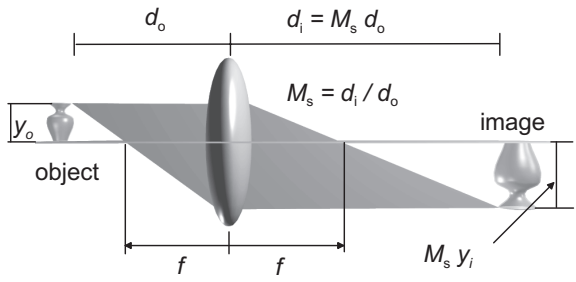


Figure 21.1: *Imaging with a single lens. The image of a vase is formed using a single lens. The distance of the vase from the lens is d_o, the distance of the image from the lens is d_i. Increasing the distance d_o decreases the distance d_i and changes the magnification M_s.*

21.2 Telecentricity

21.2.1 Single lens

An optical system with a single lens (Fig. 21.1) has a magnification M_s

$$\frac{1}{d_i} + \frac{1}{d_o} = \frac{1}{f} \quad \longrightarrow \quad d_i = \frac{d_o f}{d_o - f} \qquad M_s = \frac{d_i}{d_o} = \frac{f}{d_o - f} \tag{21.2}$$

that depends on the distances of the image d_i and the object d_o from a lens that has a focal length f; M_s is dependent on d_o, so that unless the distance of the object from the lens is known, the single lens system is of limited use when performing indirect measurements in an object by looking at its image.

21.2.2 Two lenses

Instead of single lens systems, one usually employs arrangements that consist of two lenses (or two groups of lenses) in a telecentric arrangement [1, 2]. The basic idea of *telecentricity* is that these two lenses share a common focal plane and that they are used to form an image. In a Keplerian arrangement (Fig. 21.2) the common focal plane is between the lenses. A diaphragm is centered around the common focus on the optical axis. It defines the beam diameter and, together with the focal length, the numerical aperture of the optical system and therefore also its resolution. In a telecentric system, the *lateral* (M) and *axial magnifications* (M^2) are independent of the position of the object along the

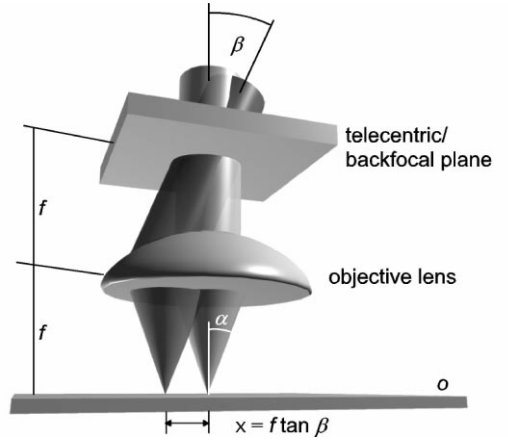


Figure 21.2: *Characteristic properties of a telecentric system. All beams pass the aperture diaphragm in the backfocal plane as planar waves. The tilt angle β determines the distance of the focus from the optical axis. The focal length f of the objective lens and the diameter of the entrance aperture $2a$ determine the opening angle α. The numerical aperture is the product of the angular aperture $\sin\alpha$ and the refractive index of the medium n. A planar wave tilted by an angle β has a focus in the object plane at a distance $x = f\tan\beta$ from the optical axis.*

optical axis (Fig. 21.3):

$$\begin{aligned} x_i &= M x_o, \\ z_i &= M^2 z_o, \\ \tan\alpha_i &= M^{-1}\tan\alpha_o \end{aligned} \tag{21.3}$$

If an object is not in focus and its boundaries are not clearly visible, its size will not change. The light distribution of the image of a single point object will spread as the sample is moved away from the focal plane, but the center-center distance of the images of two point objects will not change. This distance is a function only of the center-center distance of the objects and the magnification of the optical system. Three-dimensional imaging can only be understood by strictly adhering to the principle of telecentricity [3]. All optical systems that measure distances in an object, and of course all microscopes, use telecentric arrangements.

A microscope objective lens has to be regarded as a compound lens consisting of at least of two simple lenses with the focal lengths f_1 and f_2. If the magnification of a lens is 100× and the image is 160 mm away from the lens, it is reasonable to assume that the focal length of the smaller lens (which is closer to the object) is about 1.6 mm. When

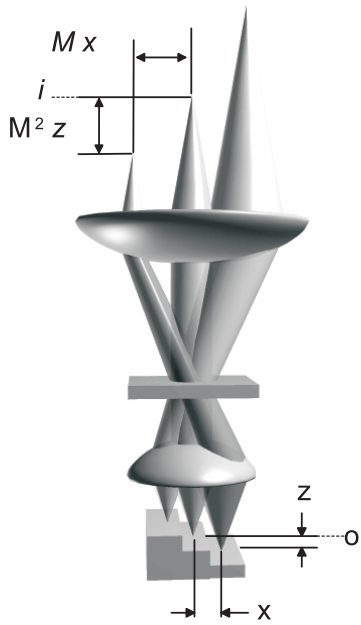


Figure 21.3: *Three-dimensional imaging in a telecentric system. A pair of angles encodes the lateral positions. A divergence or convergence angle defines the position of the emitter along the optical axis. The lateral distance Mx in the image plane i is independent of the position of the emitter along the optical axis. In every telecentric system the axial magnification is the square of the lateral magnification M. If the objects are located in different planes with a distance z, then their images have a distance along the optical axis M^2z. Please also note that the out-of-focus beams are expanded in the image plane i where the detection pinhole of a confocal microscope is located.*

using an immersion system the ratio of the respective refractive indices has to be taken into account.

Infinity-corrected optics are no exception. The tube lens (e. g., in an Axio/ICS-microscope from Carl Zeiss) probably has a focal length of around 160 mm and shares a common focal plane with the microscope objective lens corrected for infinity. The telecentric system then consists of the microscope objective and the tube lens.

The telecentric plane is also referred to as the backfocal plane of the microscope objective lens. If an object is in the focal plane of a lens, its *Fourier-transformed* image is found in its conjugate focal plane or backfocal plane. The second lens takes the inverse Fourier transform of the Fourier transform, and hence forms a real image in its backfocal plane. In a microscope this is the primary image plane, and is conjugated to

the object plane. The important property of the Fourier transform is that every position in the object plane has an associated pair of angles in a planar wave passing the backfocal plane.

If one looks carefully at the layout of a microscope one will realize that pairs of lenses are used to form images. The objective lens and the tube lens form an image of the object in the image plane. The condenser of the lamp and the tube lens form an image of the lamp in the backfocal/telecentric plane of the microscope objective lens. The ocular and the lens in the eye of the observer form an image of a conjugate image plane in the retina of the eye.

21.2.3 Position of scanning system

The correct position of the scanning system is obviously crucial for the performance of a *scanning laser microscope* [3, 4]. The scanner performs two functions: 1) The initially stationary illumination beam is tilted and hence the focus is moved laterally in two dimensions in the focal plane (Fig. 21.4); and 2) the light reflected/emitted by the sample is deflected towards the detection pinhole. Because there is practically no time delay (reflection is instantaneous and the fluorescence decay lifetime is on the order of nanoseconds), incoming and outgoing light will always follow the same optical path.

The Fourier transform relationship between the object or image planes and the telecentric plane tells us that each position $[x, y]^T$ in the object plane is associated with a pair of angles $[\varphi, \theta]^T$ in the telecentric plane and vice versa:

$$[x, y]^T \longrightarrow [\varphi, \theta]^T \qquad [\varphi, \theta]^T \longrightarrow [x, y]^T \tag{21.4}$$

By putting the scan mirrors into planes that are conjugate to the telecentric plane and by tilting the mirrors by φ, θ in orthogonal axes, a light spot can be moved to any position (x, y) in the object plane within the field of view of the microscope objective. In an ideal system both mirrors will be placed in planes that are conjugate to the telecentric plane Fig. 21.5. However, it is usually sufficient to have both mirrors very close to each other with the telecentric plane between them. An alternative is to tilt a single mirror in two axes [3, 5]. In any case the deflected light beam must be collimated.

Confocal microscopes using such a pair of mirrors observe only a single point in an object at a time. Therefore, they do not record an image. To get an image one must either move the beam relative to the object [6] or the object relative to the beam while recording the intensity as a function of the their relative position [7, 8, 9, 10]. In a practical instrument, the beam is moved laterally in the focal plane of the instrument while the sample is moved along the optical axis.

The two mirrors are mounted on very accurate motors (galvanometers), which allow almost arbitrary changes of angle as well as the speed at which an angle is reached. A large angle is equivalent to a large field. Thus, changing the range of accessible angles controls the field size [3, 4, 11]. The off-axis and on-axis precision (wobble and jitter) of a typical galvanometer is good enough to allow positioning of the beam with a precision that exceeds 5 nm in the field of a lens with a lateral magnification of 100.

In most inverted microscopes the axial movement is achieved by moving the lens relative to a fixed stage while in most upright microscopes the stage is moved relative to a fixed optical system. Because the axial displacement moves a larger mass, it is in general much slower than the lateral movement. For this reason, and also to allow a more accurate control of the axial displacement, in some microscopes the mounting element for the sample is movable independently using additional galvanometers.

21.3 Theory of three-dimensional imaging

In light microscopy, resolution is defined by the extent of the intensity of the *point spread function* (PSF), which is a mathematical description of the intensity distribution of the image of a point source in the focal region. The smaller the extent of the intensity PSF, the better the distinction of separate points and hence the resolution of the microscope. An image is calculated from the disturbance of the electric field caused by the object ([12], Chapter 4). A function $u(x', y')$ that is proportional to this disturbance function is convoluted with the amplitude PSF $h(x - x', y - y')$

$$u_1(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(x', y') h(x - x', y - y') \, \mathrm{d}x' \, \mathrm{d}y' \tag{21.5}$$

The theory makes two assumptions: invariance and linearity. Invariance essentially means that a single function exists that describes the image formation process for every point in the object. Linearity means that the image formation is independent of the object.

Although it is not obvious at first sight, this principle can be extended to the three-dimensional case. The integral Eq. (21.5) is then performed in three dimensions instead of in two and it is the 3-D disturbance of an electric field that has to be calculated. The integral can be abbreviated using the convolution operator

$$u_1 = u_0 * h \tag{21.6}$$

Two cases must be distinguished ([13], p. 39). Reflection will in general maintain the coherence of the electric field, and the intensity I_1^{refl} is derived by taking the modulus squared of u_1

$$I_1^{\text{refl}} = u_1 u_1^* = |u_1|^2 = \left| u_0 * h \right|^2 \tag{21.7}$$

Fluorescence, on the other hand, does not maintain coherence. Fluorescent emission is proportional to intensity and not to amplitude. The emission is therefore calculated by convoluting the modulus square of the PSF and the modulus square of the electric field disturbance due to the object:

$$I_1^{\text{fl}} = |u_0|^2 * |h|^2 \tag{21.8}$$

21.4 Confocal microscopy

Confocal microscopy (CM) plays an important role in three-dimensional imaging of thick samples because of its depth discrimination capability (Fig. 21.3). This is due to the combination of point illumination and the restriction of detection by a pinhole in front of a light-sensitive detector [14]. The depth discrimination is usually interpreted as an improved axial resolution. Due to the usually lower signal-to-noise ratio the improved lateral resolution is of no practical importance.

21.4.1 Confocal fluorescence microscopy

A confocal microscope is usually based on a conventional microscope. It contains an objective lens, a stage for the sample, a light source and a detector. If we work with fluorescent light, two filters are instrumental. The dichroic filter separates the illumination light with which the fluorophore is excited from the fluorescent light that is emitted by the fluorophore; the second filter separates the emitted fluorescence from scattered excitation light (Fig. 21.4). The light path is best understood if one looks first at the excitation and then at the emission light. The light source is usually a laser. The laser light is focused into a pinhole, deflected by a dichroic mirror into the objective lens and focused inside the specimen. Most of the light passes through the specimen, but a small fraction is absorbed by the fluorophore, which then emits fluorescent light. The fluorescent light is emitted in all directions with an almost equal probability so that the lens collects only a small fraction of the fluorescent light. This fraction passes the dichroic mirror and is focused on a pinhole in front of a detector.

The images of the fluorophores outside the plane in which the point detector is located (Fig. 21.2) are expanded, and hence only a small fraction of their light passes the pinhole and enters the detector. The

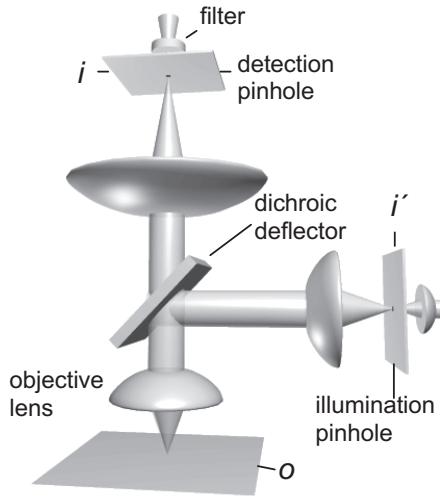


Figure 21.4: *Principle layout of a beam/object scanning confocal fluorescence microscope. A laser beam is focused into an illumination pinhole (in the plane i'), collimated and deflected towards a microscope objective lens, which focuses the light into the object. The emitted light is collected by the objective lens, passes the dichroic mirror and is focused into the detection pinhole in the image plane i. This pinhole allows the light emitted in the plane of focus o to pass and discriminates against all out-of-focus light (see also Fig. 21.3).*

detector pinhole thus discriminates against light that is not emitted in the plane of focus. The importance of the pinhole becomes even clearer when it is removed and a detector with a large sensitive area is used. Discrimination does not occur, and instead all of the fluorescent light that is collected by the objective lens contributes to the signal. Such an optical arrangement is not confocal, and behaves like any conventional fluorescence microscope.

Another view is to regard the objective lens as the device that forms an image of the illumination pinhole and the detection pinhole in their common conjugate image plane, the object plane o. Only the fluorophores that are in the volume shared by the illumination and detection PSFs are excited and detected. Therefore, in order to calculate the confocal PSF one calculates the illumination intensity PSF, the detection intensity PSF and multiplies the two functions

$$|h_{\mathrm{cf}}(x,y,z)|^2 = |h_{\mathrm{det}}(x,y,z)|^2 \, |h_{\mathrm{ill}}(x,y,z)|^2 \tag{21.9}$$

In many cases the illumination and detection intensity PSFs are quite similar and a reasonable first approximation of the confocal intensity

PSF is to assume it is the square of the illumination intensity PSF

$$|h_{\mathrm{cf}}(x,y,z)|^2 \approx \left(|h_{\mathrm{ill}}(x,y,z)|^2\right)^2 \tag{21.10}$$

If one looks at the full width at half maximum (FWHM) value of the confocal intensity PSF, the *confocal fluorescence microscope* (CFM) has its lateral and axial resolutions improved by about a factor of $1/\sqrt{2}$ over a conventional microscope. The zero positions (location of the first minimum) are of course identical in the PSFs, so that using this definition the CFM has no improved resolution. It should not be forgotten that a conventional fluorescence microscope has an axial resolution for point-like objects that is not much worse than that of the CFM.

To fully appreciate the CFM one has to understand the integrated intensities of the illumination and the confocal intensity PSFs

$$I_{\mathrm{ill,\,int}}(z) = \int_{r=0}^{r=\infty} |h_{\mathrm{ill}}(r,z)|^2 2\pi r \,\mathrm{d}r \tag{21.11}$$

For a real system this function is independent of z, which reflects the conservation of energy. The square of the integrand, however, is not conserved. Thus, the integral

$$I_{\mathrm{cf,\,int}}(z) = \int_{r=0}^{r=\infty} \left(|h_{\mathrm{ill}}(r,z)|^2\right)^2 2\pi r \,\mathrm{d}r \tag{21.12}$$

has a maximum in the focal plane. This is the explanation for the depth discrimination capability of a CFM [10, 15].

The best illustration of this effect is to record the intensity as one focuses through the coverslip into a thick layer (or sea) of fluorophore dissolved in the immersion medium of the lens

$$I_{\mathrm{cf,\,sea}}(z_0) = \int_{z=-\infty}^{z=z_0} \int_{r=0}^{r=\infty} \left(|h_{\mathrm{ill}}(r,z)|^2\right)^2 2\pi r \,\mathrm{d}r \,\mathrm{d}z \tag{21.13}$$

where z_o is the position of the coverslip-fluorophore interface. The slope and intensity variations in the shape of the sea response can be used to characterize the resolution of many confocal microscopes.

The sea response is unique to the CFM. A conventional fluorescence microscope has no such property, and as long as no phase information is available no computational methods are able to reconstruct the transition into the fluorophore layer from wide-field images.

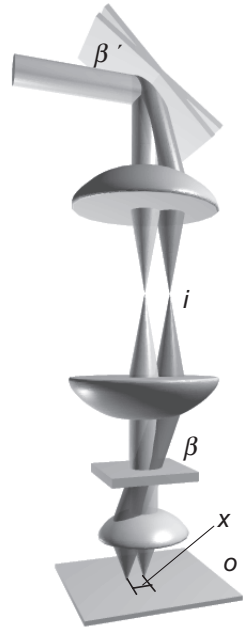


Figure 21.5: *Location of the scan unit in a laser scanning microscope. In a confocal beam scanning microscope, mirrors located in planes conjugate to the telecentric system are tilted and thus deviate the beam by an angle β' that is proportional to β. This causes a light spot to move in the image plane i.*

21.4.2 Resolution

Ultimately the resolution of any optical instrument is determined by its contrast transfer function. Even if resolution is specified as the full width half maximum (FWHM) of the PSF, it is the contrast that determines the performance of an instrument. The *lateral resolution* can be determined by measuring the size of small particles or by measuring distances between two particles or two lines. An estimate for the lateral point resolution Δx of a confocal microscope in terms of the FWHM is [16]

$$\Delta x = \frac{1.22}{\sqrt{2}} \frac{\lambda}{N.A.} \tag{21.14}$$

The point resolution can thus be expected to be about 40 % better than in a conventional microscope (compare Eq. (21.1) and Eq. (21.14)) if two flat samples are compared.

The *axial resolution* is much more complicated to measure. In principle, a single point could be observed in a plane parallel to the optical axis and both axial and lateral resolution could be derived from a single

x/z-image. An estimate for the axial resolution Δz for point objects in terms of the FWHM is

$$\begin{aligned} \Delta z_{\mathrm{fl}} &= 1.5n\frac{\lambda}{N.A.^2} \\ \Delta z_{\mathrm{ref}} &= 1.0n\frac{\lambda}{N.A.^2} \end{aligned} \tag{21.15}$$

for fluorescing and reflecting objects, respectively. The axial resolution for a reflecting object is thus about 30 % better than for a fluorescing object. Note that the resolution improves with decreasing refractive index n of the immersion system. This means that an air lens with a numerical aperture of 0.95 has a higher axial resolution than an oil immersion lens with a numerical aperture of 1.4. The factor of 1.5 is an example when the fluorophore FITC is illuminated at a wavelength of 488 nm and detected around 520 nm. It depends on the ratio of the excitation and emission wavelengths and will increase when fluorophores with a large Stokes shift are observed, but the estimate is probably sufficient for practical purposes.

21.4.3 Confocal reflection microscopy

Confocal reflection microscopy is used to determine the surface topology of reflecting and scattering objects. However, most objects do not behave like perfect mirrors. Light will be reflected from different layers in the object. This is, for example, the case in the slightly transparent silicon oxide layers, which cause multiple images and axial interference fringes. A solution to this problem is incoherent illumination using white light [17]. The incident light might also be scattered, which makes it behave more like a fluorophore. There are also claims that interference can be used to improve the lateral and axial resolution [18, 19, 20, 21] but this has not had an impact on commercially available instruments.

21.4.4 Recording stacks of images and three-dimensional reconstruction

Because the main advantage of the confocal microscope is its depth discrimination, one usually does not record a single image but a stack of images. The center position of each image is identical but the focal plane is moved through the object. Objects within a slice the thickness of which is given by the finite axial resolution of the microscope, cannot be distinguished. The axial spacing of the images in the stack should therefore be at least half the axial resolution. One main problem is that the excitation light penetrates the whole sample. It will not only excite the fluorophores in the focal plane but also those above and below.

This can mean that fluorophores are consumed before they have been observed. This problem can be partially avoided by using two-photon excitation.

To correctly digitize the image, a simple rule of thumb is to use at least four (N=4) but not more than sixteen (N=16) picture elements per side of an Airy disk. Using Eq. (21.14) we can calculate the distance between two picture elements

$$\Delta x_{\text{pix}} = \frac{1.22}{\sqrt{2}} \frac{\lambda}{N.A.} / N \tag{21.16}$$

Assuming a numerical of 1.4 and a wavelength of 500 nm this distance lies between 77 nm and 19 nm. If the field consists of 512 elements or lines, the field size is somewhere between 40 µm and 10 µm. This is by no means large but smaller than found in most applications. However, as pointed out elsewhere, it is required to collect all the information that is available from a physicist's point of view.

The recording time for a picture depends on the dwell time, for example, the time the scanner resides in a position and collects light. It usually varies between 1 µs and 100 µs. In confocal fluorescence microscopy this is sufficient to collect between 10 and 30 photons. The recording time per picture will be at least

$$512 \frac{\text{picture elements}}{\text{line}} \times 512 \frac{\text{lines}}{\text{image}} \times 1\mu\text{s} = 0.26s \tag{21.17}$$

However, the scanners have dead times until they are accelerated and they are often used only along one direction. For short dwell times the time per image will be about a factor of four larger; for dwell times around 10 µs probably less than a factor of two larger; and negligible for long dwell times. But it is not uncommon to spend 30 s to 1 min per image when fixed specimens are observed and a good signal-to-noise ratio (i.e., a large number of photons) is more important than rapid recording. When stacks are recorded the stage is moved, and the time the stage requires to settle or the time required to measure the correct position can be quite large.

However, apart from theta and 4Pi microscopes (Section 21.6), all light microscopes have better lateral than axial resolution. The problem in image reconstruction is that the volume elements have different lateral and axial extents. This makes the reconstruction complicated since side-on (90°) views tend to appear smeared.

21.4.5 Confocal transmission microscopy

It has also become clear that not all contrasts will show depth discrimination in a confocal arrangement. Transmission contrasts (implemented using two lenses) [7, 16] usually depend on absorption and on scattering.

Only when the signal is (at least partially) due to scattered light will the confocal contrast have an improved lateral resolution (e. g., phase contrast and differential interference contrast). The axial resolution defined by the sea response is only present in fluorescence and scattering light microscopy.

21.4.6 Disk scanners

A serious alternative to laser scanning microscopy is the use of scanning disks [22] that are located in an image plane. These have a number of holes that transmit the light of lamps [23]. In its simplest form the same hole is used for illumination and detection [24]. One rotation of the disk covers the whole field of view at least once, which is observed either directly or recorded using a camera.

When using a laser instead of the lamp lens arrays replace the holes in a disk and provide a very efficient and fast confocal microscope [25]. Apart from certain compromises that allow for efficient observation, and a poorer background discrimination, the properties of such systems are described in the same way as those of confocal microscopes [26].

