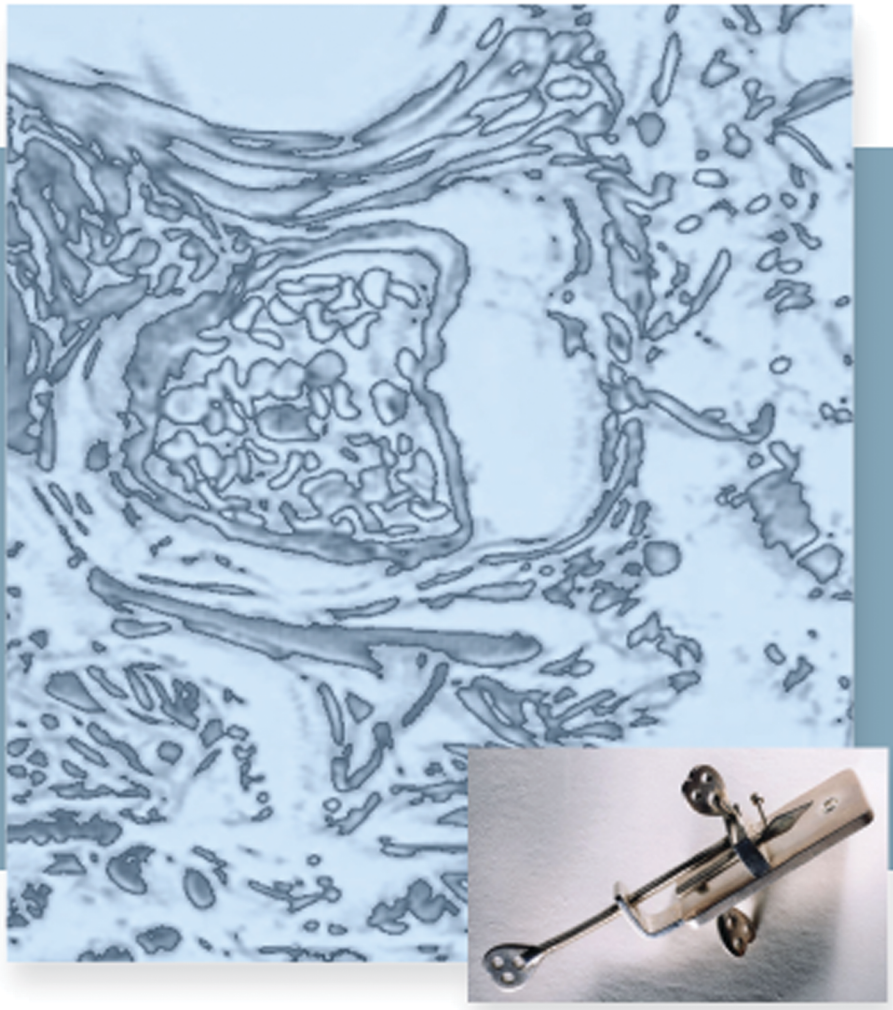
SERIES IN OPTICS AND OPTOELECTRONICS

The Limits of Resolution



Geoffrey de Villiers
E. Roy Pike

CRC Press
Taylor & Francis Group
A TAYLOR & FRANCIS BOOK

The Limits of Resolution

SERIES IN OPTICS AND OPTOELECTRONICS

Recent titles in the series

The Limits of Resolution
Geoffrey de Villiers and E. Roy Pike

Polarized Light and the Mueller Matrix Approach
José J Gil and Razvigor Ossikovski

Light—The Physics of the Photon
Ole Keller

Advanced Biophotonics: Tissue Optical Sectioning
Ruikang K Wang and Valery V Tuchin (Eds.)

Handbook of Silicon Photonics
Laurent Vivien and Lorenzo Pavesi (Eds.)

Microlenses: Properties, Fabrication and Liquid Lenses
Hongrui Jiang and Xuefeng Zeng

Laser-Based Measurements for Time and Frequency Domain Applications: A Handbook
Pasquale Maddaloni, Marco Bellini, and Paolo De Natale

Handbook of 3D Machine Vision: Optical Metrology and Imaging
Song Zhang (Ed.)

Handbook of Optical Dimensional Metrology
Kevin Harding (Ed.)

Biomimetics in Photonics
Olaf Karthaus (Ed.)

Optical Properties of Photonic Structures: Interplay of Order and Disorder
Mikhail F Limonov and Richard De La Rue (Eds.)

Nitride Phosphors and Solid-State Lighting
Rong-Jun Xie, Yuan Qiang Li, Naoto Hirosaki, and Hajime Yamamoto

Molded Optics: Design and Manufacture
Michael Schaub, Jim Schwiegerling, Eric Fest, R Hamilton Shepard, and Alan Symmons

An Introduction to Quantum Optics: Photon and Biphoton Physics
Yanhua Shih

Principles of Adaptive Optics, Third Edition
Robert Tyson

The Limits of Resolution

Geoffrey de Villiers
University of Birmingham

E. Roy Pike
King's College London

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A TAYLOR & FRANCIS BOOK

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

Printed on acid-free paper
Version Date: 20160802

International Standard Book Number-13: 978-1-4987-5811-6 (Hardback)

Library of Congress Cataloging-in-Publication Data

Names: De Villiers, Geoffrey, author. | Pike, E. R. (Edward Roy), 1929-
author.
Title: The limits of resolution / Geoffrey de Villiers and E. Roy Pike.
Other titles: Series in optics and optoelectronics (CRC Press) ; 22.
Description: Boca Raton : CRC Press, Taylor & Francis Group, CRC , [2016] |
Series: Series in optics and optoelectronics ; 22 | Includes
bibliographical references and index.
Identifiers: LCCN 2016013914| ISBN 9781498758116 | ISBN 1498758118
Subjects: LCSH: Resolution (Optics) | Inverse problems (Differential
equations)--Numerical solutions. | Differential equations,
Partial--Improperly posed problems. | Functional analysis. | High
resolution imaging.
Classification: LCC QC355.3 .D4 2016 | DDC 535--dc23
LC record available at https://lccn.loc.gov/2016013914

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

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

Contents

Preface

The purpose of this book is to bring together some basic ideas underlying the theory of resolution, with particular emphasis on the area of linear inverse problems. Resolution is a key concept in experimental physics, where it limits what we can determine about the physical world we inhabit and where many of the inverse problems are linear.

The degree of resolution can be said to quantify the level of detail in an object which may be recovered reliably. Having said that, since the days of Lord Rayleigh, resolution has continued to be a confusing and controversial subject with various prevailing resolution criteria. Much of this confusion is due to not making explicit the prior information or assumptions one is using, or, more generally, not being clear about what problem one is solving.