21.4.7 Conventional recording and subsequent deconvolution

The idea of deconvolution is best understood when one considers transfer functions. *Deconvolution* essentially divides the Fourier transform of the image by the Fourier transform of the PSF and (since higher frequencies are less efficiently transferred than lower ones) amplifies the higher frequencies [27, 28, 29]. Due to noise this procedure is not straightforward and estimation as a function of the spatial frequency is required [30]. In addition, the PSF must be estimated. This can be done in a separate experiment [31, 32, 33] or calculated during the deconvolution process [34]. A perfect deconvolution would produce a rectangular transfer function. Its Fourier transform is a sinc function, causing ringing visible in the edges of the reconstructed image. Therefore, additional filters are employed that smooth the transfer at higher frequencies.

Because conventional fluorescence microscopy has a constant integrated intensity, it is unable to resolve planes perpendicular to the optical axis. Deconvolution of conventional images works well with point-like objects such as spheres and collections of spheres. Using integrating CCD cameras one starts with images having a high dynamic range. However, producing information about a smaller volume than is optically resolvable during the computational process means that this high dynamic range is sacrificed. A deconvolution computation always increases the noise in an image.

21.4.8 Confocal microscopy as a function of time

The observation of fixed specimens puts no limit on the time one needs to record an image. The observation can in principle be performed until the fluorophore is completely consumed. However, in all experiments as a function of time, one looks at the same area or volume more than once and tries to record spatiotemporal variations [35]. This can have severe consequences for the energy one has available for the observation [36] and the resulting images.

To minimize bleaching of the fluorophore, the energy per picture element must be as low as possible. This means that the intensity of the illuminating laser has to be low and the observation period per picture element (dwell time) has to be short. However, because the resolution ultimately depends on the signal-to-noise ratio, it will become worse as less energy is used in the illumination process. Additionally, it should be noted that the illumination and fluorophore degradation products can harm the object, which is particularly important when taking time-series images of living cells.

21.5 Index mismatching effects

A serious problem that cannot be neglected is spherical aberration due to mismatching of refractive indices. This can be particularly significant when high-resolution oil immersion objective lenses are used to observe specimens that are embedded in an aqueous medium. It is also observed when the refractive index varies inside large specimens, so that recording conditions that may be valid in one spot may not work in others. This problem is important for quantitative microscopy. The usual case (high N.A. oil immersion lens, aqueous embedding medium) causes a shift of the actual focal plane towards the lens and hence a decrease in the recorded axial distances. A decrease of the maximal intensity as well as an increase of the axial FWHM as one moves the focal plane farther away from the refractive index transition plane degrades the image quality. For example, in [37], 10 μm below the transition plane the axial FWHM was twice as large as under perfect conditions.

The literature on this topic is quite rich [38, 39] and the effects are well understood. But despite a number of efforts [40, 41] it is unlikely that such effects will ever be completely correctable. The only reasonable solution to this very serious problem is to use water immersion lenses whenever possible as a means to evade the problem. The disadvantage is a lower resolution close to the coverslip (i. e., transition plane), but the advantage is, of course, a uniform, undistorted view of the complete specimen [42].

21.6 Developments in confocal microscopy

The important role confocal fluorescence microscopy has in modern research is due entirely to its axial resolution, that is, its depth discrimination capability, which allows three-dimensional imaging. However, because a typical microscope objective covers only a small fraction of the full solid angle of 4π and therefore focuses only a small segment of a spherical wavefront, the axial resolution of a confocal microscope is always poorer than the lateral resolution. The observation volume in any single-lens microscope is therefore an ellipsoid elongated along the optical axis.

21.6.1 Multiphoton illumination

The transition of a fluorophore from the ground into an excited state can be accomplished by absorbing two photons each having half the energy of the gap [43]; but the probability for such a process is quite low and high intensities are required. The existence of this effect was proven as soon as lasers became available [44].

The importance of this effect for microscopy is that the fluorescence emission intensity $I_{2h\nu}$ after *two-photon excitation* (TPE) is proportional to the square of the excitation intensity

$$I_{1h\nu} \propto I_{exc} \quad I_{2h\nu} \propto I_{exc}^2 \tag{21.18}$$

The PSF of a microscope based on TPE is therefore the square of the illumination intensity PSF [45]. The TPE microscope has the same properties as a CFM but it does not require a detection pinhole

$$|h_{2h\nu}(x,y,z)|^2 = \left(|h_{\text{ill}}(x,y,z)|^2\right)^2 \tag{21.19}$$

By adding a point detector the resolution is further improved [46]

$$|h_{2h\nu,\text{cf}}(x,y,z)|^2 = \left(|h_{\text{ill}}(x,y,z)|^2\right)^2 |h_{\text{det}}(x,y,z)|^2 \tag{21.20}$$

Denk et al. [47, 48] were the first to describe a microscope based on TPE. Modern microscopes use lasers with short pulses in the femtosecond range and peak pulse powers in the kilowatt range, although TPE has also been reported with picosecond and cw lasers [49, 50]. Nearly all important dyes can be excited by TPE with laser lines between 700 and 1100 nm [51, 52]. The wavelengths are, however, twice as long as those used for single photon excitation.

The longer wavelength is the reason why the resolution of the TPE microscope is worse than that of a CFM, which is only partially compensated by the $1/\sqrt{2}$-term due to the squaring effect. Besides other advantages over single-photon absorption, a confocal two-photon fluorescence microscope can exhibit the resolution of a confocal fluorescence microscope operating in the UV below 380 nm [46].

21.6.2 Multiple lenses

Recent developments of new instruments use two or more objective lenses to observe the sample. The light distribution in the focal plane (e. g., by interference) [53, 54] or the spatial arrangement of the illumination and detection PSFs in a confocal microscope (i. e., the illumination PSF can be shifted or rotated relative to the detection PSF [55]) can be modified.

4Pi. In *4Pi-confocal fluorescence microscopy* [54] the sample is coherently illuminated and/or observed via two opposing objective lenses. Light interferes in the sample and/or detector and leads to a sub-structured PSF with several maxima along the optical axis. The axial FWHM of the central peak in the 4Pi-confocal PSF is smaller than the FWHM of the confocal PSF, and the observation volume is effectively reduced by a factor of two. This improvement has been demonstrated experimentally for fluorescence and scattered light microscopies [56]. A similar, nonconfocal, arrangement that makes use of interference to increase the axial resolution is standing-wave fluorescence microscopy [53].

Theta. A new microscopic setup [55] uses two objective lenses to illuminate the sample and to collect fluorescence emission at an angle to the illumination axis. The resolution enhancement stems from the alignment of the lenses: the detection axis is approximately orthogonal to the illumination axis. Therefore, the overlap of the illumination and detection PSFs are reduced and the lateral resolution of the objective lens becomes dominant. Thus, the microscope has high axial and lateral resolutions in all directions.

21.7 Resolution versus distance

So far, we have looked at the resolution, that is, at two objects that emitted light of the same wavelength. Therefore, an image can only show the sum of both objects. If, however, two point-like objects emit at distinct wavelengths, two independent images can be recorded, which will each show a single object. The exact location of each object can be calculated using intensity-weighted center of mass equivalents, and the distance of any two objects can be determined with a noise-limited precision [57]. The same idea also applies to single-particle tracking in video sequences [58]. The task here is to determine distances from intensity-weighted center of mass equivalents in independently recorded images. Such distances can be below 20 nm. Another example is the *photonic force microscope* [59], which uses the position of a single bead to determine a three-dimensional structure. The position of the bead inside the focal volume can be determined with a precision that is most likely below 10 nm. The position of topological structures in an object can

hence be determined with a resolution around 15 nm. An important example for distance measurements is the determination of the surface topology of integrated circuits using, for example, confocal reflection microscopes [60]. The height differences of planes that are sufficiently far apart can be determined with an unlimited precision. Here the surface roughness, the precision with which the position of the object can be measured along the optical axis, the reflectivity of the surface, and the coherence of the light source [17] all limit the resolution.

In the case of position determination of small objects, the localization accuracy of these objects is given by the error of the coordinates for the intensity maximum. This intensity maximum corresponds to the intensity bary centre. Therefore, the standard deviation of the intensity bary center coordinates of a series of measurements can be used to express the localization accuracy of an object. In a biological specimen it has been found that this can be estimated to about a tenth of the corresponding PSF-FWHM. Thus, the accuracy of distance determination for objects, that are more than 1 FWHM apart and possess the same spectral signature, is considerably better than the optical resolution (as low as ±20 nm) [61, 62].

21.8 Perspectives of three-dimensional light microscope

Microscopy is by no means a mature technological area. Every aspect is influenced by recent developments: cameras will replace oculars and film, new solid state lasers and lamps based on LEDs will have dramatic impacts, computers will control the position and state of many currently fixed elements, optical elements will combine refractive and diffractive properties, and filters will be tunable by electric fields; these are just a few of the developments that we can expect. Another important development is the use of water immersion lenses. This will allow a better depth penetration, which could increase the number of slices per stack and hence the recording time. At the same time it adds to the amount of recorded data and the requirements for efficient data processing. Numerical methods that extract the desired information from an image will become more readily available. Their impact is not easily determinable since they compete with technological developments. The correct procedure is probably to understand the imaging process and the performance of the many elements involved and to use such methods to either correct low performance or to find new compromises.

Acknowledgments

I would like to thank Dr. Frank-Martin Haar for preparing the figures. Dr. Jim Swoger, Dr. Frank-Martin Haar, and Mr. Nicholas J. Salmon were so kind to critically review the manuscript.

21.9 References

[1] Streibl, N., (1984). Fundamental restrictions for 3-D light distribution. *Optik*, **66**:341–354.

[2] Streibl, N., (1985). Three-dimensional imaging by a microscope. *J. Opt. Soc. Am.*, **A 2**:121–127.

[3] Stelzer, E. H. K., (1995). The intermediate optical system of laser-scanning confocal microscopes. In *Handbook of Biological Confocal Microscopy*, J. B. Pawley, ed., pp. 139–154. New York: Plenum Press.

[4] Stelzer, E. H. K., (1997). Three-dimensional light microscopy. In *Handbook of Microscopy: Applications in Materials Science, Solid-State Physics and Chemistry*, S. Amelinckx, D. v. Dyck, J. v. Landuyt, and G. v. Tendeloo, eds., pp. 71–82. Weinheim, New York, Basel, Cambridge: VCH-Verlag.

[5] Slomba, A. F., Wasserman, D. E., Kaufman, G. I., and Nester, J. F., (1972). A laser flying spot scanner for use in automated fluorescence antibody instrumentation. *Journal of the Association for the Advancement of Medical Instrumentation*, **6(3)**:230–234.

[6] Wilke, V., (1983). Laser scanning in microscopy. *SPIE Proc.*, **397**:164–172.

[7] Marsman, H. J. B., Stricker, R., Resandt, R. W. W. v., Brakenhoff, G. J., and Blom, P., (1983). Mechanical scan system for microscopic applications. *Rev. Sci. Instr.*, **54(8)**:1047–1052.

[8] Stelzer, E. H. K., Marsman, H. J. B., and van Resandt, R. W. W., (1986). A setup for a confocal scanning laser interference microscope. *Optik*, **73(1)**: 30–33.

[9] Voort, H. T. M. v. d., Brakenhoff, G. J., Valkenburg, J. A. C., and Nanninga, N., (1985). Design and use of a computer controlled confocal microscope for biological applications. *Scanning*, **7**:66–78.

[10] Wijnaendts van Resandt, R. W., Marsman, H. J. B., Kaplan, R., Davoust, J., Stelzer, E. H. K., and Stricker, R., (1985). Optical fluorescence microscopy in three dimensions: microtomoscopy. *J. Microsc.*, **138(1)**:29–34.

[11] Stelzer, E. H. K., (1994). Designing a confocal fluorescence microscope. In *Computer-assisted Multidimensional Microscopies*, P. C. Cheng, T. H. Lin, W. L. Wu, and J. L. Wu, eds., pp. 33–51. New York: Springer.

[12] Born, M. and Wolf, E., (1980). *Principles of Optics*. Oxford: Pergamon Press.

[13] Wilson, T. and Sheppard, C. J. R., (1984). *Theory and Practice of Scanning Optical Microscopy*. London: Academic Press.

[14] Sheppard, C. J. R. and Choudhury, A., (1977). Image formation in the scanning microscope. *Opt. Acta*, **24(10)**:1051–1073.

[15] Cox, I. J., Sheppard, C. J. R., and Wilson, T., (1982). Super-resolution by confocal fluorescent microscopy. *Optik*, **60(4)**:391–396.

[16] Brakenhoff, G. J., Blom, P., and Barends, P., (1979). Confocal scanning light microscopy with high aperture immersion lenses. *J. Microsc.*, **117(2)**:219–232.

[17] Hell, S., Witting, S., Schickfus, M. v., Resandt, R. W. W. v., Hunklinger, S., Smolka, E., and Neiger, M., (1991). A confocal beam scanning white-light microscope. *J. Microsc.*, **163(2)**:179–187.

[18] Hamilton, D. K. and Sheppard, C. J. R., (1982). A confocal interference microscope. *Opt. Acta*, **29(12)**:1573–1577.

[19] Sheppard, C. J. R., Hamilton, D. K., and Matthews, H. J., (1988). Confocal interference microscopy. In *Scanning Imaging Technology*, Vol. 1028, pp. 92–95. Los Angeles: SPIE Press.

[20] Wilson, T. and Juskaitis, R., (1994). Scanning interference microscopy. *Bioim.*, **2**:36–40.

[21] Wilson, T., Juskaitis, R., Rea, N. P., and Hamilton, D. K., (1994). Fibre optic interference and confocal microscopy. *Opt. Commun.*, **110**:1–6.

[22] Nipkow, P., (1884). *Elektrisches Teleskop*. Germany: Kaiserlisches Patentamt (Deutsches Patentamt).

[23] Petrán, M., Hadravsky, M., Egger, M. D., and Galambos, R., (1968). Tandem-scanning reflected-light microscope. *J. Opt. Soc. Am.*, **58**:661–664.

[24] Kino, G. S., (1995). Intermediate optics in Nipkow disk microscopes. In *Handbook of Biological Confocal Microscopy*, J. B. Pawley, ed., pp. 155–165. New York: Plenum Press.

[25] Yin, S., Lu, G., Zhang, J., Yu, F. T. S., and Mait, J. N., (1995). Kinoform-based nipkow disk for a confocal microscope. *Appl. Opt.*, **34(25)**:5695–5698.

[26] Sheppard, C. J. R. and Wilson, T., (1981). The theory of the direct-view confocal microscope. *J. Microsc.*, **124**:107–117.

[27] Agard, D. A., (1983). A least-squares method for determining structure factors in three-dimensional tilted-view reconstructions. *J. Mol. Biol.*, **167(4)**:849–852.

[28] Agard, D. A., (1984). Optical sectioning microscopy: cellular architecture in three dimensions. *Ann. Rev. Biophys. Bioeng.*, **13**:191–219.

[29] Agard, D. A. and Sedat, J. W., (1983). Three-dimensional architecture of a polytene nucleus. *Nature*, **302**:676–681.

[30] Shaw, P. J. and Rawlins, D. J., (1991). The point-spread function of a confocal microscope: its measurement and use in deconvolution of 3-D data. *J. Microsc.*, **163(2)**:151–165.

[31] Carrington, W. A., (1994). Advances in computational fluorescence microscopy. In *Proceedings of the 52nd Annual Meeting of the Microscopy Society of America*, G. W. Bailey and A. J. Garratt-Reed, eds., pp. 926–927. New Orleans, Louisiana: San Francisco Press, Inc.

[32] Carrington, W. A., Lynch, R. M., Moore, E. D. W., Isenberg, G., Fogarty, K. E., and Fay, F. S., (1995). Superresolution three-dimensional images of fluorescence in cells with minimal light exposure. *Science*, **268**:1483–1487.

[33] Hiraoka, Y., Sedat, J. W., and Agard, D. A., (1990). Determination of three-dimensional imaging properties of a light microscope system. Partial confocal behavior in epifluorescence microscopy. *Biophys. J.*, **57(2)**:325–333.

[34] Holmes, T. J., (1992). Blind deconvolution of quantum limited incoherent imagery: maximum-likelihood approach. *J. Opt. Soc. Am.*, **A 9(7)**:1052–1061.

[35] Dirnagl, U., Villringer, A., and Einhaupl, K. M., (1992). In vivo confocal scanning laser microscopy of the cerebral microcirculation. *J. Microsc.*, **165**:147–157.

[36] Zink, D., Cremer, T., Saffrich, R., Fischer, R., Trendelenburg, M. F., Ansorge, W., and Stelzer, E. H. K., (1998). Structure and dynamics of human interphase chromosome territories in vivo. *Hum. Genet.*, **102**:241–251.

[37] Hell, S., Reiner, G., Cremer, C., and Stelzer, E., (1993). Aberrations in confocal fluorescence microscopy induced by mismatches in refractive index. *Jour. Microscopy*, **169(Pt3)**:391–405.

[38] Török, P., Hewlett, S. J., , and Varga, P., (1997). The role of specimen-induced spherical aberration in confocal microscopy. *J. Microsc.*, **188(2)**: 158–172.

[39] Török, P., Varga, P., Konkol, A., and Booker, G. R., (1996). Electromagnetic diffraction of light focused through a planar interface between materials of mismatched refractive indices: structure of the electromagnetic field II. *J. Opt. Soc. Am.*, **A 13(11)**:2232–2238.

[40] Visser, T. D., Groen, F. C. A., and Brakenhoff, G. J., (1991). Absorption and scattering correction in fluorescence confocal microscopy. *J. Microsc.*, **163**:189–200.

[41] White, N. S., Errington, R. J., Fricker, M. D., and Wood, J. L., (1996). Multidimensional fluorescence microscopy: optical distortions in quantitative imaging of biological specimens. In *Fluorescence Microscopy and Fluorescent Probes*, J. Slavík, ed., pp. 47–56. New York: Plenum Press.

[42] Hell, S. W. and Stelzer, E. H. K., (1995). Lens aberrations in confocal fluorescence microscopy. In *Handbook of Biological Confocal Microscopy*, J. B. Pawley, ed., pp. 347–354. New York: Plenum Press.

[43] Göppert-Mayer, M., (1931). Über Elementarakte mit zwei Quantensprüngen. *Ann. Phys.*, **9**:273–294.

[44] Kaiser, W. and Garrett, C. G. B., (1961). Two-photon excitation in CaF_2:Eu^{2+}. *Phys. Rev. Lett.*, **7**:229–231.

[45] Sheppard, C. J. R. and Gu, M., (1990). Image formation in two-photon fluorescence microscopy. *Optik*, **86(3)**:104–106.

[46] Stelzer, E. H. K., Hell, S., Lindek, S., Stricker, R., Pick, R., Storz, C., Ritter, G., and Salmon, N., (1994). Nonlinear absorption extends confocal fluorescence microscopy into the ultra-violet regime and confines the illumination volume. *Opt. Commun.*, **104**:223–228.

[47] Denk, W., Strickler, J., and Webb, W., (1990). Two-photon laser scanning fluorescence microscopy. *Science*, **248**:73–76.

[48] Denk, W., Strickler, J. P., and Webb, W. W., (1991). *Two-photon laser microscopy. United States Patent.* Switzerland, New York: Cornell Research Foundation, Inc.

[49] Hänninen, P. E., Soini, E., and Hell, S. W., (1994). Continuous wave excitation two-photon fluorescence microscopy. *J. Microsc.*, **176(3)**:222–225.

[50] Hell, S. W., Hänninen, P. E., Salo, J., Kuusisto, A., Soini, E., Wilson, T., and Tan, J. B., (1994). Pulsed and cw confocal microscopy: a comparison of resolution and contrast. *Opt. Commun.*, **113**:144–152.

[51] Fischer, A., Cremer, C., and Stelzer, E. H. K., (1995). Fluorescence of coumarins and xanthenes after two-photon absorption with a pulsed titanium-sapphire laser. *Appl. Opt.*, **34(12)**:1989–2003.

[52] Xu, C. and Webb, W. W., (1996). Measurement of two-photon excitation cross sections of molecular fluorophores with data from 690 to 1050 nm. *J. Opt. Soc. Am.*, **B 13(3)**:481–491.

[53] Bailey, B., Farkas, D. L., Taylor, D. L., and Lanni, F., (1993). Enhancement of axial resolution in fluorescence microscopy by standing-wave excitation. *Nature*, **366**:44–48.

[54] Hell, S. and Stelzer, E. H. K., (1992). Properties of a 4Pi confocal fluorescence microscope. *J. Opt. Soc. Am.*, **A 9(12)**:1259–2166.

[55] Stelzer, E. H. K. and Lindek, S., (1994). Fundamental reduction of the observation volume in far-field light microscopy by detection orthogonal to the illumination axis: confocal theta micrsocopy. *Opt. Commun.*, **111**: 536–547.

[56] Lindek, S., Stelzer, E. H. K., and Hell, S., (1995). Two new high-resolution confocal fluorescence microscopies (4Pi, Theta) with one- and two-photon excitation. In *Handbook of Biological Confocal Microscopy*, J. B. Pawley, ed., pp. 417–430. New York: Plenum Press.

[57] Burns, D. H., Callis, J. B., Christian, G. D., and Davidson, E. R., (1985). Strategies for attaining superresolution using spectroscopic data as constraints. *Appl. Opt.*, **24(2)**:154–161.

[58] Saxton, M. J. and Jacobson, K., (1997). Single-particle tracking: applications to membrane dynamics. *Ann. Rev. Biophys. Biomol. Struct.*, **26**: 373–399.

[59] Florin, E.-L., Pralle, A., Hörber, J. K. H., and Stelzer, E. H. K., (1997). Photonic force microscope based on optical tweezers and two-photon excitation for biological applications. *Jour. of Structural Biology*, **119**:202–211.

[60] Wijnaendts van Resandt, R. W., (1987). Application of confocal beam scanning microscopy to the measurement of submicron structures. In *Sanning Imaging Technology*, Vol. 809 of *SPIE Proc.*, pp. 101–106. B ellingham, WA: SPIE Press.

[61] Bradl, J., Rinke, B., Krieger, H., Schneider, B., Haar, F.-M., Durm, M., H., M., and Cremer, C., (1996). Comparative study of three-dimensional distance resolution in conventional, confocal-laser-scanning and axial tomographic fluorescence light microscopy. In *BIOS Europe ´96: Optoelectronic Research and Techniques: Optical Microscopic Techniques*, Vol. 2926 of *SPIE Proc.* Bellingham, WA: SPIE Press.

[62] Bradl, J., Rinke, B., Schneider, B., Edelmann, P., Krieger, H., Hausmann, M., and Cremer, C., (1996). Resolution improvement in 3-D microscopy by object tilting. *Europ. Microsc. Anal.*, **44**:9–11.

22 Magnetic Resonance Imaging in Medicine

Wolfgang G. Schreiber[1] and Gunnar Brix[2]

[1]Johannes Gutenberg-Universität, Mainz, Germany
[2]Deutsches Krebsforschungszentrum Heidelberg (DKFZ), Germany

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

22.1 Introduction

Magnetic resonance (MR) is a rapidly developing noninvasive diagnostic imaging modality in medicine. Although the principles of nuclear magnetic resonance were well known from the pioneering work of Bloch [1] and Edward Purcell in 1946 [2] it was only after Lauterbur's paper in 1973 [3] that it was realized that nuclear magnetic resonance (NMR) could be used for diagnostic imaging. This was achieved by adding to the homogenous magnetic field (in which the nuclear magnetic resonance effect takes place) small position dependent (gradient) magnetic fields. The origin of the re-emitted radiation can be traced back on the basis of the emitted frequency, which makes, in principle, imaging possible. Lauterbur's work was preceded by a patent by Damadian in 1972 [4], in which the clinical use of NMR was anticipated. These inventions triggered enormous activity in realizing NMR systems for use in hospitals and in the application of NMR techniques to medical diagnostics. The term "nuclear" is not commonly used because of its association with nuclear warfare and nuclear radiation. The widely accepted name for the new imaging technique is magnetic resonance imaging (MRI). This contribution should represent a comprehensive introduction to the concepts of magnetic resonance imaging for medical application. It is organized in the form of a systematic exploration of the principles of MRI, starting from basic physics of a particle with nuclear spin and the basic MR experiment. From a description of the main methods for spatial labeling of a spin, the standard imaging equations for the 2-D and 3-D MR pulse sequences will be derived. Further discussion will be on the latest technical development for fast and ultrafast image acquisition by pulse sequences with advanced methods for scanning $\boldsymbol{k}$-space.

22.2 Basic magnetic resonance physics

22.2.1 Particle with spin

Spin $\boldsymbol{I}$ is a fundamental property of a particle-like electrical charge or mass. Spin comes in multiples of $1/2$ and can be positive or negative. Protons, electrons, and neutrons possess spin of $\boldsymbol{I} = 1/2$. The spin can be thought of as leading to a circulating electric current and, hence, an associated magnetic moment $\boldsymbol{J} = \hbar\boldsymbol{I}$. The direct relationship between the magnetic moment $\boldsymbol{\mu}$ and the spin $\boldsymbol{I}$ is found from experiment,

$$\boldsymbol{\mu} = \gamma\hbar\boldsymbol{I} \tag{22.1}$$

where $\hbar = 1.055 \times 10^{-34}\mathrm{Ws}^2$ denotes Planck's constant divided by 2π. γ is called the *gyromagnetic ratio*, and depends on the type of particle

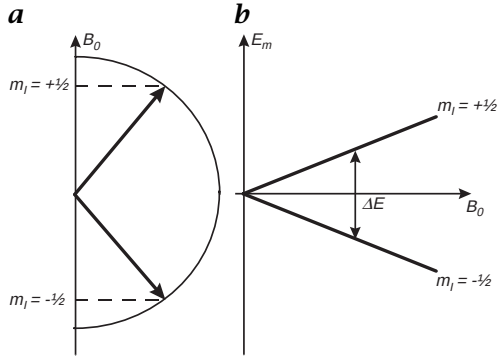


Figure 22.1: ***a*** *Orientation of a spin* $\boldsymbol{I} = 1/2$ *particle in a magnetic field along the z-axis.* ***b*** *Energy of the spin as a function of the magnetic field strength. Transitions between the* $m_I = -1/2$ *and the* $m_I = +1/2$ *energy levels can be induced by an electromagnetic field with energy* $\Delta E = \hbar\omega_0$.

or nucleus. For the proton, it is found to be $\gamma = 42.6$ MHz/T, where T is the Tesla unit of magnetic field.

In an external magnetic field $\boldsymbol{B}_0$ a spin with the magnetic moment of Eq. (22.1) has the potential energy

$$U = -\boldsymbol{\mu}\boldsymbol{B}_0 \tag{22.2}$$

Particles with spin $\boldsymbol{I} = 1/2$ can only be in two discrete orientations with respect to the external magnetic field (see Fig. 22.1a). A quantum mechanical treatment shows that the z-component of the spin can have only the distinct values $\hbar m_I$. If quantum numbers $m_I = +1/2$ and $m_I = -1/2$ are assigned to these orientations, the energy E_m of the spin in an external magnetic field $\boldsymbol{B}_0$ is

$$E_m = -\boldsymbol{\mu}\boldsymbol{B}_0 = -\hbar\gamma m_I B_0 \tag{22.3}$$

Thus two energy levels exist in the case of a spin $\boldsymbol{I} = 1/2$ particle between which transitions can be induced by absorption of a photon of energy:

$$\Delta E = \hbar\omega_0 = E_{m=+1/2} - E_{m=-1/2} = \gamma\hbar B_z \tag{22.4}$$

The photon energy depends on the gyromagnetic ratio of the particle (i. e., the type of particle observed) and of the main magnetic field strength. The energy is 8.4×10^{-8} eV for a proton in a magnetic field of 1.5 T corresponding to a frequency of 63.9 MHz. The frequency ω_0 is termed the *Larmor frequency*.

In matter an ensemble of nuclei is observed (1 mm^3 water contains 6.7×10^{19} protons). In *thermal equilibrium*, the two energy levels will

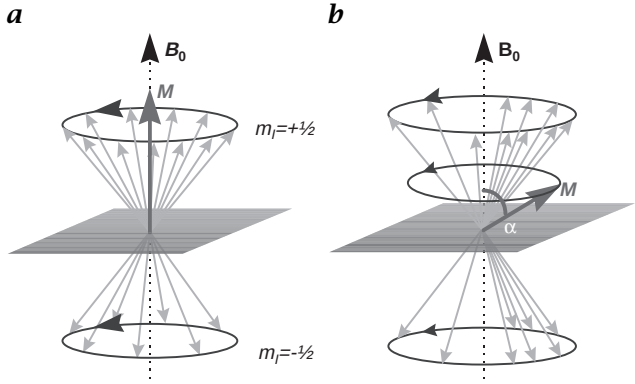


Figure 22.2: ***a*** *Magnetization* ***M*** *of an ensemble of spin 1/2 particles under conditions of thermal equilibrium. Population of the* $m_I = +1/2$ *state is higher than of the* $m_I = -1/2$ *state.* ***b*** *Ensemble of spins after irradiation with RF of energy* $\Delta E = \hbar\omega_0$*. Spins are flipped from the* $m_I = +1/2$ *to the* $m_I = -1/2$ *state, and procession takes place with a higher degree of phase coherence. As a result,* ***M*** *is rotated by an angle* α *from the z-direction.*

be populated by the protons according to the Boltzmann probability factor

$$\frac{N_{m=-1/2}}{N_{m=+1/2}} = \exp\left(\frac{-\gamma\hbar B}{k_B T}\right) \tag{22.5}$$

where $k_B = 1.38 \times 10^{-23}$ Ws/K denotes the Boltzmann factor and T the absolute temperature. Because more spins will be in energetically lower state ($N_{m=+1/2}$) than in the higher state ($N_{m=-1/2}$), a small population difference $N_{m=+1/2} - N_{m=-1/2}$ will arise in thermal equilibrium. As a consequence, a probe of N protons in a volume V has a total magnetic moment m_N. The total magnetic moment per unit volume

$$M = \frac{\Delta m_N}{\Delta V} \tag{22.6}$$

is denoted as the *magnetization*.

At body temperature the energy difference $\gamma\hbar B_0$ is about six orders of magnitude smaller than the Boltzmann energy $k_B T$. In consequence, Eq. (22.5) may be linearized yielding the magnetization in *thermal equilibrium*

$$M = \frac{\gamma^2\hbar^2}{4}\frac{N}{V}\frac{B_0}{k_B T} \tag{22.7}$$

This expression is *Curie's law*. It should be noted at this point that it is the large proton-density N/V that makes possible the observation of an MR signal from the human body. In medical applications of magnetic resonance, no signal is observed from gas-filled spaces

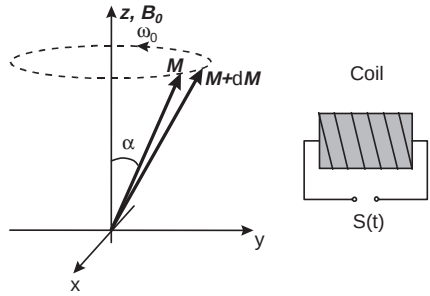


Figure 22.3: *Precession of the magnetization vector* $\boldsymbol{M}$ *around the direction of the main magnetic field* $\boldsymbol{B}_0$*. The varying transverse component of the magnetization induces in the coil an oscillatory voltage with frequency* ω_0*.*

like the lung because of the more than four orders of magnitude lower proton-density of gases, and because the proton-content in air is low[1]. Irradiating the spin system with a transverse (i. e., perpendicular to $\boldsymbol{B}_0$) high-frequency magnetic field $\boldsymbol{B}_1$ of frequency $f = \Delta E/2\pi\hbar$ leads to a transient increase of the population of the $m_I = -1/2$ at the cost of population of the mI = +1/2 state (Fig. 22.2)[2]. In addition, a higher degree of phase coherence between the spins of the particles occurs, which causes the magnetization vector to rotate by an angle α.

The exact behavior of the spin system should be treated in terms of quantum mechanics. Fortunately, the coupling of nuclear spins mutually and with the surrounding matter is weak, which allows a classical treatment of most phenomena (in the context of this section) on the basis of the equation extended with terms describing the relaxation in a phenomenological way [1].

22.2.2 Bloch equations

When magnetization $\boldsymbol{M}$ has another direction than the main magnetic field $\boldsymbol{B}_0$, a torque is exerted on the magnetization, which is perpendicular to both the magnetic field and the magnetization. The *Bloch equations* [1] describe this interaction in a phenomenological way:

$$\frac{d\boldsymbol{M}}{dt} = \gamma(\boldsymbol{M} \times \boldsymbol{B}) \tag{22.8}$$

[1] In the last years, imaging of the gas-filled spaces in the lung has been performed by using *hyperpolarized noble gases* like ^{3}He and ^{129}Xe [5, 6]. By optical pumping and spin exchange processes polarizations of up to 0.3 can be achieved for these gases. It has been demonstrated that images of the airways and of the alveolar spaces in the lung can be generated if these gases are inspired by the patient.

[2] B_1 is also referred to as an RF-field as its frequency for a main field of $B_0 = 1.5\,\mathrm{T}$ and protons is 63.9 MHz, that is, in the radiofrequency range.

The *Bloch equation* tells us that the vector $d\boldsymbol{M}/dt$ is always oriented perpendicular to the plane of $\boldsymbol{B}$ and $\boldsymbol{M}$, so the point of $\boldsymbol{M}$ moves in a circular path, it precesses around the z-axis (Fig. 22.3). The solution to Eq. (22.8) for $\boldsymbol{B}_0 = [0, 0, B_0]^T$ corresponds to a precession of the magnetization about the direction of the field at a rate

$$\omega_0 = \gamma B_0 \tag{22.9}$$

the *Larmor frequency*. If a coil is positioned in the configuration as shown in Fig. 22.3, the temporally changing magnetic transverse component of the magnetization induces in the coil an oscillatory voltage. The amplitude of the signal is proportional to the transverse component of the magnetization. The induced voltage is typically of order μV.

In practical situations we deal with three different magnetic fields, adding up to $\boldsymbol{B}$. The first is the main magnetic field in the z-direction, $\boldsymbol{B}_0 + \Delta\boldsymbol{B}$, where $\Delta\boldsymbol{B}$ denotes a field inhomogeneity of the scanner or due to the magnetic properties of the tissue in the patient. The second is the *gradient field* $(\boldsymbol{G}\boldsymbol{r})\tilde{\boldsymbol{n}}_z$ used for magnetic labeling, the position of a spin (Section 22.3.1), where $\tilde{\boldsymbol{n}}_z$ denotes the unit vector in z-direction, $\boldsymbol{G}$ is a gradient field, and $\boldsymbol{r}$ is the position. The third field is the magnetic component of the transverse magnetic field $\boldsymbol{B}_1$ used to excite the spin system and to rotate the magnetization. Therefore, the Bloch equation can be rewritten

$$\frac{d\boldsymbol{M}}{dt} = \gamma \boldsymbol{M} \times [\boldsymbol{B}_0 + \Delta\boldsymbol{B} + (\boldsymbol{G}\boldsymbol{r})\tilde{\boldsymbol{n}}_z + \boldsymbol{B}_1] \tag{22.10}$$

where the first part describes the precession of the magnetization due to the main magnetic field, and the second part describes precession due to the field inhomogeneity, the gradient field, and the RF field. If a new system of reference is introduced that rotates around $\boldsymbol{B}_0$ with the angular frequency given by Eq. (22.9), precession due to $\boldsymbol{B}_0$ is eliminated:

$$\frac{d\boldsymbol{M}'}{dt} = \gamma \boldsymbol{M}' \times [\Delta\boldsymbol{B} + (\boldsymbol{G}\boldsymbol{r})\tilde{\boldsymbol{n}}_z + \boldsymbol{B}_1] \tag{22.11}$$

where the prime denotes the magnetization in the *rotating frame of reference*. Note that z- and z'-components are equal because the rotating frame of reference is rotating around the z-axis.

It will be shown in Section 22.2.5 that by *heterodyne* signal detection, that is, mixing of the input signal with that of an oscillating reference signal of frequency ω_0, it is this equation we have to deal with in MRI. Therefore we shall make further discussion of MR on the assumption of observation in the rotating frame of reference, that is, further discussions will use Eq. (22.11) as a basis.

In the equilibrium situation with $\boldsymbol{M} = [0, 0, M_z]^T$ and when $\boldsymbol{B}_1$ and $\boldsymbol{G}$ are zero, there is no torque on the magnetization and the magnetization vector remains parallel to the main magnetic field.

The inhomogeneity term $\Delta\boldsymbol{B}$ and, as the amplitude of typical $\boldsymbol{B}_1$-fields (10 μT) is much smaller than that of the main magnetic field (1.5 T), its z-component will also be neglected. With these assumptions, the *Bloch equation* in the rotating frame of reference Eq. (22.11) can be reformulated as a matrix equation:

$$\frac{d\boldsymbol{M}'}{dt} = \gamma \begin{bmatrix} 0 & \boldsymbol{Gr} & -B_{1y'} \\ -\boldsymbol{Gr} & 0 & B_{1x'} \\ B_{1y'} & -B_{1x'} & 0 \end{bmatrix} \begin{bmatrix} M'_x \\ M'_y \\ M'_z \end{bmatrix} \tag{22.12}$$

Because all further discussion will be made in the rotating frame of reference, the primes will be omitted.

22.2.3 Excitation by radio frequency pulses

We now come back to the situation in which the spin system is excited by a transverse radio frequency (RF) field $\boldsymbol{B}_1$ (Fig. 22.2). Consider this RF field to be irradiated at frequency $\omega_0 = \gamma B_0$, and suppose the duration of the irradiation τ_{RF} is short enough to neglect the influence of the gradient field, that is, $\tau_{RF} \ll 1/\gamma\boldsymbol{Gr}$. Then, in the rotating frame of reference, the RF-field has a constant value, for example, $(B_{1x}, 0, 0)^T$. Insertion of this RF field into Eq. (22.12) leads to two coupled differential equations. Upon elimination of M_z, we find the second-order differential equation

$$\frac{d^2 M_y}{dt^2} = -\gamma^2 B_{1x}^2 M_y \tag{22.13}$$

with the well-known solution

$$M_y(t) = A \sin(\gamma B_{1x} t) + B \cos(\gamma B_{1x} t) \tag{22.14}$$

where the complex constants A and B depend on the initial conditions. A similar solution can be written for $M_x(t)$. Combining both solutions, assuming thermal equilibrium and M_0 from Eq. (22.7) as initial condition, then

$$\begin{bmatrix} M_x(t) \\ M_y(t) \\ M_z(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\gamma B_{1x} t & \sin\gamma B_{1x} t \\ 0 & -\sin\gamma B_{1x} t & \cos\gamma B_{1x} t \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ M_0 \end{bmatrix} \tag{22.15}$$

This is the well-known rotation matrix, describing a rotation of the magnetization $\boldsymbol{M}$ around the x-axis. After irradiation of duration τ_{RF},

the magnetization is rotated by

$$\alpha = \gamma B_{1x} \tau_{RF} \tag{22.16}$$

The angle α is commonly referred to as the *flip angle*. It depends on the duration of irradiation τ_{RF} and on the field strength B_1 of the RF magnetic field. In a more general notation, assuming a shaped RF pulse with nonconstant amplitude, the flip angle is:

$$\alpha = \gamma \int_0^{\tau_{RF}} B_1(t)\,dt \tag{22.17}$$

If $B_1(t)$ and τ_{RF} are chosen to rotate the magnetization by 90° (180°), this RF pulse is denoted as a 90°- (180°-) pulse. Any flip angle between 0° and 360° can be achieved by either changing the amplitude of $B_1(t)$ or the duration τ_{RF}. After the irradiation of a 90°-pulse the longitudinal magnetization has been converted completely into transverse magnetization. After a 180°-pulse along the x-direction, the y- and the z-components of the magnetization before the pulse $(M_x^0, M_y^0, M_z^0)^T$ are inverted:

$$\begin{bmatrix} M_x \\ M_y \\ M_z \end{bmatrix} \xrightarrow{180^\circ \text{ pulse}} \begin{bmatrix} M_x \\ -M_y \\ -M_z \end{bmatrix} \tag{22.18}$$

Thus, a 180° pulse is often termed as an inversion pulse.

22.2.4 $T1$- and $T2$-relaxation

In the nonequilibrium situation, when magnetization is not parallel to the z-axis, two processes exist that drive the magnetization back to the equilibrium situation. These processes are referred to as *relaxation* processes. Although relaxation should be treated by a quantum mechanical description, an intuitive macroscopic picture can demonstrate the basic mechanism of relaxation.

Consider a sample of nuclei. Even in a perfect main magnetic field these nuclei will experience somewhat different local magnetic fields because of the magnetic interaction with the magnetic moments of other nuclei or electrons. Thus, the Larmor frequency for each nucleus is somewhat different from that of the other nuclei.

Let us now define a *spin packet* as a group of spins experiencing the same magnetic field strength. All nuclei of this spin packet will precess with the same Larmor frequency. Thus, the magnetic field due to the spins in each spin packet can be represented by a magnetization vector. The vector sum of the magnetization vectors from all of the spin packets is the *net magnetization* Eq. (22.7).

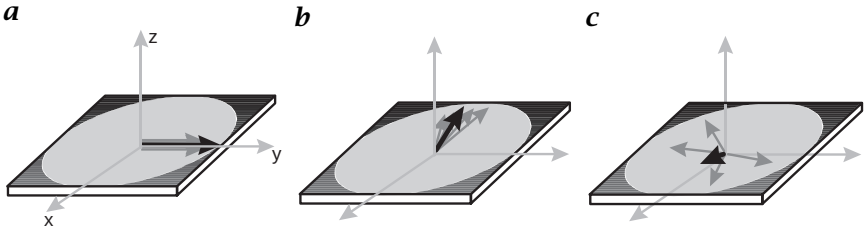


Figure 22.4: *Precession of spin packets (gray arrows) in the presence of T2-relaxation.* ***a*** *Immediately after the 90° -pulse, all spin packets are in-phase.* ***b*** *Due to slightly different local magnetic fields, the spin packets precess with different Larmor frequencies. Thus, phase dispersion occurs between spin packets. The net magnetization (black arrow) is somewhat reduced.* ***c*** *After some time T2 dispersion between the phases of the spin packets has further increased. The net magnetization is largely reduced.*

Suppose that the longitudinal magnetization has been converted completely into transverse magnetization by a 90° -pulse (Fig. 22.4a). Different spin packets will experience different local magnetic fields and, thus, will precess with different Larmor frequencies. Thus, after some time, dispersion between the phases of the magnetization vectors of the spin packets will occur (Fig. 22.4b,c). As the net magnetization is the vector sum of the magnetization vectors of the spin packets, the net magnetization is reduced. Because this relaxation process is caused mainly by interaction between the spins of individual nuclei, it is denoted as *spin-spin relaxation* or *T2*-relaxation. Using the complex notation for the transverse magnetization

$$M_T = M_x + \mathrm{i}M_y \tag{22.19}$$

spin-spin relaxation in a first approximation can be assumed to happen in a simple exponential manner:

$$M_T = M_{RF} \exp\left(-\frac{t}{T2}\right) \tag{22.20}$$

where M_{RF} denotes the transverse magnetization immediately after the RF-pulse.

The source for $T2$ relaxation is a microscopic magnetic field inhomogeneity by the magnetic interaction between the spins. A similar effect is caused by the macroscopic *field inhomogeneity* $\Delta \boldsymbol{B}$ introduced by technical imperfections of the scanner or by magnetic *susceptibility* changes within the patient. In a first approximation, it may also be assumed that these inhomogeneities result in an exponential decay of the transverse magnetization, which is now described by a time constant

$T2^*$

$$M_T = M_{RF} \exp\left(-\frac{t}{T2^*}\right) \tag{22.21}$$

with the definition

$$\frac{1}{T2^*} = \frac{1}{T2} + \gamma \Delta B \tag{22.22}$$

After irradiation with an RF field the spin system is in an excited state, from which it attempts to return to thermal equilibrium by de-excitation via a resonant irradiation (such as due to an externally applied RF field), or via the fluctuating magnetic field of nuclear spins in the surrounding, the *lattice*. (Spontaneous emission of photons is of no relevance for relaxation because the transition probability is low due to the low Larmor frequency.) Therefore, this process is termed *spin-lattice relaxation* (or, according to the time-constant describing this process, *T1 relaxation*), and can—in a first approximation—be described by a monoexponential increase of the magnetization:

$$M_z(t) = M_0 + \{M_z(0) - M_0\} \exp\left(-\frac{t}{T1}\right) \tag{22.23}$$

where $M_z(0)$ denotes the initial longitudinal magnetization.

Combining Eq. (22.20) and Eq. (22.23) with Eq. (22.12) results in the matrix formulation of the Bloch equation including relaxation processes.

$$\frac{\partial \boldsymbol{M}}{\partial t} = \begin{bmatrix} -1/T2 & \gamma \boldsymbol{Gr} & -\gamma B_{1y} \\ -\gamma \boldsymbol{Gr} & -1/T2 & \gamma B_{1x} \\ \gamma B_{1y} & -\gamma B_{1x} & -1/T1 \end{bmatrix} \begin{bmatrix} M_x \\ M_y \\ M_z \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ M_0/T_1 \end{bmatrix} \tag{22.24}$$

22.2.5 Signal detection

The MR signal is measured by *heterodyne detection*, that is, mixing the observed MR signal with the output of an electronic reference oscillator. It can be shown [7] that the high Larmor frequency is eliminated from the MR signal if the frequency of the reference oscillator is set equal to the Larmor frequency ω_0. In this case, only differences from ω_0 show up as an offset frequency in the MR signal. An exact on-resonance condition leads to a constant signal (zero frequency). Heterodyne signal detection has two consequences, which at a first glance may appear unusual: first, negative frequencies may appear; second, heterodyne mixing is inherently *phase sensitive*, thus the MR signal must be described by a complex number Eq. (22.19). The magnitude of the complex signal is calculated and displayed after image reconstruction.