The most familiar ideas about resolution come from the knowledge of optical imaging associated with our own visual system. Historically, the resolution associated with optical instruments has been assessed using the human eye, and the instruments have been modified to improve their resolving power, with the aim of making the image clearer to the eye. However, with modern data-recording methods we now have the option of recording images and then processing the data to improve the resolution further.

Whereas with optical imaging the data (or image) is often a blurred form of the object under investigation and the object is vaguely recognisable in the image, in other problems where resolution is important there is nothing in the data which is visually recognisable and there is no choice but to carry out an inversion, that is to solve an inverse problem, before one can see a recognisable object. This inversion opens a Pandora's box of problems, and the key problem is the non-uniqueness of the solution. There are typically an infinite number of solutions, all of which fit the data and these will have varying degrees of detail in them.

Resolution is then to do with trying to sort out how much of this detail is reliable and how much of it is attributable to noise on the data. Loosely speaking, the inversion process amplifies this noise, and it this which gives rise to unreliable detail in the reconstructed object. For a given level of noise on the data, some of the solutions to the inversion problem will have more unreliable details than others, and an appropriate inversion method must be chosen to weed these out.

There are a wide range of inversion methods which take the noise into account and each will pick out a different solution. There is a whole industry concerned with finding 'best' or 'optimum' solutions. It is important that *any* solution, delivered according to some explicit or implicit patent algorithm, is taken for what it is, that is, one person's choice out of an infinite number of possible fits to the given data. However, each problem, although not having a 'best' solution in the abstract will, no doubt, have a 'customer' who requires information from the data and who will usually object strongly (but irrationally!) if they are given more than one answer under the same conditions. This fact has to be faced, and pure science then has to be set aside to come to some sensible compromise.

In Fourier optics, one uses decomposition into trigonometric functions to analyse resolution via the use of 'spectral' transfer functions. In this book, we will instead use singular-value decomposition (SVD) for problems where Fourier theory is not appropriate. The SVD is ideally suited for most linear inverse problems since the 'spectrum' of singular values of a compact integral operator always accumulates to zero and the method will provide a lower 'representational entropy' than any other decomposition. The spectrum will fall more quickly below the ambient noise level and the problem will be reduced to one involving the minimum number of component functions.

A particularly striking example of this is an analysis of the optics of the optical compact disc, where we shall see in Chapter 10 that just the first two terms of the singular-value decomposition cover 99% of the response. This is in contrast to the original analysis of Hopkins and Braat of 1979, which used Fourier modulation transfer functions, and which, although only applying to the paraxial case, was a great deal more lengthy and complicated. The SVD method has much to offer over conventional Fourier optics for general use in the design of optical and, in fact, many other systems. Dennis Gabor, the inventor of holography, was famously concerned about applying Fourier analysis in cases of finite object support, which, in real life, is always the case. In Chapter 1, we discuss his proposal of 'logons' as units of time-frequency to try to deal with the problem. An SVD analysis avoids this difficulty and always provides tailored orthonormal bases over the actual object support for the calculation in hand.

We will see that the use of the singular-value decomposition also gives insight into resolution in various other approaches to solving linear inverse problems, such as the Landweber iteration. Central to the singular-function approach is the study of oscillation properties of sums of singular functions, and this forms a significant theme in this book. It will also be used to design image-plane masks for super-resolving microscopy.

Given that the problems we are interested in typically have an infinite number of solutions and there are a very large number of solution methods, coupled with the fact that modifying the experimental apparatus essentially defines a new inverse problem, it should be clear that trying to cover all aspects in a single book is inconceivable. For example, we do not cover the important subjects of optical lithography, where recent advances in resolution have been achieved using surface plasmon polaritons, and telescopy where an international consortium using a Large Binocular Telescope Interferometer has superbly resolved a volcano, Loki Patera, on the Jovian moon Io. Instead we have made a personal choice of problems and methods which we feel give insight into the concept of resolution. Inevitably others will disagree on which problems and methods should be covered, but we use authors' privilege in making a selection. Of course, we intend no slight to any persons whose work is not covered. Within our remit, we also consider how given types of apparatus can be modified to improve their resolving power, for example, using apodisation in optics and non-uniform aperture weighting in radar.

The book is based on an inter-collegiate postgraduate series of lectures in the University of London given by one of the authors (ERP) and is aimed at postgraduates and researchers. It is intended to serve as a reference text for the latter whilst having pedagogical value for the former.

There are several excellent textbooks on general linear inverse problems in existence (see, for instance, Alberto Tarantola, *Inverse Problem Theory and Model Parameter Estimation*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2005, P. C. Sabatier (ed.), *Inverse Problems: An Interdisciplinary Study*, Academic Press, Cambridge, MA, 1987, H. W. Engl, M. Hanke and A. Neubauer, *Regularisation of Inverse Problems*, Springer, New York, 2000, C. Smith and W. T. Grandy, (eds.), *Maximum-Entropy and Bayesian Methods in Inverse Problems*, D. Reidel Publishing Co., Dordrecht, Netherlands, 1985, A. K. Louis, *Inverse und schlecht gestellte Probleme*, B. G. Teubner, Stuttgart, Germany, 1989, M. Bertero and P. Boccacci, *Introduction to Inverse Problems in Imaging*, CRC Press, Boca Raton, FL, 1998, P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1998 and C. R. Vogel, *Computational Methods for Inverse Problems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2002). In contrast, this book is concerned with resolution in linear inverse problems. Many resolution techniques are covered in detail in other books (such as maximum entropy in Smith and Grandy), and although we have hoped to make intelligent comments, we have not gone into the same amount of technical detail. Instead we refer the reader to the appropriate texts. On the subject of resolution, there are various books on super-resolution, but these presuppose one knows what ordinary resolution is. One of the aims of this book is to fill that gap.

An outline of the book is as follows: Chapter 1 starts with a discussion on early lenses. Though the purpose for which these were made remains sometimes controversial, the fact is that they are

lens shaped and generally have magnifying properties, hence we feel justified in calling them lenses, unless obviously ornamental. We then move on to other historical advances in optics, such as microscopes and telescopes. Anticipating links with communication theory we also give a very brief history of telegraphy. We discuss in some detail the pioneering work of Abbe and von Helmholtz on resolution in microscopy. Key ideas in the theory of resolution such as the Rayleigh resolution criterion and the Shannon sampling theorem are then elaborated, together with the Shannon 2WT theorem and channel capacity.