Because of heterodyne mixing it is adequate in further discussion of MR to assume to be observing in the rotating frame of reference, that is, we will describe MR image reconstruction on the basis of Eq. (22.11).

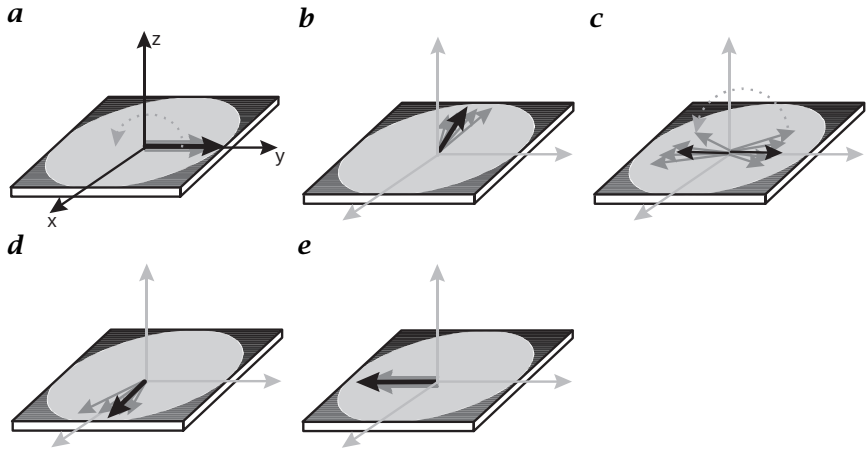


Figure 22.5: *Generation of a* spin echo. *From* ***a*** *to* ***b*** *the spin packets accumulate phase. This phase differs for the spin packets because magnetic field inhomogeneities* $\Delta \boldsymbol{B}$ *lead to differences in the Larmor frequency.* ***c*** *At* $t = TE/2$*, a 180°-pulse inverts the* y*-component of the magnetization vectors and the accumulated phase for all spin packets and, thus, inverts also the net magnetization vector.* ***d*** *Spin packets accumulate phase with the same rate as before the 180°-pulse.* ***e*** *At* $t = TE$ *all spin packets are in phase again, an echo has been generated. The net magnetization, however, is reduced only due to* $T2$ *decay.*

22.2.6 Spin echo

In the circumstance where the magnetic field inhomogeneity $\Delta \boldsymbol{B}$ can not be neglected, $T2^*$ is much shorter than $T2$ (Eq. (22.22)) resulting in severe signal loss. However, this inhomogeneity-induced dephasing is reversible by the generation of a *spin echo* [8]. The basic mechanism of generating a spin echo is shown in Fig. 22.5. It relies on the application of two RF pulses: a 90°-pulse for excitation of transverse magnetization, followed by a 180°-pulse for refocusing the dispersed *spin packets*. In the time interval between the two pulses the spin packets perform precession at their local Larmor frequency ω_0. As the magnetic field is considered inhomogeneous, the Larmor frequencies will be different for the spin packets. Thus, the accumulated phase of the spin packets at time $t = TE/2$ will also vary.

The 180°-pulse inverts ('refocuses') the y-component of the magnetization and the accumulated phase. After that pulse the spins continue to collect phase at the same rate as before the refocusing pulse. Spin packets precessing at a low ω_0 accumulate less phase than those at high ω_0. Therefore, at a time TE/2 after the 180°-pulse (i.e., at t = TE after the 90°-pulse) all spin packets are aligned again, but the net magnetization vector M is reduced due to the intrinsic T2 decay. The T2

decay cannot be refocused by the spin echo because the magnetic field strength fluctuations induced by other particles fluctuate so rapidly that the accumulated phase also changes with time.

22.3 Image acquisition and reconstruction

The principles discussed in the previous section were used for many years to characterize small probes in biochemistry, solid state physics, and materials science. It was only after Lauterbur's paper in 1973 [3] that it was realized that nuclear magnetic resonance could be used for medical imaging. This was achieved by adding to the homogeneous magnetic field (in which the nuclear magnetic resonance effect takes place) small position dependent (gradient) magnetic fields. Now the origin of the re-emitted radiation can be traced back on the basis of the emitted frequency, which makes, in principle, imaging possible. Lauterbur's work was preceded by a patent by Damadian in 1972 [4], in which the clinical use of MR was anticipated.

Lauterbur's concept was to acquire *projections* of the magnetization along different orientations as in other *tomographic methods*, and to reconstruct the images by the projection reconstruction method. In 1975, Kumar et al. [9] proposed MR imaging with orthogonal gradient fields and reconstruction by *(fast) Fourier transformation* (FFT). Today, this is the most commonly used method in medical MRI.

After introducing gradient fields and their use for spatial encoding, data acquisition schemes and image reconstruction of 2-D and 3-D Fourier-transform MRI will be described. The section will end with a discussion of basic image contrast types that can be obtained by a variation of the acquisition parameters.

22.3.1 Strategies for spatial encoding

In order to make image formation possible, spins at different locations in space must be labeled differently. The basic tool for doing so is given by the relation between the local magnetic field B_0 and the Larmor frequency ω_0 (Eq. (22.9)). If the magnetic field is not constant throughout the detected volume, the Larmor frequency of nuclei at position $\boldsymbol{r}$ will depend on the magnetic field $\boldsymbol{B}(\boldsymbol{r})$ at that location. Although it is the intention of the manufacturer and of the experimenter to make the main magnetic field $\boldsymbol{B}_0$ as homogeneous as possible, a deliberately introduced and controlled field inhomogeneity will provide a label of the position of the particle.

Gradient fields. A gradient field is a magnetic field parallel to $\boldsymbol{B}_0 = B_0\tilde{\boldsymbol{n}}_z$ ($\tilde{\boldsymbol{n}}_z$ is a unity vector in z-direction) whose magnitude varies

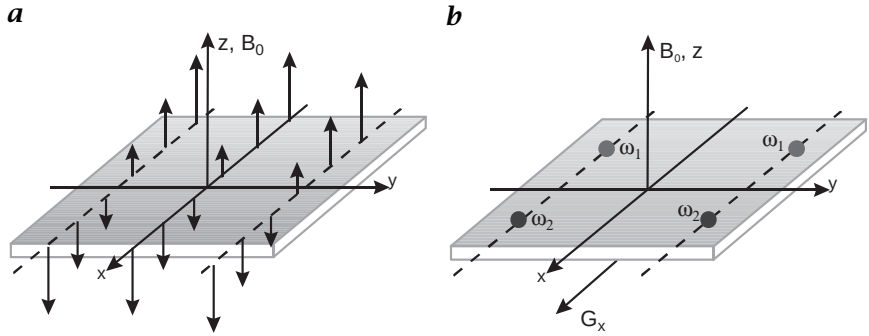


Figure 22.6: ***a*** *x-gradient. The gradient field is an additional magnetic field which is parallel to the main magnetic field* $\boldsymbol{B}_0$*. Its magnitude varies along the x-axis.* ***b*** *Spatial labeling of spins by* frequency encoding*. The x-gradient field leads to a variation of the Larmor frequencies* $\omega = 2\pi f$ *along the x-direction; Larmor frequencies for nuclei with equal x-position but different y- (or z-) position precess with equal Larmor frequencies and can not be distinguished by the x-gradient.*

along one direction in space (Fig. 22.6a), for example, along the x-direction:

$$\boldsymbol{G}_x(t) = xG_x(t)\tilde{\boldsymbol{n}}_z \tag{22.25}$$

When gradients along the x-, y-, and z-direction are turned on with a strength $\boldsymbol{G} = [G_x, G_y, G_z]^T$, the total magnetic field is a superposition of the three gradient fields and will be a function of the spatial location of the spin

$$\boldsymbol{B}(\boldsymbol{r}, t) = \boldsymbol{B}_0 + (\boldsymbol{G}(t)\boldsymbol{r})\,\tilde{\boldsymbol{n}}_z \tag{22.26}$$

where $\boldsymbol{r} = (x, y, z)^T$. It is this gradient field $\boldsymbol{G}(t)\boldsymbol{r}$ that induces the *spatial encoding* of the resonance frequency:

$$\omega(\boldsymbol{r}, t) = \gamma\,(B_0 + \boldsymbol{G}(t)\boldsymbol{r}) \tag{22.27}$$

For a fixed gradient vector $\boldsymbol{G}(t)$ two points in space, which are displaced from each other by a vector that is orthogonal to $\boldsymbol{G}(t)$ will have the same Larmor frequency (Fig. 22.6b), but, by applying a second gradient field with different orientation later these may be distinguishable. By transmitting enough different gradient fields all points may be made (almost) distinguishable, and this is what MRI imaging systems do.

Gradient fields are achieved by additional coils in the MR scanner, which generate these spatially (and temporally) varying magnetic fields. These coils are referred to as the x-, y- and z-gradient coils. The typical magnitude of the gradient fields in a modern medical MR scanner

is $G_{\max} = 24\,\mathrm{mT/m}$ with a typical slew rate of $\mathrm{d}G/\mathrm{d}t = 120\,\mathrm{mT/m/ms}$. For an object with extension of *field-of-view* (FOV) = 500 mm, the maximum additional field introduced by the gradient is 6.25 mT resulting in a maximum offset of the resonance frequencies of

$$\Delta f = \gamma G_{\max}\mathrm{FOV}/4\pi = 266\,\mathrm{kHz}$$

The receiver of the MR scanner must therefore have sufficiently high *bandwidth* to sample the signal of nuclei with these Larmor frequencies.

Slice selection. If *RF pulses* are applied without any magnetic fields in parallel, they excite protons in the whole volume seen by the RF coil. However, gradient fields can be used to restrict spatially the spins that are excited by the RF pulse to a certain slice.

Consider a z-gradient with magnitude G_z. If a slice is defined by $z_1 \le z \le z_2$, a RF pulse is needed that contains frequencies only in the range $f_1 \le f \le f_2$, where the frequencies and positions are related by $f_i = (2\pi)^{-1}\gamma G_z z_i$, $(i = 1, 2)$. Therefore, a frequency spectrum is needed that ideally has a constant amplitude for $f_1 \le f \le f_2$ and which has zero amplitude outside this region (Fig. 22.7). A pulse with these properties is the *sinc-pulse*

$$B_1(t) = B_1^{max}\frac{\sin(\gamma G_z d\tau_{RF}/2)}{\gamma G_z d\tau_{RF}/2} \tag{22.28}$$

where d denotes the *slice thickness*.

The Concept of k-space. Consider the nuclear spins at position $\boldsymbol{r}$ in the sample, occupying a small element of volume $d\boldsymbol{V}$. If the local spin density is $\rho(\boldsymbol{r})$ then there will be $\rho(\boldsymbol{r})dV$ spins in this element. In the rotating frame of reference, the *MR signal* from this volume element in the presence of a gradient $\boldsymbol{G}$ may be written as

$$S(t) = \int\int\int \rho(\boldsymbol{r})\exp[i\gamma\boldsymbol{G}\boldsymbol{r}t]\,\mathrm{d}\boldsymbol{r} \tag{22.29}$$

where the integration is over the whole sensitive volume of the receiving *coil*. Equation Eq. (22.29) makes no allowance for the decay of the signal due to transverse relaxation, an assumption that is fine provided the signal decay due to $\gamma\boldsymbol{G}\boldsymbol{r}$ is much more rapid than that due to $T2$ relaxation. The latter assumption is fulfilled in most sequences with the exception of *turbo spin echo* sequences with a high turbo-factor and *echo-planar* sequences (see Section 22.5.2).

Equation (22.29) has the form of a *Fourier transformation*. To make this more obvious, Mansfield [10] introduced the concept of a reciprocal space vector $\boldsymbol{k}$ given by

$$\boldsymbol{k} = (2\pi)^{-1}\gamma\boldsymbol{G}t \tag{22.30}$$

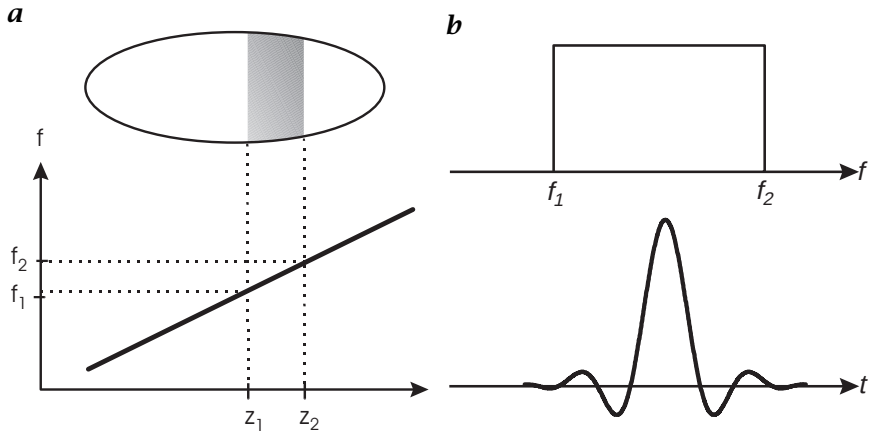


Figure 22.7: *Slice selection by RF pulses and gradient fields.* ***a*** *In the presence of a z-gradient field, the resonance frequency varies along the z-axis. The slice defined by $z_1 \leq z \leq z_2$ corresponds to Larmor frequencies $f_1 \leq f \leq f_2$.* ***b*** *To excite spins within this slice, an RF pulse with a constant amplitude at these frequencies is required (upper part). Shape of the RF pulse in the time domain with this excitation profile (lower part).*

It is clear that k-space may be traversed by changing time or changing the gradient magnitude. The direction of this traverse is determined by the direction of the gradient $\boldsymbol{G}$. In the formalism of k-space, signal and *spin density* may be written as

$$S(\boldsymbol{k}) = \int\int\int \rho(\boldsymbol{r}) \exp(2\pi \mathrm{i} \boldsymbol{k}\boldsymbol{r})\, \mathrm{d}\boldsymbol{r} = \mathcal{F}\left(\rho(\boldsymbol{r})\right) \tag{22.31}$$

and its inverse,

$$\rho(\boldsymbol{r}) = \int\int\int S(\boldsymbol{k}) \exp(-2\pi \mathrm{i} \boldsymbol{k}\boldsymbol{r}]\, \mathrm{d}\boldsymbol{k} = \mathcal{F}^{-1}\left(S(\boldsymbol{k})\right) \tag{22.32}$$

Equation (22.31) and Eq. (22.32) state that the signal, $S(\boldsymbol{k})$, and the spin density, $\rho(\boldsymbol{r})$, are connected by the Fourier transform. This is the fundamental relationship in MRI.

Let us think about the physical meaning of the k-values. Consider a gradient in x-direction with no components along the y-and z-direction. This gradient is applied during the time interval τ_{acqu} while the data are measured by the *analog-digital converter*. The exponential in Eq. (22.32) is periodic along the x-direction, because $\exp(-2\pi \mathrm{i} k_x x)$ repeats itself when $k_x x$ increases by 1. This means that the k_x value describes a wavelength: $\lambda = 1/k_x$. As the readout gradient is usually preceded by a negative gradient of duration $\tau_{acqu}/2$ in order to generate a gradient-echo (see Section 22.3.2), $\boldsymbol{k}$-space is usually measured symmetric about

the origin, that is, in the interval $[-k_{x,max}, k_{x,max}]$, with k_x given by

$$k_{x,max} = \frac{\gamma G_x \tau_{acqu}}{2} = \frac{\gamma G_x \tau_s N_x}{2} \tag{22.33}$$

and where N_x is the number of samples per readout interval τ_{acqu}, that is, the time during which data are measured and converted, and $\tau_s = \tau_{acqu}/N_x$. The largest wavelength occurring in the x-direction is

$$\lambda_{max} = \frac{2\pi}{k_{x,min}} = \frac{2\pi}{\gamma G_x \tau_s} = \text{FOV} \tag{22.34}$$

which gives the FOV of the image. Therefore, the smallest wavelength, that is, the resolution, is

$$\lambda_{min} = \frac{2\text{FOV}}{N_x} \tag{22.35}$$

It is interesting to note here that the spatial resolution defined by Eq. (22.35) can be much smaller than the wavelength of the electromagnetic radiation transmitted and received by the MR scanner in order to observe the nuclear magnetic resonance effect. While the latter is on the order of a few meters, the *spatial resolution* in MRI can, in principle, be as low as 10 μm. The higher spatial resolution is a consequence of using the resonance effect rather than diffraction limited imaging for formation of the MR image.

Frequency encoding. Consider an x-gradient. The MR signal in the presence of this gradient is

$$S(k_x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\rho(x,y,z)\exp[2\pi i k_x x]dx\right]\mathrm{d}y\,\mathrm{d}z \tag{22.36}$$

and the inverse Fourier transform is

$$\mathcal{F}^{-1}[S(k_x)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\rho(x,y,z)\,\mathrm{d}x\,\mathrm{d}y \tag{22.37}$$

This equation states that the inverse Fourier transformation of the MR signal detected in the *presence* of a gradient field may be regarded as the projection of the spin density onto the axis defined by the gradient direction (Fig. 22.8). As the signal $S(k_x)$ is sampled in time-domain at N_s discrete time points $t_i (i = 1, 2, \ldots N_s)$ with $k_x(t_i) = (2\pi)^{-1}\gamma G_x t_i$, the Fourier transform of the signal is a frequency, and the position x of a spin is coded by its Larmor frequency (Fig. 22.6b). Thus, the acquisition of the MR signal in the presence of a gradient field is termed *frequency encoding*.

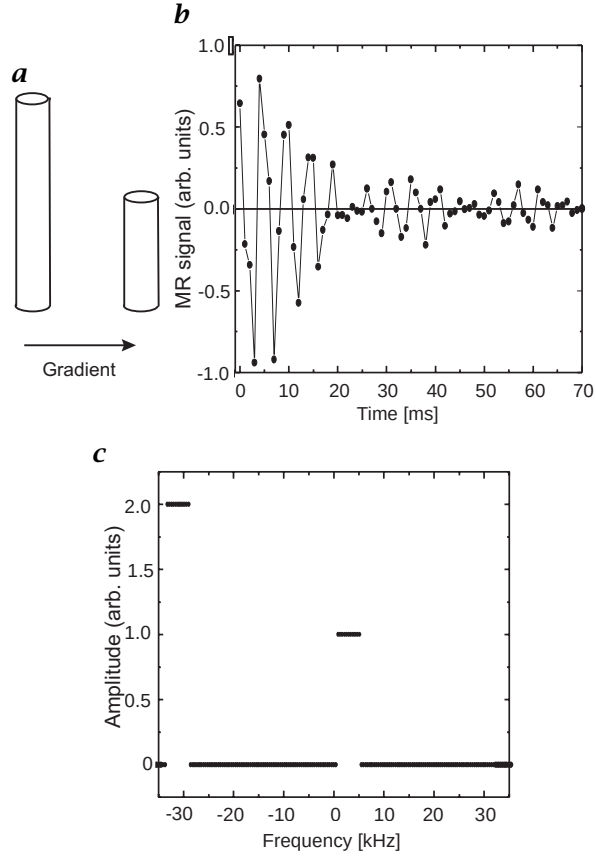


Figure 22.8: *Fourier relationship between* ***a*** *object;* ***b*** *time domain MR signal in the presence of a gradient field; and* ***c*** *reconstructed frequency spectrum.*

The direction of the gradient used for frequency encoding is usually referred to as the frequency encoding gradient. Because the MR signal is measured during the time when this gradient is switched on, the frequency encoding gradient is also called the *readout gradient*. The frequency encoding procedure is often referred to as 1-D imaging and leads to the projected distribution of spins rather than the exact spin density $\rho(x, y, z)$. In order to measure a 2-D or 3-D distribution of $\rho(x, y, z)$, spatial encoding must be performed with additional gradients in other directions. If these additional gradients are applied during the acquisition of the MR signal, a series of projections along varying orientations will be the result from which images can be reconstructed.

Phase encoding. In order to make use of fast Fourier transform algorithms, another approach was suggested by Kumar et al. [9], which relies on the application of orthogonal gradient fields.

Consider a y-gradient of magnitude G_y switched on with a duration τ_y before data acquisition. During that time the protons precess with a Larmor frequency $\omega_y = \gamma B_0 + \gamma G_y y$.

After this gradient is switched off, the protons have accumulated an additional phase, which depends on their position in y-direction (Fig. 22.9):

$$\Phi(y) = \gamma G_y y \tau_y \tag{22.38}$$

Now consider a series of N_y measurements, where a gradient in y-direction is switched on during τ_y with varying magnitudes

$$G_y^i = \left(i - \frac{N_y}{2}\right) G_y \qquad (i = 1, 2, \ldots, N_y) \tag{22.39}$$

According to Eq. (22.30), different positions in k-space are selected by the gradient:

$$k_y^i = (2\pi)^{-1} \left(i - \frac{N_y}{2}\right) \gamma G_y \tau_y \tag{22.40}$$

The MR signal acquired after each of these gradients is thus given by

$$S(k_y^i) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \rho(x, y, z) \exp(2\pi \mathrm{i} k_y^i y) \, \mathrm{d}y \right] \mathrm{d}x \, \mathrm{d}z \tag{22.41}$$

Inverse Fourier transformation of the series of signals measured with different y-gradients results in

$$\mathcal{F}^{-1}\left[S(k_y^i)\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \rho(x, y, z) \, \mathrm{d}x \, \mathrm{d}z \tag{22.42}$$

Again the inverse Fourier transform of the MR signal acquired after a series of gradients gives the projection of the spin density $\rho(x, y, z)$ onto the y-axis.

Switching on and off a gradient field for some time interval τ_y before frequency encoding and measuring the signal is termed *phase encoding*. Although this phrase may be misleading because all gradients of field inhomogeneities add phase to the magnetization, it is universally used.

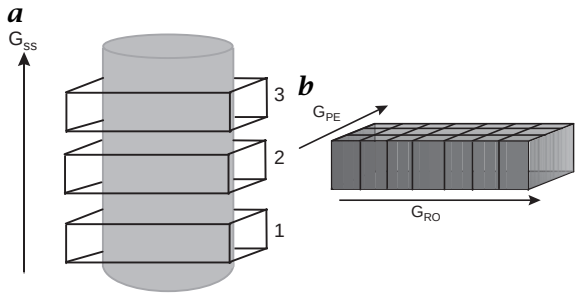


Figure 22.9: *Two-dimensional MRI.* ***a*** *In a first-step magnetization within a single slice is excited by an RF pulse in combination with a slice-selection gradient G_{ss}.* ***b*** *For that slice 2-D spatial encoding is performed with a phase-encoding gradient G_{PE} and a frequency-encoding gradient G_{RO}. If multiple slices are to be measured, this procedure is repeated for each slice.*

22.3.2 Two-dimensional magnetic resonance imaging

We will now consider how the previously described methods for spatial selection and encoding can be combined to form an image from a 3-D object.

The most commonly applied strategy to do so is first to excite a slice in the object and then to perform a *2-D spatial encoding* in that slice of excited magnetization. A 3-D volume can then be examined by repeating this process for the other slices (Fig. 22.9).

Figure 22.10 shows a pulse sequence diagram of a 2-D MRI sequence, that is, the computer program for the MR scanner defining the temporal arrangement of the various RF fields, gradient fields, and the timing of the data acquisition. The graphical visualization of a pulse sequence is termed a *pulse sequence* diagram.

Transverse magnetization is excited in a slice by a combination of a *sinc-shaped RF pulse* and a constant gradient in slice direction G_{ss}. The slice-encoding gradient induces dephasing of the magnetization in the slice direction across the finite slice. This dephasing begins to occur during the RF pulse. It may be compensated for by an additional gradient in slice-direction, where the integral of that refocusing gradient equals half the integral under the slice encoding gradient (Fig. 22.10). For convenience, the phase of the magnetization after refocusing in the slice-direction is assumed to be zero.

After the *slice selection gradient* (and overlapping with the rephasing slice selection gradient) the phase encoding gradient is applied. The pulse sequence will be repeated with different values of the phase encoding gradient. The MR signal is sampled with the gradient in frequency encoding direction (G_{RO}) switched on. Data acquisition, that

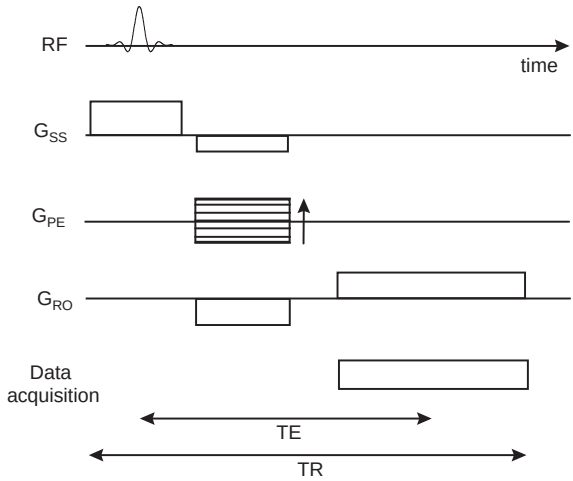


Figure 22.10: *Two-dimensional pulse sequence. RF: RF pulse. G_{ss}: slice selection gradient. G_{PE}: phase encoding gradient. G_{RO}: frequency encoding gradient. TE: echo-time (time interval between mid-RF and mid of data-acquisition). TR: sequence repetition time (time between successive repetitions of the pulse sequence with different values of the phase encoding gradient). The arrow next to the phase encoding gradient indicates that this gradient is incremented after each TR interval.*

is, measurement and digitization of the MR signal (N_{RO} data points), is performed during the frequency encoding gradient.

Although the *readout gradient* implies a position dependent phase, which is a spatial label, it also leads to a reduced MR signal (in analogy to the discussion on slice selection gradient). Thus, a gradient with opposite sign is switched on in the frequency encoding direction (in parallel to the phase encoding gradient). For best image quality the area under this gradient is usually chosen to be half the area under the readout gradient. In this case the maximum MR signal occurs during mid-data acquisition. The concept of having a signal increase after some time of gradient activity is called echo generation. As echo generation in the pulse sequence of Fig. 22.11 is performed by gradient switching, it is referred to as a *gradient-echo* pulse sequence. In another type of pulse sequence, the gradient-echo may be combined with a spin echo (Section 22.3.4). The time interval between mid-RF and mid of data acquisition is the *echo time* (TE). The length of this interval determines the influence of the amount of T_2 and T_2^* decay on the MR signal.

After data acquisition, the pulse sequence in Fig. 22.10 is repeated with a different value of the phase encoding gradient. Each acquisition of an MR signal with a single value of the phase encoding gradient is called a phase encoding step. The time between successive repetitions

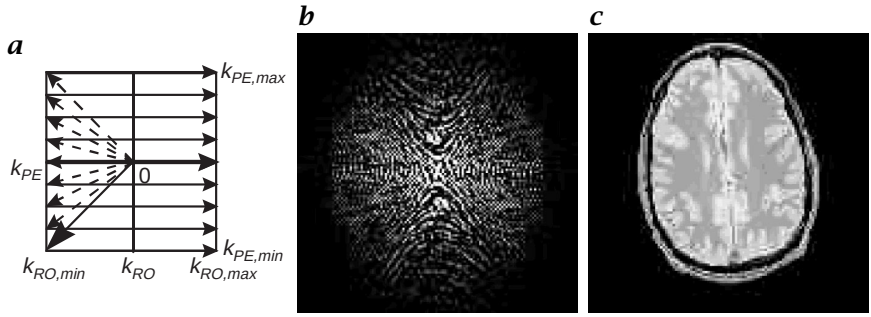


Figure 22.11: *Gradient-echo sequence.* ***a*** *Traversal of* ***k****-space. The signal is initialized at* $k = 0$ *by the RF pulse. Because of the negative first-phase encoding gradient value Eq. (22.40), the first line of* ***k****-space is measured with the minimum* $\mathbf{k}_{PE}$ *value (long bold arrow). For this value a line in* ***k****-space is traversed in frequency encoding direction from* $\mathbf{k}_{RO,min}$ *to* $\mathbf{k}_{RO,max}$ *(bold arrows). The whole measurement is repeated with different values of* $\mathbf{k}_{PE}$ *(shadowed arrows).* ***b*** *MR signal from measurement (real part of the complex MR signal).* ***c*** *The reconstructed magnitude MR image shows a human brain.*

is called the sequence *repetition time* (TR). Both TE and TR are important factors influencing the contrast in the MR image (see Section 22.4).

Following Eq. (22.31), the MR signal is

$$S(k_x, k_y) = \int_{-d/2}^{d/2} \left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \rho(x, y, z) \exp\{2\pi i(k_x x + k_y y)\}\, dx dy \right] dz \tag{22.43}$$

where d denotes the *slice thickness*. It is clear that the MR signal is the 2-D Fourier transform of the spatial distribution of spin density within the excited slice. Reconstruction of $\rho(x, y^2)$ simply requires to calculate the inverse Fourier transform

$$\rho(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} S(k_x, k_y) \exp\{-2\pi i(k_x x + k_y y)\}\, dk_x dk_y \tag{22.44}$$

where the integrated z-dependence representing averaging across the slice has been suppressed. Generally, the number of phase-encoding values and the number of data points measured during the frequency encoding gradient are a power of two and, therefore, image reconstruction can be performed on the basis of fast *Fourier transform* (FFT) algorithms. For medical applications, it is the magnitude of the complex numbers resulting from the inverse Fourier transformation that is of relevance and, therefore, phase information is usually dismissed in MRI. See Fig. 22.11 for an example.

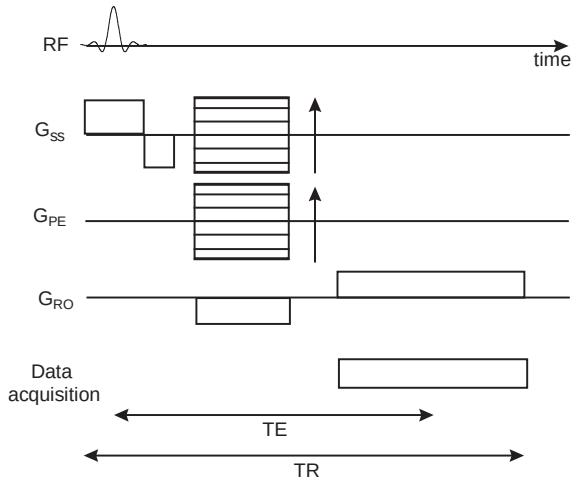


Figure 22.12: *Three-dimensional pulse sequence. Additional spatial encoding in the slice-direction is performed by a second-phase encoding gradient with N_{SS} various values. This pulse sequence is repeated $N_{SS} \times N_{PE}$ times to perform full 3-D spatial encoding.*

We have defined the slice, read- and phase-encoding direction to be in z-, x- and y-direction, respectively. However, this was an arbitrary selection. The only requirement is that the slice-, frequency-, and phase-encoding gradients are perpendicular but they do not have to be aligned with the axes of a Cartesian coordinate system. By superposition of gradient fields of the physical gradient coils of the scanner, any direction in space may be defined as the slice direction. This makes MRI extremely flexible in that the orientation of any acquired MR image may be freely selected with no restrictions. In *computed x-ray tomography* (CT) on the other hand, a slice is defined by the plane in which the x-ray tube and the detector rotate around the patient. Although the patient table may be angulated to a certain degree, only small deviations of the transverse slice orientation may be obtained in CT.

22.3.3 Three-dimensional magnetic resonance imaging

Often information is required over a large 3-D volume-of-interest. Two different approaches may be used for doing so: (i) scanning several slices across the volume with a 2-D pulse sequence as described in the previous section, or (ii) true *3-D encoded MRI*.

The 3-D imaging sequence is created from a 2-D sequence by adding a second-phase encoding gradient to phase encode the volume along the slice select direction (Fig. 22.12). Following Eq. (22.31), the MR sig-

nal $S(k_x, k_y, k_z)$ is

$$S = \int_{-d/2}^{d/2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \rho(x, y, z) \exp\{2\pi i(k_x x + k_y y + k_z z)\} \, dx \, dy \, dz \quad (22.45)$$

where d denotes the thickness of the excited slice. The MR signal is now the 3-D Fourier transform of the spatial distribution of spin density within the excited volume. *Reconstruction* of $\rho(x, y, z)$ simply requires 3-D inverse Fourier transformation

$$\rho = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} S(k_x, k_y) \exp\{-2\pi i(k_x x + k_y y + k_z z)\} \, dk_y \, dk_y \, dk_z \quad (22.46)$$

The larger number of RF excitations during a 3-D MRI scan means that the SNR is increased by the square root of the number of slices, when compared with a 2-D sequence. Therefore, thinner slices and/or smaller pixels can be measured with an acceptable SNR in 3-D imaging than in 2-D imaging. The slice thickness in 2-D MRI is in the range of 4 to 10 mm, while in 3-D MRI it may be as low as 0.5 mm. Moreover, because the excited volume may be rather thick, the duration of the RF pulse is relatively short and, thus, TE typically is shorter than in thin-slice 2-D imaging.

On the other hand, 3-D imaging requires a total measurement time of $N_{SS} \times N_{PE} \times TR$. To keep this time in an acceptable range, TR must be much shorter than in 2-D imaging (4 to 12 ms vs 400 to 4000 ms). Because of this short TR 3-D images can easily be made $T1$-weighted (see Section 22.4), but it is rather difficult to achieve $T2$-weighting. With a modern MR scanner, 3-D MR images of the human head can be measured within 6:56 min with voxeldimensions of $0.5 \times 1.0 \times 1.0\,\text{mm}^3$ ($N_{RO} = 512$, $N_{PE} = 256$, $N_{SS} = 256$, $TR = 6$ ms). In order to emphasize the 3-D nature, a pixel in MR images is commonly referred to as a voxel, that is, a volume element (see also Volume 2, Section 2.3.2).

22.3.4 Spin-echo sequence

Figure 22.13 shows the pulse sequence diagram of a 2-D *spin-echo* pulse sequence. Slice selection and phase encoding are performed in the same way as in the gradient-echo pulse sequence of Fig. 22.10. However, after the rephasing gradient in slice direction and after the phase encoding gradient, a 180°-pulse is applied in order to generate a spin echo in the middle of the data acquisition period. The readout dephasing gradient is applied with the same sign as the readout gradient because of the inversion of the magnetization by the 180°-pulse. The area

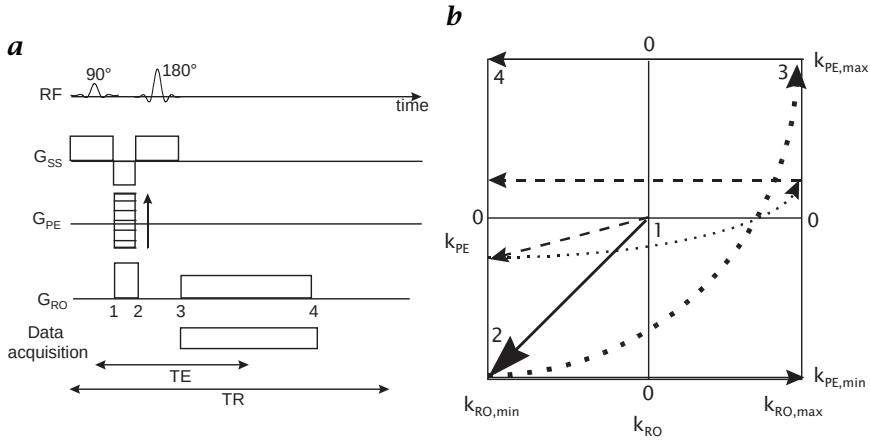


Figure 22.13: ***a*** *Spin-echo pulse sequence. A spin echo is generated at the echo-time TE by the slice-selective 90°–180° RF pulse combination. The refocusing pulse in slice direction is applied before the 180°-pulse, but it may also be applied after the 180°-pulse. Frequency encoding with generation of a gradient-echo is performed in the readout direction.* ***b*** ***k****-space diagram. Only two phase encoding steps are drawn. The 180°-pulse inverts the sign of the phase, so the sign of both the* $\mathbf{k}_{RO}$ *and the* $\mathbf{k}_{PE}$ *is mirrored through the origin (2 ⇒ 3). Then the readout gradient is switched on and a line with constant* $\mathbf{k}_{PE}$ *is described in the* ***k****-plane (3 ⇒ 4).*

under the dephasing gradient is calculated to generate the gradient-echo at the same time as the spin echo. It should be pointed out here that, although pulse sequences are usually grouped into either gradient-echo or spin-echo pulse sequences, every spin echo is combined with the formation of a *gradient echo* at the same time to achieve frequency encoding.

The spin-echo sequence is the standard pulse sequence in most imaging protocols in medicine for several reasons: The 90° pulse creates the maximum transverse magnetization from the available longitudinal magnetization and, thus, the SNR is at maximum for long TR values. In general, the TR in spin-echo sequences is chosen to be at least on the order of $T1$. Therefore, because of $T1$-relaxation, at least 63 % of the equilibrium magnetization has built up before the next 90° pulse. Furthermore, due to the formation of a spin echo, static field inhomogeneities are largely eliminated, resulting in pure $T2$ rather than in $T2^*$ as the dominant factor determining the available transverse magnetization at the time of acquisition of the MR signal. Because $T2 \gg T2^*$, the signal is often higher than in gradient-echo images with identical echo-time. The long TR can be used effectively for the acquisition of several images with different echo times (Fig. 22.14), from which $T2$ can be calculated, or for imaging of multiple slices without increasing

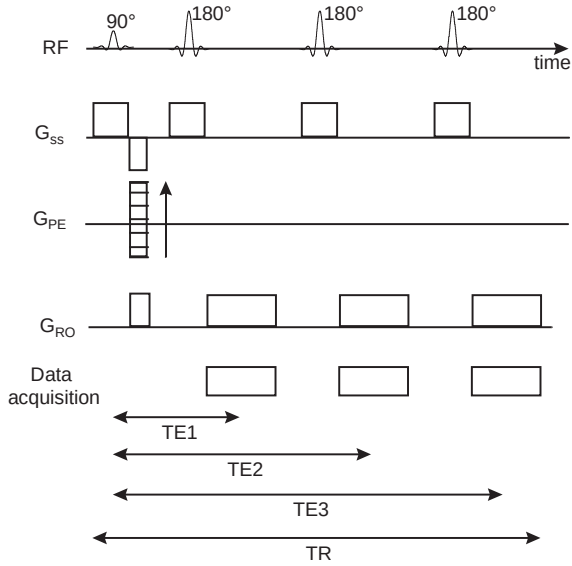


Figure 22.14: *Multiecho spin-echo pulse sequence. After transverse magnetization is excited by the initial 90° pulse a series of 180° pulses generate multiple spin echoes with variable TE. See Fig. 22.18 for images acquired with a double echo pulse sequence.*

the measurement time ("*multislice imaging*," Fig. 22.15). In medical applications the assessment of a complete volume rather than of a single slice is required, so the latter technique is the standard mode of operation of spin-echo (and of long-TR gradient-echo) imaging. In contrast to true 3-D imaging (Fig. 22.12), the measurement time does not depend on the number of individual slices acquired. However, due to the limited length of the RF pulses, the minimum slice thickness achievable with multislice imaging is on the order of 2 mm, while in practice the minimum slice thickness in 3-D imaging is given by the maximum measurement time a patient or physician is willing to accept.

22.4 Image contrast

The MR signal depends on the measurement parameters and on a variety of tissue parameters such as spin density, $T1$, $T2$, but also on macroscopic or microscopic (blood) flow, magnetic susceptibility, water self diffusion, etc. For a review of tissue relaxation times, see [11]. In order to discuss the three basic forms of *image contrast*, *spin density*, $T1$ and $T2$ (or $T2^*$), a spin-echo experiment is used as an example of how to obtain different forms of contrast. When $T2$ is replaced by

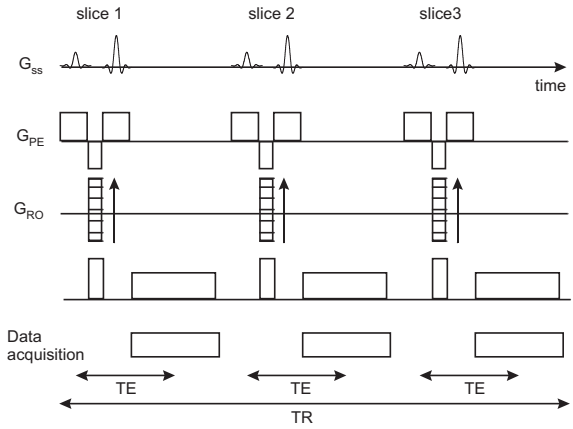


Figure 22.15: *Multislice spin-echo pulse sequence. After the first spin-echo part, the RF and gradient scheme is repeated for other slice positions. By this acquisition scheme, up to 20 slices can be measured simultaneously at a typical TR of 600 ms without increasing the measurement time.*

$T2^*$ these results are identical to those that would be found for the 90° gradient-echo experiment.

22.4.1 $T2$-, $T2^*$- and spin-density contrast

The MR signal is proportional to the magnitude M_T of the transverse magnetization vector $\boldsymbol{M}_T$ during the data acquisition interval, the latter being a function of the flip angle α, of the longitudinal magnetization before the irradiation of the RF pulse, and of the echo-time TE (Fig. 22.16)

$$M_T(t) = M_z^- \sin\alpha \exp\left(-\frac{TE}{T2}\right) \tag{22.47}$$

where M_z^- denotes the magnitude of the magnetization vector just before the RF pulse, which shall be assumed to be parallel to the z-axis. If the RF pulse irradiates a spin system in thermal equilibrium (i. e., $TR \gg T1$), then $M_z^- = M_0$ and the signal intensity depends on the spin density and on $T2$. If $TE \ll T2$, then the exponential will be close to unity and the signal will be proportional to the equilibrium magnetization. As the temperature in the human body is constant, the signal is then proportional to the spin density. If, on the other hand, $TE \approx T2$, then the signal will be affected by both the spin density and $T2$ of the tissue.

As it is not only the spin density that affects the MR signal in a pulse sequence with $TR \gg T1$ and $TE \ll T2$, images acquired with

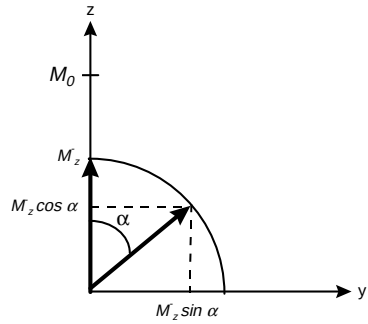


Figure 22.16: *Magnetization after an RF pulse of flip angle α. The remaining longitudinal magnetization is proportional to the cosine, the excited transverse to the sine of the magnetization before the pulse.*

these acquisition parameters are referred to as *spin-density weighted* images. In analogy, images with $TE \approx T2$ are referred to as $T2$-*weighted* images. If several images were acquired with a variety of echo-times, then a map of $T2$ or of the spin density may be calculated on a voxel-by-voxel basis by performing a fit of Eq. (22.47) to the series of images. Although this technique is commonly used in research, in diagnostic imaging the concept of using weighted images is more efficient in that the required measurement time is much shorter, and often the image contrast between normal and pathologic tissue is better in weighted images than in images calculated by signal fitting. Sometimes it may even be the combination of different parameters influencing the image contrast in weighted images that improves the contrast between normal and pathologic tissue.

Figure 22.17 shows examples of spin density and $T2$-weighted images from the human brain. In spin-density weighted images, tissue with high spin density is bright, in $T2$-weighted images, tissue and fluids with long $T2$ give a high signal.

22.4.2 $T1$-contrast

Prerequisite for acquisition of $T2$- (or $T2^*$-) and spin-density weighted images is the acquisition with $TR \gg T1$ in order to achieve thermal equilibrium before the next RF excitation. However, if $TR \approx T1$, $T1$-relaxation can not be neglected in the discussion of image contrast, because M_z^- now also depends on the time since the last RF pulse (i. e., TR) and on the $T1$ of tissue. Taking into consideration the Bloch equation for the longitudinal magnetization (Eq. (22.23)), one finds that the longitudinal magnetization just prior to the $i+1$-st irradiation of an RF

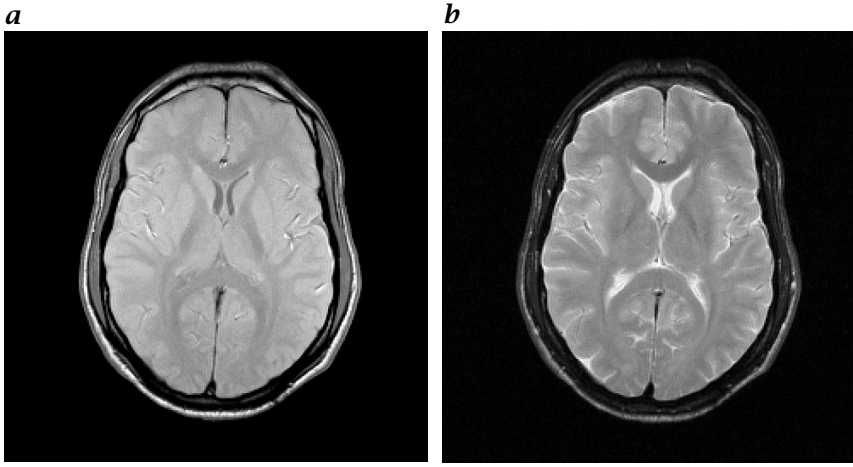


Figure 22.17: *Images from a 6-mm slice through the brain of a healthy subject obtained with a double-echo spin-echo pulse sequence (TR = 3800 ms).* ***a*** *Spin-density weighted image obtained from the first echo (TE = 22 ms). As spin density differences in biological tissue are usually low, the contrast of spin-density weighted images is often also low.* ***b*** *$T2$-weighted image obtained from the second echo (TE = 88 ms) at the same slice position. Tissue and fluids with long TE present with very high signal intensity.*

pulse with flip angle α is

$$M_z^{i+1} = M_0 + (M_z^i \cos\alpha - M_0) \exp\left(-\frac{TR}{T1}\right) \tag{22.48}$$

where M_z^i denotes the magnetization just before the i-th irradiation of the RF pulse. Combining Eq. (22.48) with Eq. (22.47) for $\alpha = 90°$ results in the signal equation of the spin-echo sequence

$$M_T = M_0 \left[1 - \exp\left(-\frac{TR}{T1}\right)\right] \exp\left(-\frac{TE}{T2}\right) \tag{22.49}$$

Thus, in order to measure an image with $T1$ as the main source of image contrast, the conditions $TR \approx T1$ and $TE \ll T2$ must be fulfilled. These images are termed $T1$-weighted images. In $T1$-*weighted* images the magnetization of tissue with short $T1$ quickly relaxes back to its equilibrium value and appears bright.