Following Gabor, who was inspired by the Smith–Helmholtz paraxial lens theorem in optics, we discuss the analogy between the Shannon 2WT theorem and the invariance of étendue in optical systems. We shall see that both 2WT and étendue apply, both in optics and in communication theory, to full rather than 'paraxial' theorems. The two subjects can now, therefore, be seen to have common mathematical roots and developments.

We will also look at basic ideas on apodisation in optics and beam-pattern design in radar.

In Chapter 2, we move into more modern concepts of resolution which began with the famous solution by Slepian in 1954 of the first-kind Fredholm integral equation which describes simultaneously time-limited and low-passband signals, applicable, for example, to coherent imaging and communication theory. This theory contains the 2WT theorem as a limiting case. We go on to cover subjects such as super-directivity and apodisation and then give a brief introduction to linear inverse problems; as examples, we consider coherent and incoherent imaging. The first of a series of papers on super-resolution in optical microscopy by Bertero, Pike and colleagues was published in Optica Acta in 1982 and is introduced here. The following body of work led to the various coherent and incoherent super-resolving microscopes built at King's College London, which will be described in Chapter 10. Chapter 2 concludes with a discussion on the quantum limits of optical resolution.

Chapter 3 is devoted to some topics in linear algebra and functional analysis which are central to linear inverse problems. Indeed these problems can be viewed as exemplars in applied functional analysis. In order to make the book more readable, however, we have relegated some of the more sophisticated mathematics to an appendix.

In Chapter 4, we study the connection between noise level and resolution, firstly, for systems described by compact forward operators (such as the finite Laplace transform) in the case where the singular functions form a Chebyshev system. Secondly, we look at convolution and Mellin-convolution operators, and we finish with solutions to linear inverse problems with discrete data.

Chapter 5 is concerned with some of the optimisation theory used in the solution of linear inverse problems. This body of mathematics is so fundamental to the subject that we include a chapter on it rather than place it in an appendix.

Chapter 6 is concerned with methods for solving linear inverse problems which are deterministic in nature; by this we mean that the additive noise on the data may be random but the object and noise-free data are supposed to be ordinary functions or vectors. The incorporation of positivity is also considered here.

Chapter 7 consists of a discussion of what might loosely be termed deterministic spectral analysis. This is basically Fourier analysis where only a finite number of Fourier coefficients are known and hence the problem of determining the full Fourier series is ill-posed. The resulting mathematical structure is fundamental to inverse problems in optics as well as to statistical spectral analysis.

Statistical methods for solving linear inverse problems are considered in Chapter 8. Here the object (solution), image (data) and noise are all postulated to be random variables of one kind or another.

In Chapter 9, we look at some important practical optical applications, including particle sizing by photon-correlation spectroscopy, in which one has to deal with the difficult problem of Laplace-transform inversion, laser Doppler velocimetry, projection tomography and diffraction tomography.

Finally in Chapter 10, we look at resolution in optical microscopy. This subject is associated with a wide range of technologies, ranging from conventional and fluorescence microscopy to the use of meta-materials as super-lenses.

Within fluorescence microscopy, we first discuss the use of super-resolving, image-plane, optical masks. Their design, calculation, appearance and use are considered.

The super-resolving performance of confocal image-plane-mask systems, was first shown theoretically in the incoherent case in 1989 in Reference 13 of Chapter 10, and both theoretically and experimentally, also in the incoherent case, in 1993 in Reference 12 and in 1998 in Reference 18. Such highly super-resolved images were further published in 2002 in Reference 20, as reproduced in figure 10.18. These resolution gains are significantly greater than those achieved in Reference 50 in 1999 using the STED technique. Comparable resolutions are 84 nm and 106 nm, respectively, taking into account the differences in wavelength and N.A.

The cover picture of this book, micrographs of canine blood cells, uses the right panel of Figure 10.20. This figure, showing confocal and super-resolved images side by side, shows very graphically the improvement in clarity and resolution over the standard confocal image gained by using such masks. The effective numerical aperture of this confocal, image-plane-mask system, using an oil-immersion lens of numerical aperture 1.3, was 2.14.

A short section follows on optical-disc systems. In the past, the compact disc was a prime target for super-resolving optics and between 2001 and 2004 a European Union consortium 'Super Laser Array Memory' (SLAM), contract no. IST-2000-26479 involving a number of partners, including King's College London, worked on various approaches under the leadership of Joseph Braat at Philips Research Laboratories, Eindhoven. Although interesting results, including theoretical prediction and computer simulation of super-resolution by the use of image-plane masks, were obtained, the project was really overtaken by the advent of streaming from solid-state memories and was not renewed.* However, as mentioned earlier in this preface, there is an important result in this section where it is shown that in an analysis of the optics of compact discs using a basis of singular functions, the first two terms cover 99% of the response.

Chapter 10 continues with a section on near-field super-lenses first described by Pendry. A full electromagnetic simulation in MATLAB® of this plasmonic phenomenon is included in Appendix F.

We then discuss recent super-resolution achieved by structured illumination, non-linear fluorophore responses and localisation of isolated single molecules, both by optical and by ultrasound microscopy.

We include various technical appendices. In Appendix A, we discuss the origins of spectacles and present new findings which upset conventional wisdom. Here we are conscious of travelling through somewhat delicate territory, and we urge interested readers to delve further and then make up their own minds. As pointed out to us by Christine De Mol at the Free University of Brussels, history is also an ill-posed inverse problem.

We are happy to acknowledge a great debt of gratitude to Mario Bertero of the University of Genoa, who has guided our physical considerations for many years with a keen mathematical eye, and to his colleagues Paula Brianzi and Patrizia Boccacci, with whom we have also worked very profitably over the same period. Professor Bertero, in fact, was a lead researcher in the high astronomical resolution of the volcano on Io referred to earlier. We also thank Christine De Mol, with whom we have collaborated extensively over many years, including a period within a NATO-funded consortium between London, Brussels and Genoa.

The same goes for Pierre Sabatier for his continued personal interest in our work and his hosting of the burgeoning inverse problems community for many years at his RCP 264 workshop in Montpellier. His acceptance of the first editorship of the UK Institute of Physics Journal *Inverse Problems* set it off on an exemplary path.

* The capacity of the current Blu-ray compact disc could now be at least doubled using image- or pupil-plane masks.

ERP thanks Nicole Ostrowsky of the University of Nice and her former colleagues, Didier Sornette and Kehua Lan, for encouragement and help in the baffling early days of our work on one of the most difficult of all practical linear inversion problems, the Laplace transform with real data, which is central to particle-size measurement by photon-correlation spectroscopy (PCS), considered in Chapter 9. We were also aided considerably in this work by the experimental and programming skills of Marvin Young, who spent a post-graduate period with us from the United States. The major application of PCS has been by Malvern Instruments Ltd. in the United Kingdom, who licensed our early patents and now market particle-characterisation instrumentation worldwide.

Our thanks are also due to our former colleagues at RSRE, Malvern, in particular, John McWhirter, in the thick of it from the beginning, Ian Proudler, Kevin Ridley, Alan Greenaway (formerly at Heriot-Watt University), the late David Broomhead (formerly of Manchester University), Greg King, Robin Jones, Eric Jakeman, Peter Pusey, John Rarity (now at the University of Bristol) and Brian Roberts. We also thank John Walker (now at the University of Nottingham), and past King's College postgraduate students Richard Davies, Ben McNally, Gerard Hester, Deeph Chana, Fabienne Penney (née Marchaud) and Pelagia Neocleous, who have made significant contributions to our work on the inversion of a series of band-limited transmission problems in optics, sonar and radar. Ely Klepfish has applied and developed our methods to difficult analytic-continuation problems in theoretical physics which arise in high-temperature superconductivity theory, to some problems of the early universe and even a joint foray with us more recently into financial applications. Jan Grochmalicki, Ulli Brand and Shi-Hong Jiang took the brunt of the experimental work and supercomputing needed in confocal microscopy and compact discs and we have benefited from many interactions with the optical storage group at Philips Research Laboratories in Eindhoven.

Finally, significant contributions have been made, particularly to our software suites, by summer internship students over a number of years, Emmanuel Cohen, Florent Colas, Sebastian Demoustier, Xavier Esnault, Akil Hlioui, Didier Laval, Eric Pailharey, Cyril Polinacci, and Christophe Ramananjaona from the Ecole Supérieure d'Optique at Orsay and Krzysztof Roszkowski from the Technical University of Warsaw.

We are grateful for support over the years from a number of funding bodies, particularly the UK Engineering and Physical Sciences Research Council, the European Commission, the U.S. Army Research Office, DARPA, the NATO Scientific Affairs Division and the Royal Society.

Thanks are also due to Dan Smith and Paul Sykes at Aspect Printing in Malvern for help and support with the preparation of this book as well as much support from Luna Han and Jill Jurgensen at Taylor & Francis Group.

We thank Ian Proudler and Mario Bertero for kindly reading and providing constructive criticism of parts of the manuscript. Mistakes undoubtedly remain, for which the authors accept full responsibility. Finally, special thanks are due to ERP's wife, Pamela Pike, for much tea and sufferance over the years taken to put this book together.

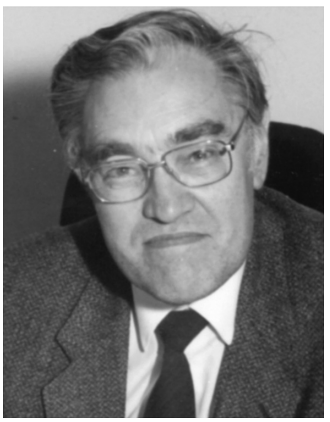
Geoffrey de Villiers
E. Roy Pike

Authors

(Photograph by Lisa Roberts Photography, Malvern, UK.)

Geoffrey de Villiers received a first class honours degree in physics from Durham University in 1980 and also the J. A. Chalmers Prize in physics. He earned a DPhil in theoretical physics from Oxford University in 1984. He joined the Royal Signals and Radar Establishment in Malvern in 1983, now part of QinetiQ Ltd. He left QinetiQ Ltd. in 2011 and is currently a research fellow in the Department of Physics and Astronomy and an honorary senior research fellow in the School of Electronic, Electrical and Systems Engineering at the University of Birmingham.

His specialism is linear inverse problems with particular emphasis on singular-function methods and resolution enhancement. He has worked on a wide variety of practical inverse problems in photon-correlation spectroscopy, radar, sonar, communications, seismology, antenna-array design, broadband array processing and computational imaging. His current research interests are in inverse problems in gravitational imaging and ionospheric physics.



Professor E. Roy Pike has first degrees in both mathematics and physics and a PhD in x-ray diffraction from University College, Cardiff. He won a Fulbright Scholarship to the Physics faculty at MIT, returning to the UK Royal Signals and Radar Establishment in Great Malvern, where he rose to chief scientific officer, with a visiting professorship of mathematics at Imperial College, London. He was appointed Clerk Maxwell Professor of Theoretical Physics at Kings College, London, later to become also head of its School of Physical Sciences and Engineering. His main research fields are theoretical physics, quantum optics (founding the journal *Quantum Optics*, now incorporated into *Journal of Physics B*) and the burgeoning mathematical discipline of inverse problems (founding the journal *Inverse Problems*, published by the UK Institute of Physics), now in its 31st year. He is interested in software and wrote the first draft of MathML for the World Wide Web. He was founder and first chairman of Stilo International plc. (Stilo.com), a web software company.

Dr. Pike's awards include the Royal Society Charles Parsons Prize, the McRobert Award of the Confederation of Engineering Institutions, the Annual Achievement Award of the Worshipful Company of Scientific Instrument Makers, the Civil Service Award to Inventors and the Faraday Medal of the Institute of Physics. He has published more than 300 papers, and authored, edited or

jointly edited 14 books. He has been hon. editor of *Journal of Physics A*, *Optica Acta and Quantum Optics* and board member of *Inverse Problems and Optics and Laser Technology*. He is currently a series editor for *Optics and Quantum Electronics* for Taylor & Francis. He is a fellow of the Royal Microscopical Society, a fellow and former vice-president of the Institute of Physics, a fellow of the Institute of Mathematics and Its Applications, a fellow of Cardiff University, a fellow of King's College and a fellow of the Royal Society.

Chapter 1

Early Concepts of Resolution

1.1 Introduction

1.1.1 Early History

It is not unreasonable to suppose that early humans were aware of the concept of resolution. As soon as they had acquired the power of speech, they would have realised that the ability to resolve visual detail at long distances differed significantly from one individual to the next. One might speculate that those with the keenest long sight would have been deployed in watching for prey or an imminent attack from a neighbouring tribe. The Romans considered short-sightedness a permanent defect that reduced the market value of a slave, relative to that of keen-sighted and literate ones, and older people would be read to by younger ones.