The signal equation for the spin-echo sequence Eq. (22.49) may also be used for a gradient-echo pulse sequence with $\alpha = 90°$ -and when $T2$ is replaced by $T2^*$. For comparison of a spin-echo with a gradient-echo image see Fig. 22.18.

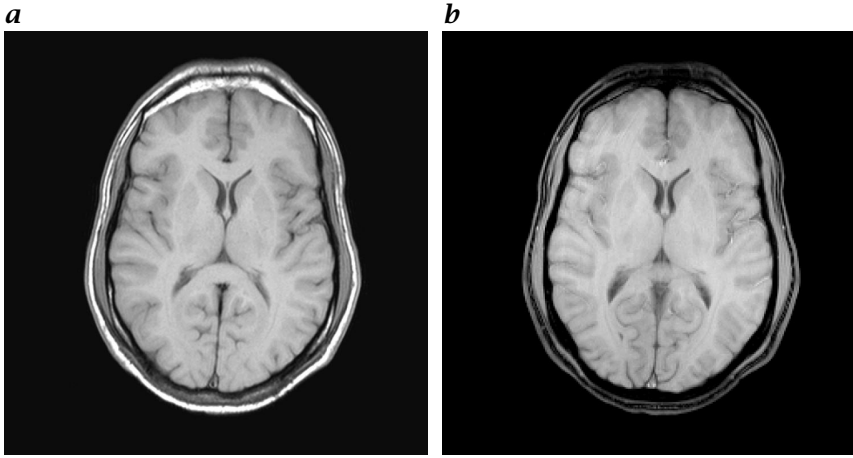


Figure 22.18: *T1-weighted images (TR = 580 ms) from the same subject as in Fig. 22.17. The measurement time was 4:55 min for each sequence.* ***a*** *Spin-echo image (TE = 15 ms).* ***b*** *Gradient-echo image (TE = 7 ms, $\alpha = 90°$). Overall image quality is better in the spin-echo image. On both images fluids have low image signal due to the long T1 of several seconds. Normal brain tissue T1 is between 600 and 1000 ms at $B_0 = 1.5$ T. Note the better contrast in the spin-echo image.*

22.5 Fast imaging methods

The *acquisition time* of *T*1- and *T*2-weighted spin-echo images is on the order of several minutes. If high spatial resolution is to be obtained the number of phase encoding steps and, thus, the measurement time will quickly come in an unacceptable range of 10 min and more. In addition, in the presence of slow motion (e. g., breathing of the patient) it may be of advantage to measure while the patient can hold his breath. For such applications the image acquisition time should be less than about 30 s.

These demands require strategies to reduce the scan time (this will be presented in this section). One strategy is to reduce the *TR* and, in order to keep enough longitudinal magnetization before the next RF excitation, the flip angle simultaneously, a technique called fast low-angle shot (FLASH) imaging [12]. Another strategy is to reduce the number of RF excitations required for an image by generating multiple differently phase-encoded gradient or spin echoes after a single RF pulse. These techniques are referred to as *turbo spin echo* or, in the case of one or only a few RF excitations per image, *echo-planar pulse sequences*. These approaches will be described in the following sections.

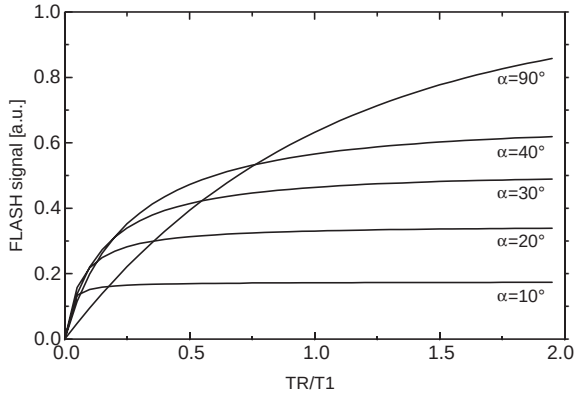


Figure 22.19: *MR-signal of FLASH acquisition as a function of $TR/T1$. Note that at short TR the FLASH signal is higher than that of a pulse sequence with $\alpha = 90°$.*

22.5.1 FLASH imaging

Gradient-echo imaging presents a method to reduce scan times significantly by simultaneously reducing the TR and the flip angle α of the RF-pulse. Assuming that no transverse magnetization exists before the next RF excitation, the general signal equation for a *gradient-echo* pulse sequence with flip angle α results from Eq. (22.48) and Eq. (22.47) for $i \to \infty$:

$$M_T = M_0 \frac{(1 - \exp(-TR/T1)) \sin\alpha}{1 - \cos\alpha \exp(-TR/T1)} \exp\left(-\frac{TE}{T2*}\right) \tag{22.50}$$

The MR signal in dependence of TR and flip angle is shown in Fig. 22.19. For short repetition times higher MR signal is obtained by measuring with a flip angle of less than 90°. Thus, when the flip angle is reduced, TR may be shorter than $T1$. The optimum flip angle with regard to the MR signal is denoted as the Ernst angle

$$\alpha_{opt} = \arccos\left[\exp\left(-\frac{TR}{T1}\right)\right] \tag{22.51}$$

As a consequence of the reduced TR, image acquisition times may be shortened significantly from several minutes in spin-echo imaging to a few seconds if repetition times of 10 - 20 ms are used. This type of gradient-echo imaging is termed *FLASH imaging* ("Fast Low Angle Shot," [12]).

Figure 22.20 shows FLASH images in a patient before and after administration of a paramagnetic *contrast agent*. Paramagnetic contrast agents by their presence lead to a reduction of the $T1$ of tissue and,

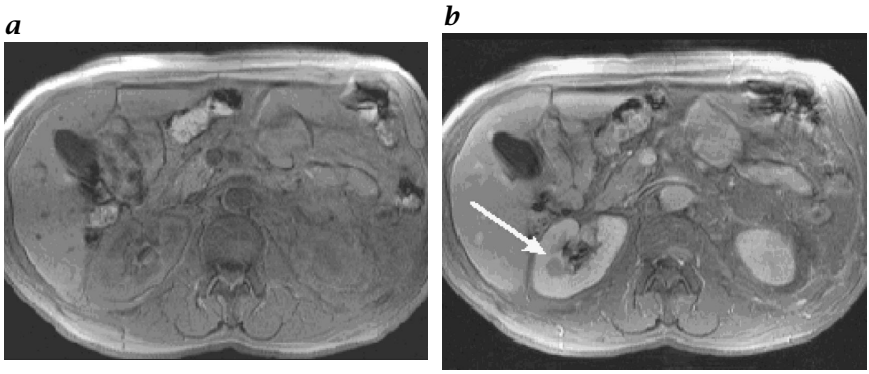


Figure 22.20: ***a*** *T1-weighted FLASH image (TR = 180 *ms*, TE = 4 *ms*, $\alpha = 80°$) from the kidney in a patient with a kidney tumor. Measurement time was 20 s, images were acquired during a single breathhold.* ***b*** *Image from the same patient after administration of a contrast agent. The contrast agent leads to a reduction of $T1$ and, hence, to an increase of the signal intensity. While the contrast agent is delivered to normal kidney tissue, the kidney tumor (arrow) does not enhance. Thus, the tumor is well seen on image b, while it can hardly be detected on image* ***a****.*

thus, an increase of the MR signal. Differences in the amount of contrast agent delivered to tissue caused by pathology result in an enhanced contrast between the pathologic and the normal tissue.

22.5.2 Sequences with advanced *k*-space trajectories

Strategies have been developed to reduce the long acquisition times in conventional spin-echo imaging by more effectively spending the excited transverse magnetization by segmented *k*-space acquisition, where a reduced number of RF excitations is followed by a sequence of multiple differently phase-encoded spin- or gradient-echoes. In its extreme, an image may be reconstructed after a single excitation, for example, in *echo-planar imaging* (EPI). Both techniques will be described in the following two sections.

Turbo spin echo. Quality of spin-echo images is often better than in gradient-echo images (Fig. 22.18). Thus, techniques have been developed to reduce the scan times of spin-echo sequences with keeping the advantage of refocusing static magnetic field inhomogeneities by 180° pulses.

The *turbo spin-echo* pulse sequence [13] is a special variant of the multi-echo spin-echo pulse sequence of Fig. 22.14. A series of signals with varying echo-time was measured in that sequence after the excitation by a 90° -pulse. All echoes were measured after the same

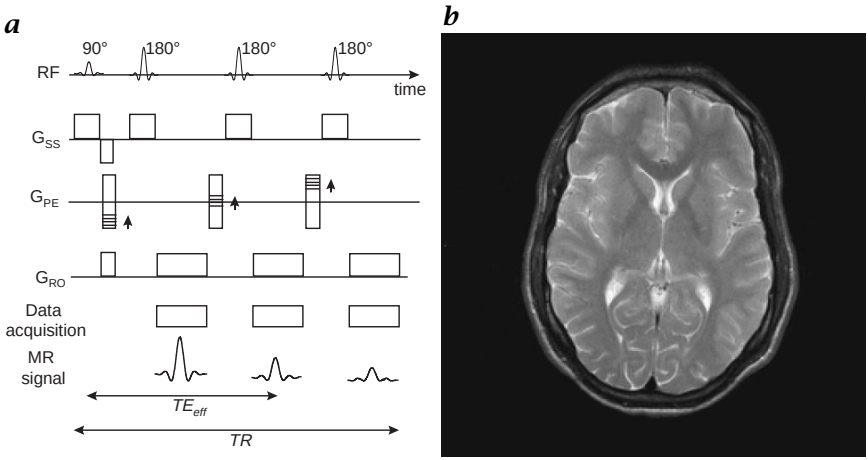


Figure 22.21: ***a*** *Turbo spin-echo pulse sequence. Multiple phase-encoding steps are performed for the echos. Because in this example, the second echo is phase encoded with* $G_{PE} \approx 0$*, the effective echo time* TE_{eff} *corresponds to the echo-time of the second echo.* ***b*** $T2$*-weighted turbo spin-echo image from the same subject as in Fig. 22.17 (*$TR = 5200$ *ms,* $TE = 128$ *ms, turbo factor 23). Although the measurement time was only 52 s, excellent image quality and* $T2$ *contrast is obtained.*

phase-encoding gradient. In the turbo spin-echo pulse sequence, multiple echoes are acquired, but these echoes are acquired with different phase-encoding gradients (Fig. 22.21a). The number of echoes acquired within one TR interval is the *turbo factor* n_{TSE}.

As the phase-encoding steps are acquired with varying TE, the echo-time of a turbo spin-echo sequence is not well defined. The low spatial frequencies in Fourier space determine the large-scale signal intensity of an image. Thus, an *effective echo time* TE_{eff} may be defined as the time from the 90° -pulse to the time of the echo where the phase-encoding gradient is zero ($k_{PE} = 0$).

The decay of the MR signal during the various phase-encoding steps causes *blurring* in the image. The degree of blurring depends on the $T2$ of tissue and on the turbo factor. In tissue where $T2$ is long, turbo factors of up to 30 may be used without significant loss of resolution.

The number of phase encodings required for the acquisition of an image is N_{PE}/n_{TSE}. Therefore, the acquisition time of a turbo spin-echo sequence is reduced by the factor n_{TSE} as compared with an otherwise identical spin-echo pulse sequence. As the measurement time for a high resolution $T2$-weighted image is reduced from about 10 min to 1 min or less, the turbo spin-echo sequence has become the preferred sequence for spin-density- and $T2$-weighted imaging today.

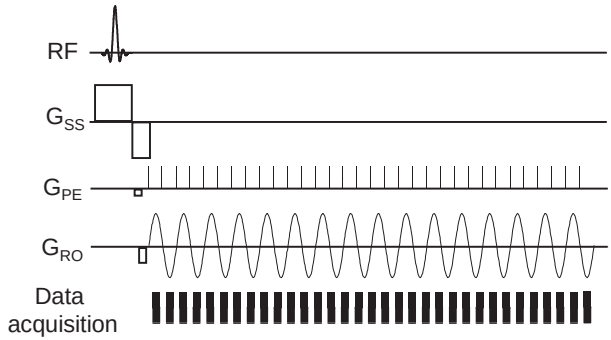


Figure 22.22: *Echo-planar pulse sequence. After the RF pulse, all lines in* ***k****-space are acquired by refocusing multiple gradient-echoes with a sinusoidal gradient. The phase encoding gradient is replaced by short "blips". For a* ***k****-space diagram see Fig. 22.23a.*

Echo-planar imaging. The idea of measuring multiple echoes after a single excitation can also be applied to gradient-echo imaging. In analogy to the turbo spin-echo sequence a turbo gradient-echo sequence can be generated by simply refocusing multiple gradient-echoes after one RF excitation pulse. As the time for refocusing of the gradient-echoes is much shorter than the 180°-pulses in the turbo spin-echo sequence, all phase encoding steps may be acquired after a single excitation (Fig. 22.22). As a consequence, acquisition times are even shorter (100 ms and less). This mode of data acquisition is commonly referred to as echo-planar imaging. Because all lines of ***k***-space are acquired after a single RF excitation, it is also termed a "*single-shot*" technique and is usually referred to as *echo-planar imaging* [10].

Although the concept is simple, technical demands, in particular regarding the gradient system, are severe. In order to reduce power requirements, the gradient coils may be part of a resonant circuit. Therefore, gradient-echoes may be generated under a sinusoidal gradient, although conventional trapezoidal gradients can also be used if somewhat longer acquisition times are accepted. Because the phase encoding must be performed between two gradient-echoes, and as it would require applying a normal frequency encoding gradient, it is replaced by a series of short gradient-pulses, each advancing k_{PE} by the required amount.

The MR scanners capable of routinely performing echo-planar imaging have been commercially available only since 1996. Because of the limited image quality (Fig. 22.23b) EPI is used extensively for research but not as much for diagnostic purposes.

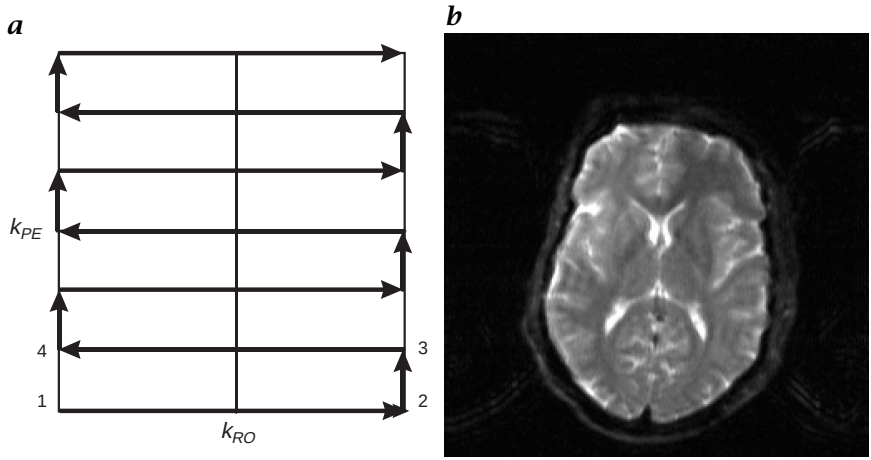


Figure 22.23: *Single-shot echo-planar imaging.* ***a*** ***k****-space is traversed after a single RF pulse. Short phase encoding "blips" perform the step* $2 \rightarrow 3$. ***b*** *EPI image of the brain, acquired in approximately 120 ms. As the TR is virtually infinite, the image is strongly T2-weighted (TE = 70 ms). Due to the single RF excitation and T2 decay during the acquisition of the multiple gradient-echoes, the image is somewhat blurry in comparison to the T2-weighted turbo spin-echo image in Fig. 22.21.*

22.6 Overview of quantitative applications

In the last sections, an introduction to the main concepts of modern medical MRI has been presented. All described methods have found an immense variety of applications in the assessment of tissue morphology. Common to all of these applications, however, is that they are based mainly on qualitative differences of physical and physiologic parameters, resulting in the concept of "weighted" images rather than representing quantitative images of a single physical parameter alone. It is the aim of this section to point out a few more quantitative applications of medical MRI, which may be of interest to the reader of this handbook.

With the hardware used in medical MRI, an MR signal can be obtained only from nuclei of fluids and of substances dissolved in fluids. Gases cannot be measured because of the low spin density (compare, however, the remark made in Footnote 1 of this chapter), and solids because of their extremely short T2. Whereas in the human body the proton concentration is much higher (110 M) than that of any other nucleus, and because its gyromagnetic ratio γ is very high, medical applications of MRI are performed almost exclusively at the resonance frequency of the proton, although any nucleus with spin $\boldsymbol{I} = 1/2$ may in principle

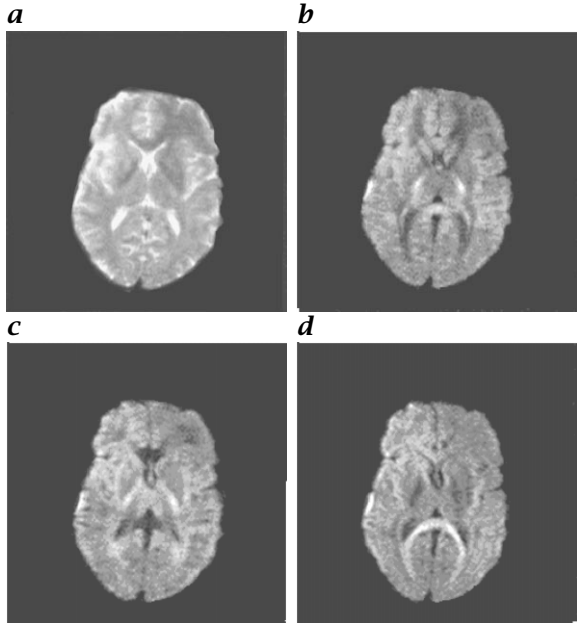


Figure 22.24: *Diffusion-weighted images of the brain in a healthy subject.* ***a*** *Reference image without diffusion gradient, and images with diffusion sensitizing gradients in* ***b*** *vertical,* ***c*** *horizontal, and* ***d*** *in the through-plane direction. In the diffusion-weighted images a high signal intensity represents a low diffusion coefficient. Note the intensity variation of brain structures depending on the orientation of the diffusion gradients. It represents different orientations of nerve fibers. All images were measured with otherwise identical imaging parameters and at identical slice positions. Images* ***b, c, d*** *are displayed with the same range of gray values; for image* ***a*** *different scaling was used due to its higher overall signal intensity.*

be used. Nuclei used so far are—among others—^{31}P, ^{19}F, ^{13}C and ^{23}Na [14].

Paramagnetic substances lead to a reduction of $T1$ of the fluid and, if they diffuse across the blood vessel wall, of the tissue. Because *tumors* often have abnormal blood supply, vessel permeability, and distribution volume for the paramagnetic substance, this inhomogeneous distribution of paramagnetic contrast agent is the basis for contrast between diseased and normal tissue in $T1$-weighted images (see Fig. 22.20). Moreover, measurements have shown that the relaxation rate $R1$ of tissue ($R1 = 1/T1$) and the concentration C of a paramagnetic substance are related by $R1 = R10 + \alpha C$, where $R1_0$ denotes the tissue relaxation rate without the paramagnetic substance and α a constant depending on the substance. This relation may also be used to calculate the concentration of a paramagnetic substance from the MR

signal, and—using an underlying mathematical model—may be used to calculate the *fractional distribution volume* of the substance within a voxel or the *permeability* of a vessel for that substance [15]. This principle has also been used in environmental physics to visualize and measure the 3-D distribution of *flow* both in soil and in *porous media* [16].

The MRI also has the potential to measure macroscopic and microscopic motion. Quantification of *macroscopic flow* will be addressed in greater detail in Section 23.2.2. *Self-diffusion*, that is, microscopic thermal motion, can be measured by introducing a pair of gradients between the RF pulse and the data acquisition interval. The two gradients of this pair are identical but have an opposite sign. Thus, a spin ensemble in stationary tissue is dephased by the first gradient and refocused completely by the second gradient, resulting in no signal change when compared to the identical sequence without that bipolar gradient pair. If significant diffusion occurs between the two gradients, the refocusing by the second gradient will be incomplete due to the displacement of the molecule between the first and the second gradient and, therefore, signal attenuation will be observed. As signal attenuation depends on the direction of the bipolar diffusion gradient, orientation dependent diffusion may be observed (Fig. 22.24). Typical diffusion coefficients in normal human brain are about $10^{-3}\,\mathrm{mm}^2/\mathrm{s}$ [17].

22.7 References

[1] Bloch, F., (1946). Nuclear induction. *Phys. Rev.*, **70**:460–474.

[2] Purcell, E. M., Torrey, H. C., and Pound, R. V., (1946). Resonance absorption by nuclear magnetic resonance in a solid. *Phys. Rev.*, **69**:37–38.

[3] Lauterbur, P., (1973). Image formation by induced local interactions: Examples employing nuclear magnetic resonance. *Nature*, **242**:190–191.

[4] Damadian, R., (1972). Apparatus and Method for Detecting Cancer in Tissue. U.S. Patent 3789832.

[5] Albert, M., Cates, G., Drelehuys, B., Happer, W., Saam, B., Springer, C., and Wishnia, A., (1994). Biological magnetic resonance imaging using laser-polarized ^{129}Xe. *Nature*, **370**:199–200.

[6] Middleton, H., Black, R., Saam, B., Cates, G., Cofer, G., Guenther, R., Happer, W., Hedlund, L., Johnson, G., Juvan, K., and Swartz, J., (1995). MR imaging with hyperpolarized ^{3}He gas. *Magn. Reson. Med.*, **33**:271–275.

[7] Callaghan, P., (1991). *Principles of Nuclear Magnetic Resonance Microscopy*. Oxford: Clarendon.

[8] Hahn, E., (1950). Spin echos. *Phys. Rev.*, **20(4)**:580–594.

[9] Kumar, A., Welti, D., and Ernst, R., (1975). NMR Fourier zeugmatography. *Jour. Magn. Reson.*, **18**:69–83.

[10] Mansfield, P., (1977). Multi-planar image formation using NMR spin echoes. *J Physics*, **C10**:L55.

[11] Bottomley, P., Foster, T., Argersinger, R., and Pfeifer, L., (1984). A review of normal tissue hydrogen NMR relaxation times and relaxation mechanisms from 1-100 MHz: Dependence on tissue type, NMR frequency, temperature, species, excision, and age. *Med Phys*, **11**:425–448.