In the Islamic world, one finds quoted in a number of sources; see, for example [1] that Abbas Abu al-Qasim ibn Firnas ibn Wirdas al-Takurini (810–887 CE) of Islamic Spain developed a way to produce very clear glass and with it made 'reading stones' which were manufactured and sold throughout Spain for over two centuries. Reading stones were transparent planoconvex 'lenses' placed on a manuscript to magnify the script.

Ibn Firnas was an Andalusian polymath who has a crater 'Ibn Firnas' on the Moon named in his honour, as well as the Ibn Firnas Airport in northern Baghdad. It would seem that further Umayyadian research might throw light on these claims. Indeed, a green glass weight from the Umayyad Dynasty, dated as early as 743 CE, is owned by the Walters Art Museum in Baltimore, United States.

Reading stones have been made throughout the ages, particularly of rock crystal, a hard, clear, crystalline or of fused quartz, SiO_2, or of beryl, beryllium aluminium cyclosilicate, from which the German word for spectacles, Brillen, is derived. Such lenses have also been used, over a long period of history, to focus the rays of the sun to make fire.

We would like to make it clear that figured transparent objects which can be used for the eye lenses of statues, magnification of script, decorative jewellery, making fire and later in history for microscopes, telescopes, ophthalmic correction, etc., are all called 'lenses' in this book. This should not be taken as any disrespect to any of these subfields and we use the qualifications reading stones, burning glasses, magnifiers, spectacles, etc., when being specific. As we shall explain later, a 'lens' is supposed to look like a transparent lentil and only in microscopy is that sometimes the case.

The history of ancient crystal and glass lenses in the middle-eastern Egyptian, Greek and Roman worlds has been thoroughly researched by specialist scholars in modern times. In the following, we try to summarise many of these researches which are scattered throughout the literature.

Rock-crystal lenses date back to over 4500 years ago in the early-kingdom period of Egypt of the IV/V dynasties. Skills developed for working very hard stone, such as granite, obsidian and flint requiring fine abrasive materials, were adapted for cutting and polishing hard quartz.

These planoconvex lenses were used in funerary (Ka) statuary to represent realistic eyes. With a smaller, highly concave section ground and polished in the centre of the plane back face, they have the eerie optical property of following the observer around when viewed from different angles.

Examples of such lens-eyed statues, which were well-preserved hidden in the masonry of mastaba burial chambers, have been excavated and can be seen today: Le Scribe Accroupi in the Louvre museum in Paris, the priest Kapunesut Kai with his son and daughter in the Egyptian museum in Cairo and, in the same museum, the IV-dynasty statues of the Prince Rahotep and his wife*; see, for example, Enoch [2].

Dr. Enoch also discusses the discovery of a planoconvex rock-crystal lens in the right eye of a celebratory carved-stone libation vessel (a rhyton), in the form of a bull's head, dated about 1500 BCE, found at the Little Palace of Knossos, now at the Heraklion Archaeological Museum in Crete. The back-plane face of the lens has a delicately painted face, intended to be viewed, magnified, through the lens.

In another contribution, Enoch [3] states that 'It is generally agreed that the first lenses had their origin in the Near East or Eastern Mediterranean basin area. The earliest lens that will be considered in this article originated in Crete during the Minoan era, roughly 3,500 years ago'.

He also states [4] that about 400 BCE, magnifying lenses appeared in Greece, and later in Rome. They were used in jewellery and in decorative settings.

Enoch's professional work, of course, encompassed all lenses in the general sense we adumbrated earlier.

Aristophanes, 419 BCE [5] also says

> 'Have you ever seen a beautiful, transparent stone at the druggist's with which you may kindle fire?'

There is an interesting possibility that in the third century BCE Archimedes (287–c. 212 BCE) used such a lens for magnification as a monocle wired to his head [6]. At that time, this would almost certainly have been of rock crystal.

The oldest piece of glass which can be positively dated was made somewhere in Mesopotamia in the third millennium BCE, but it was the Romans who began to use glass for architectural purposes, with the discovery of clear glass (through the introduction of manganese oxide) in Alexandria around 100 CE [7]. We relate Seneca the Younger's remarks on water-filled glass lenses in Hellenistic Roman times in Appendix A.

Emperor Nero (37–68 CE) famously spent a small fortune bringing mosaic and coloured glass to Rome from Alexandria, a Roman territory at that time. The invention of mosaic glass and of glass blowing† which allowed glass vessels to be made efficiently took place in important Egyptian glass factories in the Ptolemaic Roman periods, 332 BCE–395 CE. Glass furnaces were located near the raw-material sites in Wadi Natrun and on the shores of Lake Mariout near Alexandria.

The Catholic 'White Fathers' excavating in Carthage from 1897 found a pair of round spherical 'lenses' in a sarcophagus from the fourth century BCE [8]. They were held in the Museum Lavigerie, which was renamed in 1956 to Carthage National Museum.

A Roman glass lens of this period found at the Akhmim site in upper Egypt is on display at the Ashmolean museum in Oxford. This lens has been described recently by Jane Draycott [9] together with references to and records of 10 glass lenses having been recovered through archaeological excavation from sites throughout Egypt: 2 from the San-el-Hagar site called Tanis, now in the British museum, 3 from Hawara and 4 from Karanis (Kom Ushin).

The Karanis lenses are on display in the Kelsey Archaeological Museum at the University of Michigan where they are classified as 'writing implements'. Francis Kelsey, a classics professor at the University of Michigan at the turn of the last century, was responsible for the excavations and the impressive collection of antiquities in this museum [10].

Heinrich Schliemann recovered some 50 bronze-age, planoconvex, rock-crystal lenses during his excavations at Homeric Troy [11]. Schliemann's colleague, Dörpfeld, who dated the objects

* Beautiful colour images of all these works can be found by using their names and locations in a simple web search.

† Glass-blowing was invented by the Phoenicians between 400 and 50 BCE.

to ca. 2200 BCE, accepted that the largest among them were used as magnifying glasses. Sines and Sakellarakis in 1987 [12] wrote that 'there are now 23 ancient lenses on display in the Archaeological Museum in Heraklion and many more in storage there'.

Pliny the Elder in his *Natural History* of 79 CE [13] writes of a practical use for crystal lenses:

> I find it stated in medical authors that crystal balls placed opposite to solar rays are the most useful contrivance for cauterizing the human body.

In the same work (of 37 books in 10 volumes), one finds 111 appearances of the word 'glass' and 916 of the word 'eye' but no confluence of the two. However, among many other suggestions, one also finds that

> The heads and tails of mice, reduced to ashes and applied to the eyes, improve the sight.

Perhaps the most famous ancient lens specimen extant, claimed to be the oldest known lens in the world* [14], is the planoconcave 'Nimrud' rock-crystal lens in the British Museum. This was discovered in the throne room of King Sargon II's Assyrian palace of Nimrud (Kouyunjik-mound ruins, Nineveh, near present-day Mosul, Iraq) in 1850 by Sir Austen H. Layard. It dates back to the seventh century BCE, at which time Nineveh is said to have been the largest city in the world. This lens is slightly oval (40 mm length, 35 mm breadth) with a thickness of 23 mm and a focal length of 114 mm. Layard himself [15] stated that

> its properties could scarcely have been unknown to the Assyrians, and we have consequently the earliest specimen of a magnifying and burning glass.

He also added in a footnote that Sir David Brewster, FRS, had examined the lens and noted that

> its convex surface had been fashioned on a lapidary's wheel, or by some method equally rude and it gives a tolerably distinct focus at a distance of $4\frac{1}{2}$ inches from the plane side.

Sir David's conclusion was that its most likely purpose was magnification. In an article on the lens [16], Sir David again concluded by assigning reasons why it should not be looked on as an ornament, but rather as a true optical lens. The British Optical Association's W.B. Barker investigated this from an optician's perspective in 1930 and concluded that the shape, small size (1.6×1.4 in.) and level of workmanship (probably ground on a flat lapidary's wheel) dictated against this being an ornamental item or a mere burning glass. He suggested that the lens would neatly cover the human orbital aperture and produce magnification suitable for near work.

In a Presidential Address before the Optical Society of America, in Rochester, NY, 24 October 1921, on the occasion of the Helmholtz Memorial Celebration, James Southall, who at the time was President of the Optical Society of America, stated:

> It is possible that magnifying glasses were used by the Chaldeans about six thousand years ago.† The cuneiform characters on the tablets found by Layard in the ruins of Nineveh which are now in the British Museum are singularly sharp and well-defined, but so minute in some instances as to be illegible to the naked eye. Specimens of the very implements used to trace these inscriptions were found in the ruins and curiously enough glass lenses were found also. . ..‡

However, the British Museum's curator's notes, of 2012, cast doubt on Layard's statement, Sir David Brewster and Barkers' opinions and Southall's conjectures, regarding its use as a magnifier.

---

* But note the Minoan lens dated 1500 BCE and the Trojan lenses of 2200 BCE described in the previous text.

† The Chaldeans captured the Assyrian capital of Nineveh in 612 BCE. It seems that Southall really meant 600 BCE. Also the Layard lens, in fact, was of quartz not glass.

‡ From Southall [17]. With permission.

These notes hypothesise that, although it undoubtedly has optical properties, these are probably accidental and that it is much more likely that this is a piece of inlay, perhaps for furniture.

Two archaic Greek (800–480 BCE) rock-crystal lenses of shorter focus and much better quality than the Nimrud lens were found by Professor Yannis Sakellarakis in a sacred cave on Mount Ida in Crete in 1983. Much older lenses were found by Sir Arthur Evans, FRS, during his excavation of the palace at Knossos and the neighbouring Mavro Spelio cemetery, dating from ca. 1400 BCE [12,18].

Sines and Sakellarakis present as evidence for the use of lenses for magnifying purposes:

> the fine detail of Roman gold-glass portrait medallions, the discovery of a glass lens in the house of a Roman engraver in Pompeii and another pair of glass lenses in the house of an artist in Tanis.*

They also claim that further support for this thesis would be the difficulty in making seals and coin dies with the unaided eye. Natter, who was a skilled eighteenth century gem carver, argued in 1754 [19] that the best examples of carved gems by Greek artists would have required magnification. The pioneering eighteenth century German archaeologist and writer Johann Joseph Winckelmann in his History of Art (1776) [20] also drew the same conclusion from the most minute carving of ancient Greek gems. Excavations of predynastic carved ivory knife handles dating from some 5000 years ago also raise the same questions.

There are many museum pieces of the following centuries of such concentrators for burning or even for decoration.

Another set of lenses which have given rise to much speculation are the Visby lenses found in Viking graves on the island of Gotland, Sweden [21]. They were investigated scientifically by a small team in 1997 [22]. These are made of rock crystal and dated from the eleventh or twelfth century CE. Some of them have a silver mounting and may have been used as pendants, though it is believed that the lenses are much older than the mountings; others are unmounted and show no signs of use as jewellery. Their place of manufacture is uncertain, though it could have been Byzantium. Somehow, these lenses were made with the ideal elliptical focussing shape 500 years before Descartes showed how to calculate this mathematically. They are bi-aspheric lenses of sufficiently high quality that it has even been suggested that they could have been used in a telescope, again five centuries before the accepted date of invention of such instruments. They were more likely to have been reading stones, that is, they were placed over manuscripts to magnify the letters. They could also have been burning glasses for igniting fires.

Improvement or clearing of the vision of the human eye using various eye salves or collyria has occurred over the ages but the first use of lenses for spectacles to correct defective vision and hence improve resolution is still a subject of debate. Although reading glasses are thought to be one of the most important inventions of the past 2000 years [23], we have still to agree with Vasco Ronchi [24] who wrote in 1946:

> Much has been written, ranging from the valuable to the worthless, about the invention of eyeglasses; but when it is all summed up, the fact remains that the world has found lenses on its nose without knowing whom to thank.

The story continues to the present day with a number of specialised books devoted to this topic, for example, Rosenthal [25], Ilardi [26], Rosen [27] and Willach [28].

For interested readers, we have outlined briefly some of the complicated history of this question in Appendix A. Whatever the origin of spectacles, we can say with some certainty that lenses have been around since the ancient Egyptians of 2500 BCE, the ‘Golden Age’ of the old kingdom, *vide supra*. As we discuss in the appendix, they might have been invented in China many centuries BCE although this is disputed, used in monocle form by Archimedes in the third century BCE

* These latter, housed now in the British Museum, were excavated ca. 1885 by the British Egyptologist Sir William Flinders Petrie, FRS. They are dated ca. 147 CE.

(as mentioned earlier), discovered in the eleventh century in the Arab world (see the poem of Ibn Hamdis which we quote in this appendix) or in the thirteenth century in Germany, long suspected* (see the black-letter Gothic attribution to around 1270 in Figure A.2), or as has been conventionally believed by a number of scholars, as late as the end of the thirteenth century in northern Italy, also described in Appendix A.