[12] Haase, A., Frahm, J., Matthei, D., Hänike, W., and Merboldt, K., (1986). FLASH imaging. Rapid NMR imaging using low flip-angle pulses. *J. Magn. Reson.*, **67**:258–266.

[13] Henning, J., Nauert, A., and Friedburg, H., (1986). RARE imaging, a fast imaging method for clinical MR. *Magn. Reson. Med.*, **3**:823–833.

[14] Gilles, R. J., (1994). *NMR in Physiology and Biomedicine*. San Diego: Academic Press.

[15] Tofts, P., (1997). Modelling tracer kinetics in dynamic Gd-DTPA MR imaging. *J. Magn. Reson. Imag.*, **7**:91–101.

[16] Greiner, A., Schreiber, W., and Brix, G., (1997). Magnetic resonance imaging of paramagnetic tracers in porous media: quantification of flow and transport parameters. *Water Resour Res.*, **33**:1461–1473.

[17] Ulug, A. M., Beauchamp, N., Bryan, R. N., and van Zijl, P., (1997). Absolute quantitation of diffusion constants in human stroke. *Stroke*, **28**:483–490.

23 Nuclear Magnetic Resonance Microscopy

Axel Haase, Jan Ruff, and Markus Rokitta

Physikalisches Institut, Universität Würzburg, Germany

23.1 Introduction

Nuclear magnetic resonance (NMR[1]) imaging has focused almost entirely on medical diagnosis. Widespread applications to the study of the anatomy, function, and biochemistry of different organs are used today. These NMR images are obtained using the ^{1}H NMR signal of water and lipid protons; NMR is based on the application of a high static magnetic field and a dynamic (radiofrequency) magnetic field. The (NMR-)frequency is proportional to the static magnetic field. An additional magnetic field gradient is used for NMR imaging. One obtains a radiofrequency dispersion proportional to spatial distribution of ^{1}H NMR-signal in the object. The proportionality constant is given by the magnetic field gradient strength.

[1]Both acroyms, NMR and MR, are used for nuclear magnetic resonance, compare Chapter 22

Handbook of Computer Vision and Applications
Volume 1
Sensors and Imaging

ISBN 0-12-379771-3/$30.00

In clinical NMR imaging static magnetic fields of up to 1.5 T and gradient strengths of up to 25 mT/m are used. Under these circumstances, a routine spatial resolution in three dimensions of 1 mm can be achieved. NMR imaging can also be performed in NMR spectrometers having a small bore (approximately 9 cm) superconducting magnetic field strength of up to 11.7 T and a gradient strength up to 2000 mT/m. Here, the spatial resolution can be improved by a factor of at least 100. NMR imaging in this regime is called "*NMR microscopy*" [1].

During the past decade many useful biological applications of NMR microscopy have been described. This paper will describe a few typical examples in plant physiology and animal studies. Although the spatial resolution is far inferior compared to other microscopy experiments, e. g., light microscopy and electron microscopy, it has a few important advantages. NMR microscopy needs no sample preparation. It can be performed in vivo in a controlled physiological condition. The applied static and dynamic magnetic fields are completely non-invasive and non-destructive and no biological side-effects have been described so far. NMR microscopy can be performed repeatedly on the same object in vivo and time dependent changes become visible.

The technique is able to produce anatomical images and in addition to measure spatially resolved important biophysical and biochemical data. NMR relaxation times change the NMR signal intensity in a well-defined way and hereby observe the molecular dynamics of water and lipid molecules. The result is a high image contrast which is completely under experimental control without the need of extra staining or contrast media. The NMR signal is further very sensitive to all kinds of *motion*. Therefore microscopic motions like the molecular diffusion can be observed and quantitatively evaluated. Macroscopic movements like the flow of water and blood can be measured and vessels clearly identified. Due to the fact that the NMR experiments can be repeated on the same object in short time intervals, the detection of biological function is feasible.

NMR spectroscopy is an important analytical tool in chemistry. The molecular structure produces a (chemical) shift of the NMR frequency which is characteristic for a molecule. The NMR spectrum is therefore useful to detect molecular structures or identify and quantify molecules in a mixture of different components. In vivo NMR spectroscopy in intact biological objects presents a non-invasive biochemical analysis which can be performed in addition to NMR imaging.

The aim of this chapter is to describe the technique and principal biological applications of NMR microscopy. NMR-spectroscopy and the combination with imaging is a further fascinating method which is however beyond the scope of this contribution.

23.2 Methodology

23.2.1 NMR microscopy

NMR-imaging is a method that creates almost simultaneously NMR signals from different positions within the object (see Chapter 22). The NMR frequency is proportional to the static magnetic field. If this field varies linearly across the sample, the NMR frequency also changes. The superposition of the static magnetic field with a field gradient will, therefore, result in an NMR spectrum in which the resonance frequency gives the position of the NMR signal in the object. As magnetic field gradients can be produced independently in all three dimensions, NMR imaging is able to perform 3-D images of the object. The spatial resolution is hereby dependent on the magnetic field gradient strength. In principle, extremely high gradient strengths are applicable for imaging with an atomic resolution [2]. In more practical applications in biological objects, the spatial resolution is limited due to low signal-to-noise ratio (SNR) of the NMR signal and molecular diffusion.

The signal-to-noise ratio is proportional to the number of nuclei producing the NMR signal and to the square of the static magnetic field strength. Due to technical and physical limitations, a detectable NMR signal needs at least on the order of 10^{12} nuclei. Therefore, the minimum image volume will be on the order of $(10\ \mu\text{m})^3$ when the ^{1}H NMR signal of water is measured. However, this spatial resolution can be improved when higher magnetic fields or optimized NMR detection systems are used. It should be noted that an improvement of the spatial resolution by a factor of 10 needs an increase of the SNR by a factor of 10^3!

The spatial resolution is further limited by molecular diffusion effects [3]. In order to detect an NMR signal, a certain time interval is necessary, which is of the order of 1 ms. Assuming a molecular self-diffusion coefficient of water of 2.3×10^{-9} m^2/s, water molecules travel by Brownian motion a mean distance of approximately 1 μm in 1 ms. Therefore, the spatial resolution can not be improved beyond 1 μm due to this diffusion effect.

Both effects, the low SNR and the molecular diffusion, limit the spatial resolution of NMR microscopy to values of more than approximately 10 μm. The magnetic field gradient for imaging must be of the order of 100 mT/m or higher and the static magnetic field should be as high as possible to increase the SNR. Typical magnetic field values range between 7 T and 14 T, a factor of 10 larger than clinical magnets. In order to keep the measuring time of a single NMR signal short, the switching times of the magnetic field gradient have to be on the order of 0.1 ms. These technical conditions limit NMR microscopy to magnets having a small bore size of 15 cm or less.

Biological and medical applications of NMR microscopy are, therefore, relevant for the investigation of small objects, such as small plants and seedlings in botany, and mice, rats, or perfused organs in basic medical research. All imaging techniques discussed in Chapter 22, including fast methods, are feasible. In addition to the high image contrast with respect to NMR relaxation times, spin-density, magnetic susceptibility and the methods for 2-D and 3-D imaging, the technique is able to measure flow velocities and to perform *NMR angiography*.

23.2.2 NMR flow measurements

Flow NMR imaging can be performed using two different kinds of experiments: the time-of-flight effect and the measurement of NMR signal phase changes. It has been shown in Section 22.3 that the image acquisition needs the measurement of a series of NMR signals, which are separated in time by the repetition time TR. During this time interval TR, which is often shorter than the spin-lattice relaxation time T1, the spin system will not recover to the high longitudinal magnetization. Magnetization flowing into the slice of detection during TR will, therefore, exhibit a higher signal intensity as compared to the stationary spin system. This time-of-flight effect is responsible for the high signal intensity of blood vessels. A simple threshold in data processing will, therefore, show only high signal intensities, that is, blood vessels. By the acquisition of a series of 2-D images, a 3-D NMR angiogram can be reconstructed.

A further mechanism, especially for quantitative flow velocity measurements, is the flow dependent NMR signal phase change. For the following, we assume a constant flow velocity v parallel to the direction of a field gradient G. During time interval t, moving spins will change their NMR frequency. This will give a phase change $\Delta\phi$ of the NMR signal of these spins according to:

$$\Delta\phi = \gamma G t^2 v \tag{23.1}$$

where γ is the gyromagnetic ratio of the nucleus. The phase change can be measured quantitatively and is proportional to the flow velocity. The other parameters are under direct experimental control. An NMR signal phase can be transferred into a signal intensity and will be detected as a flow image contrast. For quantitative flow velocity measurements, one has to note that in most vessels in plants or in blood vessels, laminar flow with a parabolic flow profile exists. In most cases, the spatial resolution of NMR imaging is not sufficient to detect this flow profile. Here, an averaged mean flow velocity v_{avg} will be detected. Under these circumstances the flow dependent signal S is given by:

$$S \propto \sin^2(\gamma G t^2 v_{\mathrm{avg}})/(\gamma G t^2 v_{\mathrm{avg}}) \tag{23.2}$$

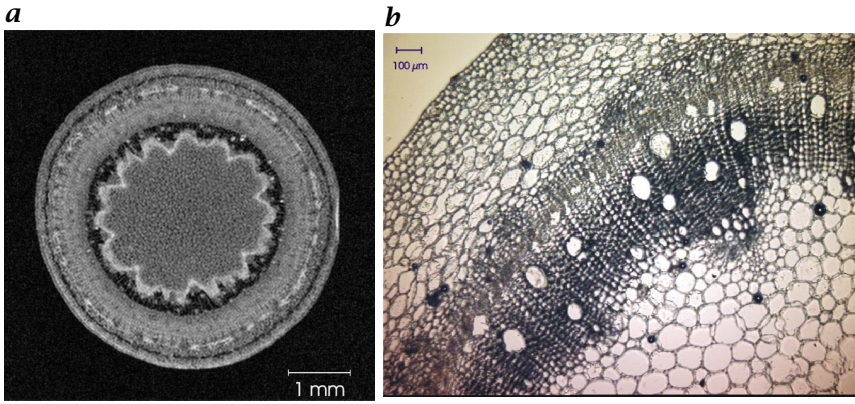


Figure 23.1: *Stem of Ricinus communis:* ***a*** *2-D cross-sectional NMR image with a spatial resolution of 23 µm.* ***b*** *2-D optical micrograph of a similar area as in* ***a****.*

For practical quantitative determinations of flow velocities, a series of NMR images have to be measured with different values of the gradient strength G to measure the function given in Eq. (23.2) precisely.

Finally, it should be noted that both methods for flow imaging use the intrinsic characteristics of the NMR signal. Flow measurements using NMR need no staining or the application of contrast media and are, therefore, completely noninvasive. In combination with fast NMR imaging, time-dependent changes of flow velocities can be observed.

23.3 Applications to plant studies

NMR studies in plant research are extremely rare at the moment, although the non-invasive character of the method makes it possible to observe anatomy, biochemistry and water transport simultaneously [1, 4]. In the following, we will show a few typical studies which can be performed on intact plants in vivo [5].

23.3.1 Materials and methods

All experiments were performed on a 7 T Bruker BIOSPEC horizontal bore magnet with a bore size of 20 cm. The magnetic field gradient system was able to achieve a gradient strength of up to 200 mT/m in a short rise time of 0.2 ms. The plants were installed horizontally in a homebuilt climate chamber with full climate control and the possibility of measuring transpiration and assimilation of the plant simultaneously with NMR flow studies [4]. The plant can be illuminated.

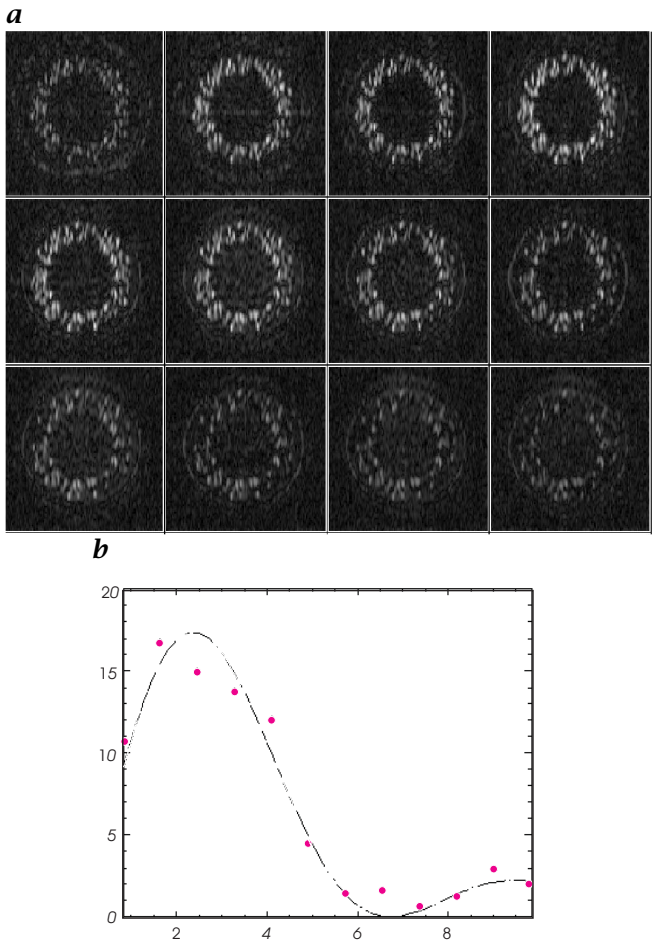


Figure 23.2: ***a*** *Twelve flow velocity weighted NMR images of the Ricinus communis plant stem with 12 different flow encoding gradient strengths. The gradient strength increases from top left to lower right.* ***b*** *Signal intensity vs gradient strength for a selected image element of the NMR data shown in* ***a****. The curve gives a fit of the experimental data according to Eq. (23.2).*

A Helmholtz-type radiofrequency coil with a diameter of 20 mm, tuned to the ^{1}H NMR frequency of 300 MHz, was used as transmitter and receiver coil for investigation of the plant stems. The NMR imaging sequence consists of two parts: flow encoding and imaging. The spatial resolution of the imaging experiment was 23 μm with a slice thickness of 600 μm. For quantitative flow studies, the spatial resolution had to be on the order of 200 μm to keep the measuring time short. The flow measurement has been done with phase encoding. Here, a stimulated

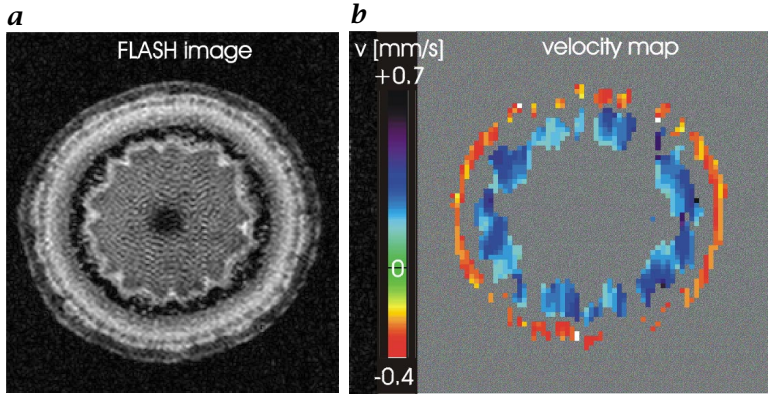


Figure 23.3: ***a*** *Cross-sectional NMR image of the plant stem of Ricinus communis.* ***b*** *Calculated flow velocity image using the image data shown in Fig. 23.2. Blue flow data are flow velocities directed from the roots to the leaves (xylem) and red data are flow velocities in the opposite direction; (see also Plate 11).*

echo experiment with two gradient pulses separated by 200 ms was applied. The gradient strength had to be stepped through with eight different values to observe the function given in Eq. (23.2) as precisely as possible.

23.3.2 Results

Figure 23.1a shows a 2-D cross section of the plant stem with a spatial resolution of 23 μm. The cortex, phloem, xylem, and pith parenchyma can be clearly differentiated and even single cells are visible. For comparison, Fig. 23.1b shows a part of an optical micrograph of this plant. It is clearly visible that the spatial resolution is far superior compared to the NMR image; however the main anatomical details are all identified in the NMR image.

Figure 23.2a shows a series of NMR images having a reduced spatial resolution for flow measurements. The flow encoding magnetic field gradient parallel to the axis of the plant stem was varied from image to image. It can be seen that the signal intensity in different portions of the plant is changed. The signal change is better visualized in Fig. 23.2b for one image element. Here, the curve is the best fit of the experimental data according to Eq. (23.2). From this curve, the average flow velocity and flow direction can be calculated. In Fig. 23.3, a calculated flow velocity map is shown. It displays an inner ring of xylem vessels having a flow direction from the roots to the leaves and flow velocities on the order of 0.5 mm/s and an outer ring of phloem vessels with lower flow velocities and an opposite flow direction.

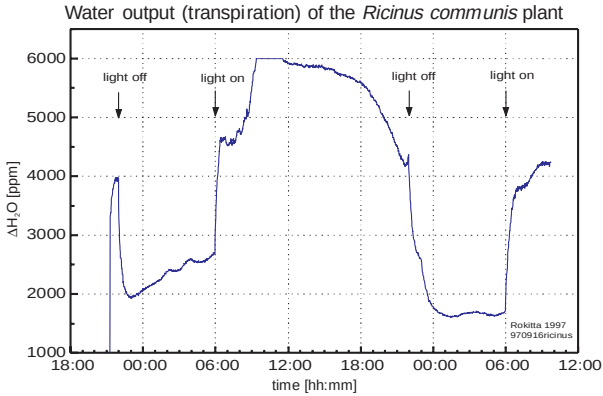


Figure 23.4: *Time-dependent transpiration rate of the plant Ricinus communis, following different illumination conditions and measured within the climate chamber.*

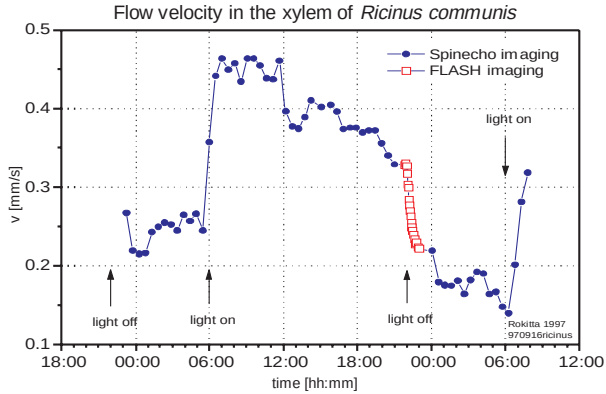


Figure 23.5: *Time dependent flow velocity changes as measured by NMR and averaged over the whole xylem area. The red squares are flow data measured using FLASH NMR (time resolution of 3.5 min), the blue circles are flow data measured with spin echo NMR imaging (time resolution 21 min).*

The experiment detected time-dependent flow velocity changes. In Fig. 23.4, the change of the plant transpiration is shown under varying illumination conditions. For example, high transpiration is observed when the light is switched on and transpiration changes immediately when the light is switched off. In Fig. 23.5 the flow velocities averaged over the whole xylem area as measured by the NMR technique are displayed. The averaged flow velocity increases from approximately 0.25 to 0.4 mm/s when the plant is illuminated and decreases when the light goes off. The measurement time of this experiment can be as short as

3 min when fast imaging sequences such as FLASH imaging are used. The time dependent behavior of the transpiration and flow change are very similar. The differentiation of flow velocities in different vascular bundles and the phloem will be important for future studies.

23.4 Applications to animal studies

Nuclear magnetic resonance imaging in basic medical research, especially in animal studies, is often performed in high magnetic fields. The high spatial resolution of NMR microscopy makes it feasible to investigate even small animals like mice. New animal models based on gene-overexpression, or -mutation are appearing rapidly and NMR microscopy might be an excellent method to study these animals. This is especially the case in the very popular mouse models. In the following, we will show NMR experiments in mice for the investigation of the heart [6]. These experiments will demonstrate that NMR microscopy can be applied even when motion is present (e.g., a beating heart or respiration).

23.4.1 Materials and methods

The NMR experiments were performed on a 7 T Bruker BIOSPEC horizontal bore magnet as described in the preceding. The NMR system was equipped with an extra set of gradient coils with a maximum strength of 870 mT/m and a rise time of 0.28 ms. A bircage Bruker NMR coil was used as a radiofrequency excitation and detection NMR coil. Imaging was performed using an ECG triggered cine FLASH. The repetition time was dependent on the R-R interval of the ECG signal. Typically, 12 frames per heart cycle with a minimum TR of 4 ms were acquired. To increase the SNR, four averages of the image data were needed. The spatial resolution of the 2-D images was 117 μm, slice thickness was 1 mm. The total measuring time was 100 s for one slice, or 10 to 13 min for 8 contiguous slices covering the whole heart.

Male C57bl/6 mice (mean mass of 17.8 g) were anesthetized with Isofluorane (4 % per l/min O_2 for initiation, 1.5 % as maintenance dose during the experiment). The ECG wires were attached to the front paws. The ECG trigger signal was taken from a homebuilt ECG unit [7]. During the NMR experiment, the mouse had to be kept normothermic by a warming pad.

23.4.2 Results

Transaxial and coronal images offer a detailed view of the heart anatomy. In Fig. 23.6, a coronal view is displayed showing high signal intensity

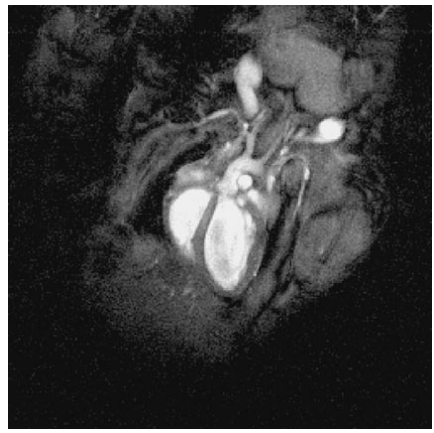


Figure 23.6: *Coronal NMR image of the mouse heart. The ECG-triggered image was acquired in diastole. The spatial resolution is 117 µm.*

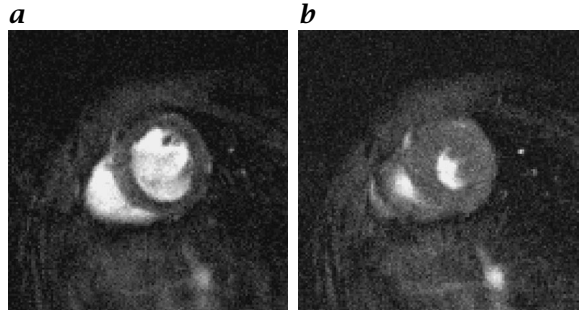


Figure 23.7: *Transaxial NMR images of the mouse heart:* ***a*** *in diastole; and* ***b*** *in systole. For movies of the beating heart see /movies/23 on the CD-ROM (mouse.mov).*

of the water-filled left and right ventricles, the heart muscle, and small vessels. Transaxial images are given in Fig. 23.7a,b for systole and diastole. The changing volume of the ventricles is clearly visible. On the basis of an intensity threshold, different parts of the anatomy can be segmented. In Fig. 23.8, the left ventricle muscle and cavity are shown in systole and diastole.