As we also review in this appendix, Roger Bacon, in 1268, following the writings of the eleventh-century *Book of Optics* [29] by the Arab polymath Ibn al-Haytham (known in the West as Alhazen) also described in detail how, in planoconvex form, they were used as 'reading stones', placed as magnifiers with the plane side onto a manuscript. The earlier Spanish work described above was unknown to Alhazen and Bacon.

Devices such as spectacles are designed to correct for defective vision. Other optical instruments are designed to image objects at scales for which our eyes are not suited in order to satisfy our curiosity about the natural world.

The word 'lens' itself is Latin for the genus of the lentil plant and it has been hypothesised by Michael Hutley [30] that it was adopted in the seventeenth century when Robert Hooke and Antonie van Leeuwenhoek made small hemispheres of glass which resembled the lentil seed for their microscope objectives.

Biconvex microscope and telescope objectives became widely used, and although perhaps not always as small as a lentil seed, the name was eventually applied to all optical imaging objectives and oculars.

Keen sight no longer held its former importance. New unimaginable worlds were revealed to the inventors and users of microscopes and telescopes which aided human vision and the comprehension of the universe we inhabit. From van Leeuwenhoek, who first observed micro-organisms in Delft lake water† and Galileo, who discovered that the Milky Way was not just a nebulous cloud but made up of individual stars, the importance to the human race of this extraordinary historical explosion in optical resolving power cannot be overestimated.‡

A photograph of an actual van Leeuwenhoek microscope owned by the Royal Society in London is shown in Figure 1.1 and is used in our cover jacket. The specimen is placed on the end of the sharp

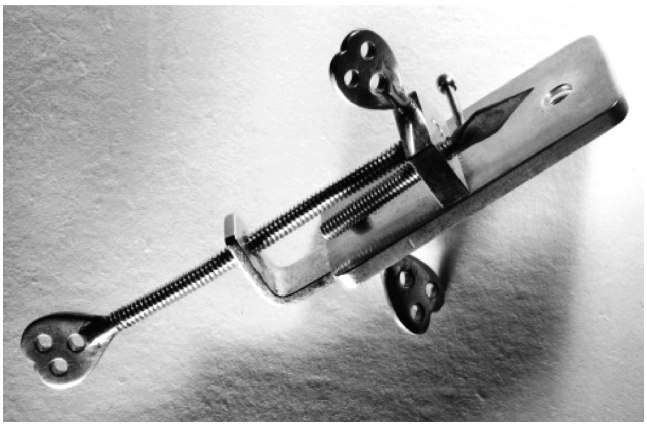


FIGURE 1.1: A van Leeuwenhoek microscope. (Courtesy of the Royal Society, London, UK.)

* See the reference to Carl Barck in Appendix A.

† For his pioneering work, van Leeuwenhoek was elected to a Fellowship of the Royal Society of London in 1680.

‡ For a modern, eloquent and scholarly review of the significance of these advances, combined with the Copernican heliocentric view of the solar system, see Scharf [31].

tip and the hand-polished lens is sandwiched between the two brass bassplates. The three screws adjust the specimen position and the focus.

We have discussed the simple microscopes of Hooke and van Leeuwenhoek. The end of the sixteenth century saw the creation of a compound optical microscope, possibly by the Dutch spectacle maker Zacharias Janssen with the aid of his father Hans. This, however, has been disputed and other spectacle makers, Hans Lippershey, in the same town, Middelburg, who filed a patent in 1608 which was not granted and Jacob Metius of Alkmaar, an instrument maker who filed a patent two weeks later, also not granted, may also be credited with the invention. An interesting history can be found in King [32].

Since then, the microscope has been developed to such an extent that, in order to obtain finer and finer resolution, many modern microscopes do not use light at all for imaging. Electron microscopy, acoustic microscopy, scanning tunnelling microscopy and atomic force microscopy are all examples falling into this category. Fluorescence microscopy is an example of advanced optical microscopy. With each step change in resolving power come new scientific discoveries.

The ultimate limits in the quest for resolving smaller and smaller detail are explored in particle colliders, where the structure within nucleons can be probed. In the parton model [33], as the momentum transfer in the collision increases, finer detail is seen in the sense that more partons, which today are thought of as quarks, anti-quarks and gluons, are seen.

There is also a controversy around the invention of the refracting telescope. Lippershey applied for a Belgian patent for it in 1608 but two other claimants, Janssen and James Mettius, came forward, and the patent was not granted.

The reflecting telescope was invented by Isaac Newton ca. 1670. The telescope has also evolved considerably from its early days. Examples which use different parts of the electromagnetic spectrum from visible light are radio, infrared and x-ray and γ-ray telescopes. Resolution is the key performance measure for all these optical instruments. It is also critically important for imaging devices such as radars, sonars and seismic arrays.

In the latter applications, it is possible to track the phase of the waves, whereas in the visible region of the spectrum, due to the short wavelength, this was not possible until the advent of the laser in 1960. After this time, the so-called ‘coherent’ imaging at laser frequencies became possible by mixing the optical image with a reference beam of the same frequency (homodyning) or heterodyning with a frequency-shifted beam.

The concept of resolution is, however, much more widely applicable. The subject of telegraphy, that is, the sending of messages over long distances, has developed in parallel with optics and, as we shall see, there are many parallel concepts in the two fields. Early telegraphy systems involved such ideas as signal beacons, heliographs and smoke signals. Among later inventions were naval flags and semaphore.

Nowadays, telegraphy has come to mean electrical telegraphy, whereby electrical signals are transmitted down a communication channel. Until recently, this channel was typically the air or a metal wire. Nowadays, optical fibre is commonly used. The origins of electrical telegraphy are discussed in detail in Munro [34] from whom we quote the following:

> The first suggestion of an electric telegraph on record is that published by one ‘C. M.’ in the *Scots Magazine* for February 17, 1753. The device consisted in running a number of insulated wires between two places, one for each letter of the alphabet. The wires were to be charged with electricity from a machine one at a time, according to the letter it represented. At its far end the charged wire was to attract a disc of paper marked with the corresponding letter, and so the message would be spelt. ‘C. M.’ also suggested the first acoustic telegraph, for he proposed to have a set of bells instead of the letters, each of a different tone, and to be struck by the spark from its charged wire.

The identity of 'C. M.', who was probably Charles Morrison of Greenock, and many other interesting schemes are discussed in the subsequent pages of Munro's book, which can be easily read at the URL cited in our reference list.*

Other notable early work is a 1795 paper by Salvá i Campillo [35], 'Report on electricity applied to telegraphy', on an electrostatic telegraph involving the use of Leyden jars. A second report by the same author followed in 1804, 'Second report on galvanism as applied to telegraphy', [36]. This report discussed the use of Voltaic piles. An early electrochemical telegraph was demonstrated by the German inventor Samuel Thomas von Sömmering in 1809, based on [36].