Due to the fact that time-dependent changes of the anatomy have been observed using the NMR technique, a few important functional parameters can be calculated, including the left ventricle mass (LVM), ejection fraction (EF), cardiac output (CO), and stroke volume (SV). On the basis of eight animal studies in a pilot experiment, we observed the following values:

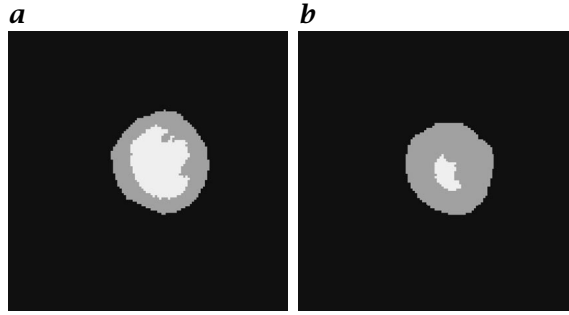


Figure 23.8: *Segmented NMR images according to a signal threshold using the image data shown in Fig. 23.7 from the left ventricle:* ***a*** *in diastole; and* ***b*** *in systole.*

LVM:	76.1 ± 6.5 mg	EF:	68.6 ± 6.6 %
CO:	11.2 ± 2.4 ml/min	SV:	30.5 ± 4.6 µl

There was a high reproducibility of the NMR experiments with low intraobserver (5 % mean value), interobserver (5 % mean value), and interstudy variability (3 % mean value).

23.5 Discussion

Examples of application of NMR microscopy, both in plant physiology and in animal studies, demonstrate the feasibility and accuracy of the *in vivo study* of the anatomy and function of intact biological systems. The method is well suited to a noninvasive investigation of intact plants and live animals. The spatial resolution can be improved up to approximately 20 µm in stationary objects or 100 µm in moving parts in animals. All anatomical details can be visualized with high image contrast. There is no artifact due to sample preparation as is the case in other microscopy studies. The NMR imaging has the capability to measure the object with a 3-D technique. The 3-D information is available in a digital format, which makes further quantification of volumes, surfaces, and distances feasible.

The imaging experiment can be repeated rapidly, depending on the available SNR. In principle, the time resolution can be as short as 4 ms, as demonstrated by ECG-triggered imaging of the mouse heart. It is important to note that the time resolution can be adapted to the needs in the different biological studies, that is, in the millisecond range for cardiac imaging or in the minute range for time dependent flow measurements in plants. Therefore, NMR microscopy is an ideal technique for in vivo functional studies in biological and medical research.

23.6 References

[1] Callaghan, P., (1991). *Nuclear Magnetic Resonance Microscopy*. Oxford: Clarendon Press.

[2] Yannoni, C., Züger, O., Rugar, D., and Sidles, J., (1996). Force detection and imaging in magnetic resonance. In *Encycl. of Nucl. Magn. Reson.*, pp. 2093–2100. New York: John Wiley & Sons.

[3] Brandl, M. and Haase, A., (1994). Molecular diffusion in NMR microscopy. *J. Magn. Reson. Series B*, **103**:162–167.

[4] Kuchenbrod, E., Landeck, M., Thürmer, F., Haase, A., and Zimmermann, U., (1996). Measurement of water flow in the xylem vessels of intact maize plants using flow-sensitive NMR imaging. *Bot. Acta*, **109**:184–186.

[5] Rokitta, M., Zimmermann, U., and Haase, A., (1999). Fast NMR flow measurements in plants using NMR imaging. *J. Magn. Reson. In press*.

[6] Ruff, J., Wiesmann, F., Hiller, K.-H., Voll, S., von Kienlin, M., Bauer, W., Rommel, E., Neubauer, S., and Haase, A., (1998). Microscopic magnetic resonance imaging for non-invasive quantification of myocardial function and mass in the mouse. *Magn. Reson. Med.*, **40**:43–48.

[7] Rommel, E. and Haase, A., (1995). An ECG trigger unit optimized for fast rat heart imaging. In *Proc. SMR/ESMRMB, 3rd Annual Meeting, Nice*, p. 938.

Index

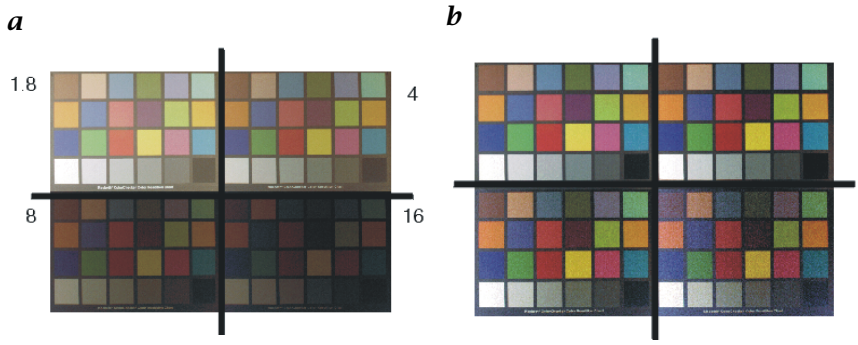


Plate 1: ***a*** *Log image of McBeth chart with f-stops as indicated;* ***b*** *same as* ***a*** *but normalized to black and white for each quarter; (see also Fig. 8.10, p. 232)*

Plate 2: *Noiseless McBeth chart generated synthetically for comparison with Fig. 8.10; (see also Fig. 8.11, p. 232)*

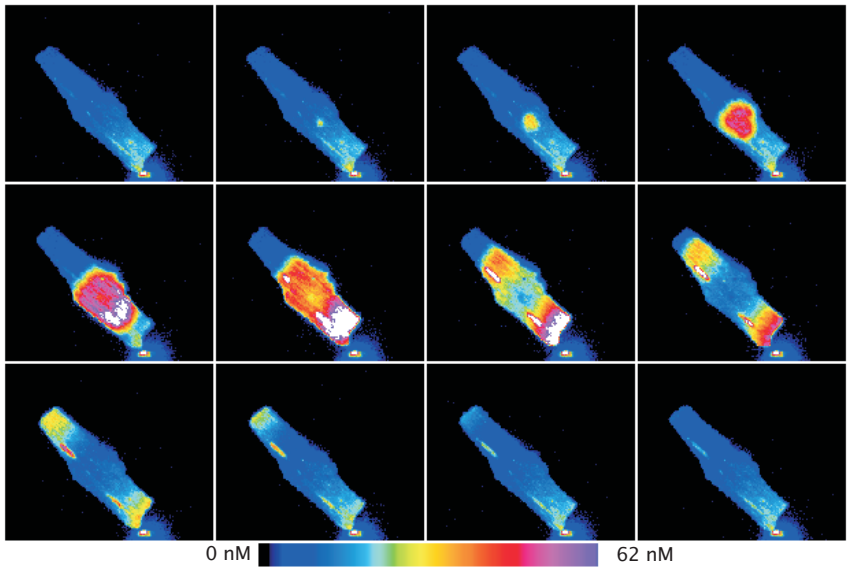


Plate 3: *Example of a fast spatially resolved Ca^{2+}-image sequence. Cardiac myocytes were labeled with the Ca^{2+}-sensitive indicator Fluo-3 (2 µM) and spontaneous Ca^{2+}-waves propagating inside the myocyte can be seen. The sequence was recorded with a MERLIN system and an Astrocam frame transfer camera (Life Science Resources Inc., Cambridge, UK) with an EEV37 CCD chip read out at 5.5 MHz to capture the images. [Figure courtesy of Dr. B. Somasundaram, Life Science Resources Inc. and Dr. N. Freestone of Babraham Institute, Babraham, Cambridge UK]; (see also Fig. 12.3, p. 334)*

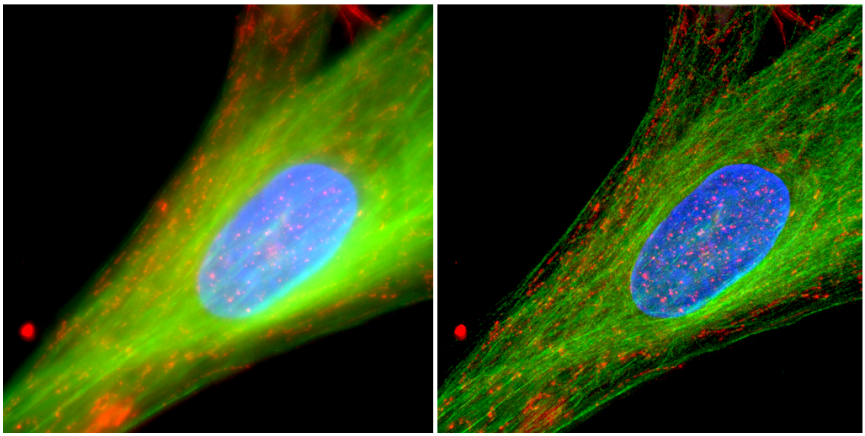


Plate 4: *Example of a nearest-neighbors deblurring algorithm. The image of a human skin fibroblast consists of three separate 12 bit gray-scale images, each recorded with a different fluorescent dye. The cell has been processed for double immunofluorescence and counterstained with Hoechst 33258 for DNA. Microtubules in the cell are localized with a monoclonal IgG antibody to beta-tubulin followed by a secondary antibody tagged with FITC. Mitochondria are localized with a monoclonal IgM antibody to a novel protein, followed by a secondary antibody tagged with Texas Red. For FITC, excitation filter = 485 nm, with a barrier filter at 530 nm. For Texas Red, the excitation filter was 560 nm, with a 635 nm barrier filter. Traditional UV filters were used for the Hoechst dye. Images were captured at six depths at 1 µm steps and each color channel was deblurred separately resulting in the deconvolved image at the right side. [Figure courtesy of Dr. R. Zinkowski, Molecular Geriatrics Corp., Vernon Hills, IL and Dr. Chris MacLean, VayTek Inc., Fairfield, IA]; (see also Fig. 12.4, p. 335)*

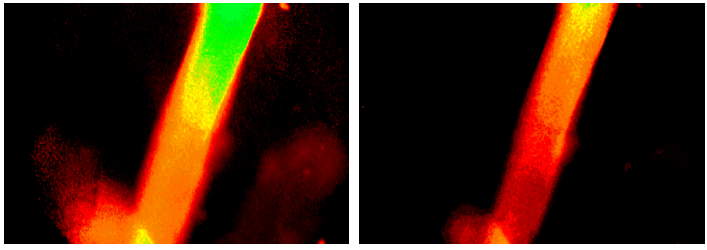
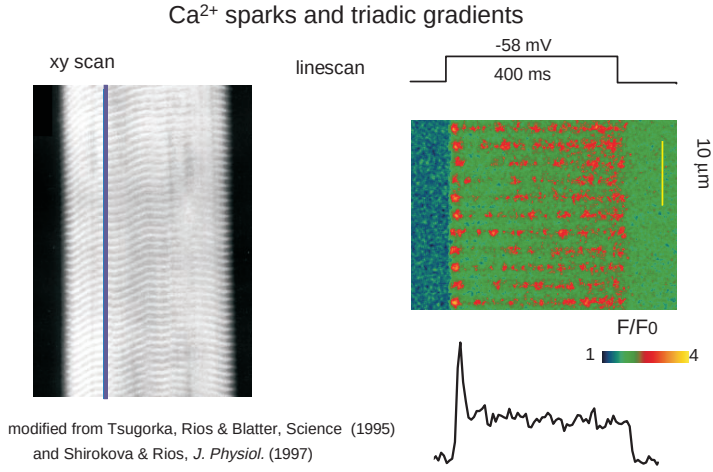


Plate 5: *Example of a no-neighbors deblurring algorithm. A muscle fiber from Xenopus laevis M. lumbricalis with a diameter of 100 µm was stained with the ratiometric Na^+-indicator SBFI-AM (340 nm/380 nm ratio). The deblurred image on the right contains less out-of-focus information than the original ratio image on the left; (see also Fig. 12.5, p. 336)*

a



b

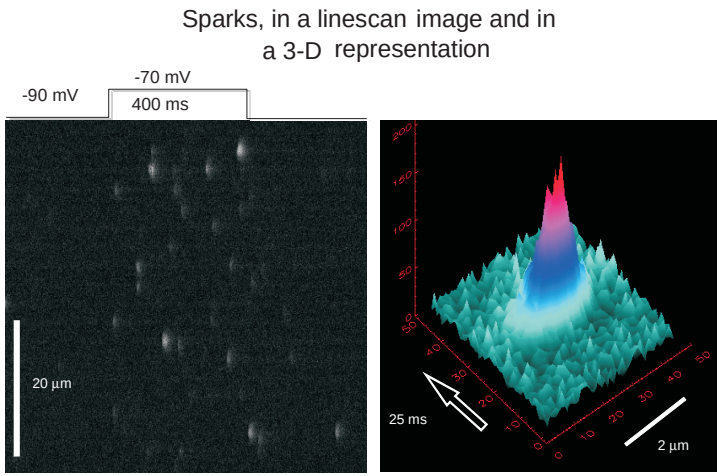


Plate 6: *Ca^{2+}-sparks measured in skeletal muscle fibers from Rana pipiens with the fluorescent indicator Fluo-3 under voltage clamp conditions:* ***a*** *line-scan image of Fluo-3 fluorescence upon 400 ms depolarization, F/F_0 is the normalized fluorescence;* ***b*** *3-D representation of a spark as reconstructed from the linescan image data. [Figure courtesy of Prof. E. Rios, Rush University, Chicago, IL, USA]; (see also Fig. 12.6, p. 338)*

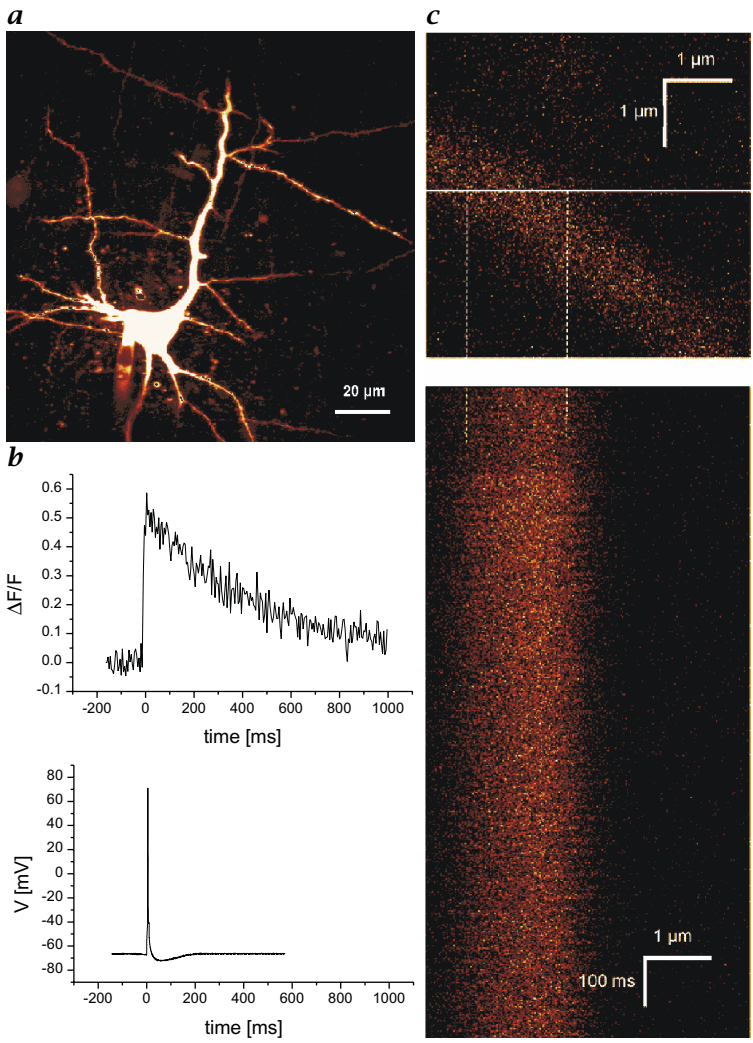
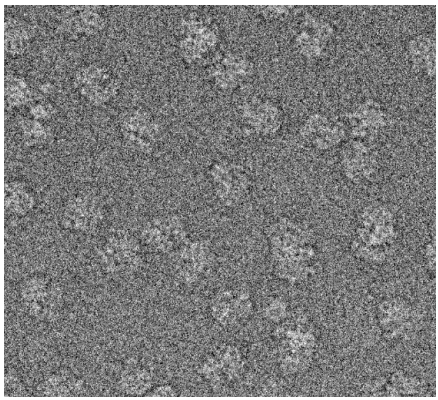


Plate 7: *Example of two-photon microscopy in brain slices.* ***a*** *A neocortical layerV pyramidal cell in a rat brain slice was filled via somatic whole-cell patch pipettes with the calcium indicator Calcium Green-1 (100 µM) or Oregon Green 488 BAPTA-1 (100 µM);* ***b*** *upper trace: calcium fluorescence transient evoked by a single backpropagating dendritic action potential; lower trace: Electrophysiological recording of the AP with somatic whole cell recording in current-clamp mode;* ***c*** *Linescan through a basal dendrite: fluorescence was recorded in linescan-mode. Upper picture: The line in the xy-image shows the position of the linescan. Lower picture: The linescan had a length of 1160 ms. All points in one line between broken lines were averaged. (Figure courtesy of Helmut Köster, Max-Planck-Institut für Medizinische Forschung, Heidelberg); (see also Fig. 12.7, p. 340)*

a



b

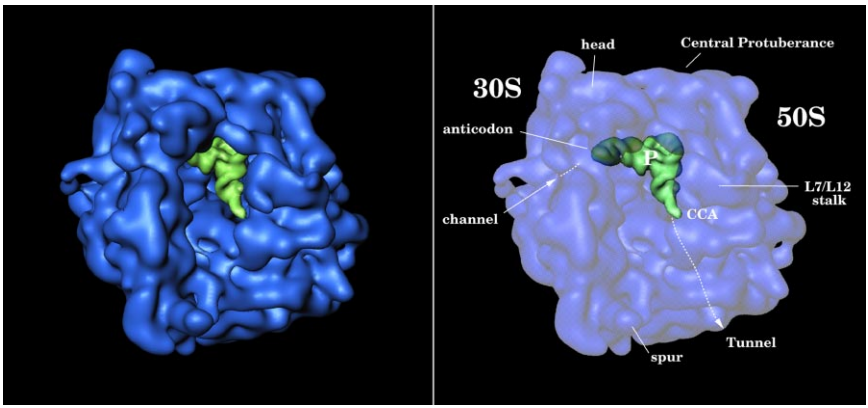


Plate 8: ***a*** *70S ribosomes from Escherichia coli, visualized by cryo-electron microscopy. Electron optical magnification 50,000×.* ***b*** *The ribosome at 15 Å, reconstructed from 30,000 projections obtained by cryoelectron microscopy, shown with a tRNA in the P-site position as experimentally found. The anticodon of the tRNA is in the vicinity of the channel that is thought to conduct the messenger RNA, while the acceptor end (marked CCA) is seen to point toward the opening of the tunnel that is believed to export the polypeptide chain [Prepared by Amy Heagle and Joachim Frank, Laboratory of Computational Biology and Molecular Imaging, Wadsworth Center]; (see also Fig. 13.12, p. 376)*

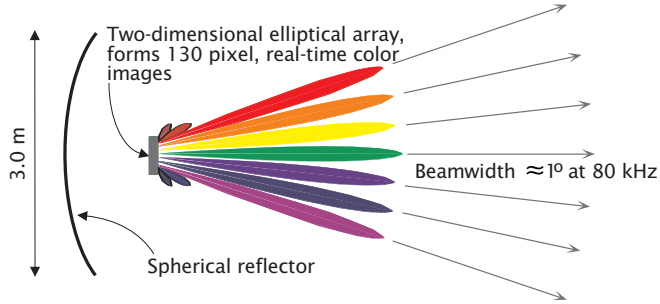


Plate 9: *Schematic showing the spherical reflector, array head, and fan of beams; (see also Fig. 15.3, p. 419)*

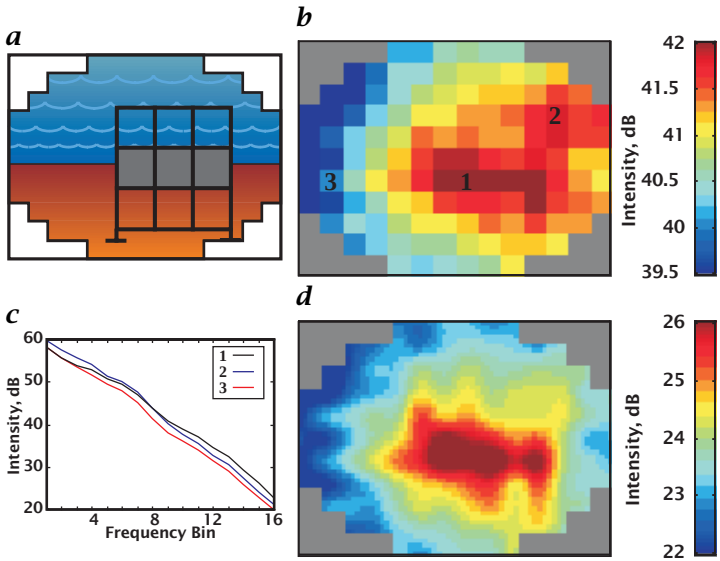


Plate 10: ***a*** *Simulated view of the rectangular bar target mounted vertically on the seabed;* ***b*** *acoustic daylight image of the bar target from raw intensity data;* ***c*** *spectra of the pixels labeled 1, 2, and 3 in* ***b****;* ***d*** *interpolated acoustic daylight image of bar target from the same data used in* ***b****; (see also Fig. 15.4, p. 421)*

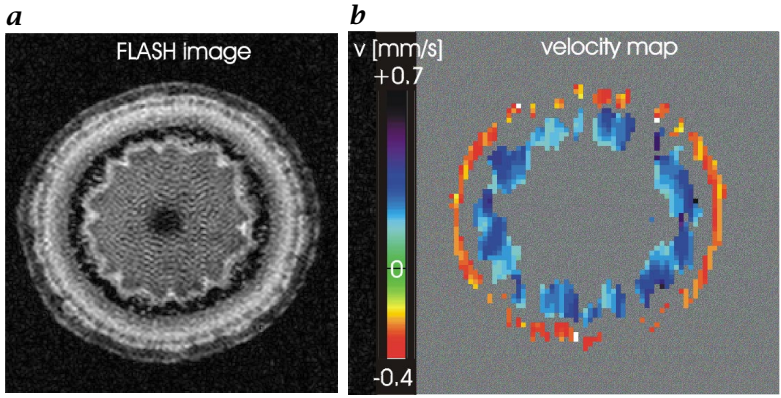


Plate 11: a *Cross-sectional NMR image of the plant stem of Ricinus communis.* ***b*** *Calculated flow velocity image using the image data shown in Fig. 23.2. Blue flow data are flow velocities directed from the roots to the leaves (xylem) and red data are flow velocities in the opposite direction;* *(see also Fig. 23.3, p. 607)*