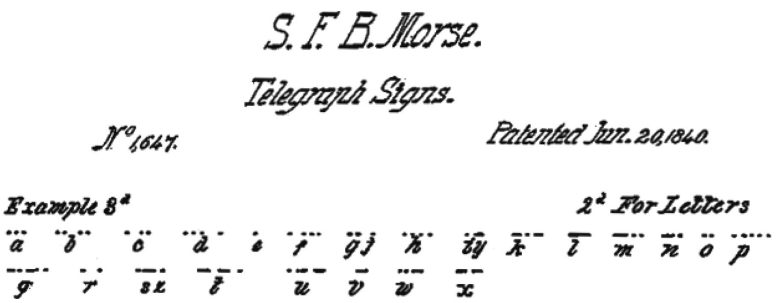
According to Munro, designs of electrostatic telegraphs by Ralph Wedgwood in 1806 and Sir Francis Ronalds in 1816 were submitted to the British Admiralty but were rejected on the grounds that semaphore was sufficient for the country. In fact, Ronalds built his telegraph in the garden of his house in Hammersmith, London, and a commemorative plaque was later raised there by the Institution of Electrical Engineers, see the article by his great-great-great niece, [37].

Munro's book also includes the description of more realistic studies following Oersted:

> 'In 1820 the separate courses of electric and magnetic science were united by the connecting discovery of Oersted, who found that a wire conveying a current had the power of moving a compass-needle to one side or the other according to the direction of the current. Laplace, the illustrious mathematician, at once saw that this fact could be utilised as a telegraph, and Ampère, acting on his suggestion, published a feasible plan. Before the year was out, Schweigger, of Halle, multiplied the influence of the current on the needle by coiling the wire about it. Ten years later, Ritchie improved on Ampère's method, and exhibited a model at the Royal Institution, London. About the same time, Baron Pawel Schilling, a Russian nobleman, still further modified it, and the Emperor Nicholas decreed the erection of a line from Cronstadt to St. Petersburg, with a cable in the Gulf of Finland but Schilling died in 1837, and the project was never realised.'

Gauss and Weber are credited with the first practical system used in Göttingen in 1833. Cooke and Wheatstone constructed the world's first commercial telegraph in England in 1837. Their first experimental line ran between Euston and Camden Town railway stations and this was followed by a commercial line between Paddington and West Drayton stations. Independently, in America, Morse was responsible for the first single-wire telegraph, patented in 1840, as well as the code which bears his name, which became the standard code for telegraphy. In Figure 1.2, we show a portion of the patent of 1849 for the original Morse code.

In telegraphy, the concept of resolution is strongly related to channel capacity, that is, the capacity of the channel to transmit information. Consider, for example, the information as encoded in a sequence


S. F. B. Morse.

Telegraph Signs.

No 1,647. Patented Jun. 20, 1840.

Example 8d 2d For Letters

a b c d e f g j h i k l m n o p

q r s t u v w x

FIGURE 1.2: Morse-code patent.

* This ebook is for the use of anyone anywhere at no cost and with almost no restrictions whatsoever. You may copy it, give it away or reuse it under the terms of the Project Gutenberg License included with this ebook or online at www.gutenberg.org.

of pulses. In a band-limited communication channel, if the pulses are too closely spaced, then they will tend to merge with each other, leading to problems with decoding the information carried.

As pointed out by Elias et al. [38], there are strong parallels between the theory of electrical communication channels and optical imaging, and these will form the basis for notions of resolution upon which we will concentrate in this book. In particular, an analogy is made in these papers between the treatment in Fourier space of the convolution of an electrical signal with an integrating memory function by multiplication of their two temporal Fourier spectra, and that for an optical image which is convolved with a point-spread function in two spatial dimensions, that is, with products of wave-number spectra rather than of temporal spectra.

1.1.2 A Human Perspective on Resolution

Each person's mental picture of resolution is based, to a greater or lesser extent, on experience of their own visual system. It is therefore worth considering this further to see what insights we may gain from it.

Real-world scenes tend to consist of regions with more or less uniform texture divided by discontinuities which we call edges. There are also typically pointlike features within a scene. However, we do not see sharp edges and points as such; we always see blurred versions of them. As far as our eyes are concerned, the stars, with one notable exception, should appear as points. A simple test on a clear night will demonstrate that, in fact, they appear as small patches of light. What we see when looking at a point of light is known as the point-spread function or impulse response of our visual system.

Because of this blurring, we can never be exactly sure what we are looking at. Our inability to resolve detail beyond a certain scale means that an infinite number of possible scenes will give rise to the same sensation within our brains. Although as scientists we tend to believe in an objective reality, our senses do not allow us to determine this unambiguously.

In practice, we use prior information about what we are looking at to try to improve this situation. Looking at an optician's chart or a motorway sign, we can tell ourselves that the symbols come from a finite alphabet and this can help to decide what the symbols are. Similarly, if we know that the scene consists of one or two closely spaced point sources of equal magnitude, we can use this information to try to decide between the two possibilities. This problem, known as two-point resolution, has been studied by many people with a view to quantifying the degree of resolution achievable with a given optical instrument. An early two-point resolution criterion, that due to Rayleigh [39] was aimed at specifying the minimum angular distance between two-point sources in order for them to be resolved by the human eye. It should be apparent to the reader that this notion of resolution is rather restrictive and does not sit well with our everyday experience. However, there are connections between two-point resolution and the main concept of resolution we will use in this book, and these will be discussed later on.

Returning to our visual system, the limit of resolution is determined by a number of factors. The detector mechanism consists of a finite number of rods and cones, the cones being responsible for daytime vision and the rods responding to low light levels. In the central fovea (see Figure 1.3), there are no rods but a higher density of cones than in the rest of the retina. The resolving power of the eye in daylight conditions is greatest in the fovea. The resolution is inherently limited by the aperture of the eye, controlled according to the ambient light by the size of the iris. Another factor is imperfections in the accommodation process leading to long- and short-sightedness. Aberrations such as astigmatism also affect the performance of the eye.

So, to summarise what we have learnt about resolution from the human visual system, we can say that the image we see is a blurred version of reality due to a number of physical reasons. There are an infinite number of possible scenes which would give rise to the same image in our heads, and prior information can be helpful in restricting this set. We have also seen that, in common with other problems where resolution is involved, the data are sampled by a finite set of detectors.

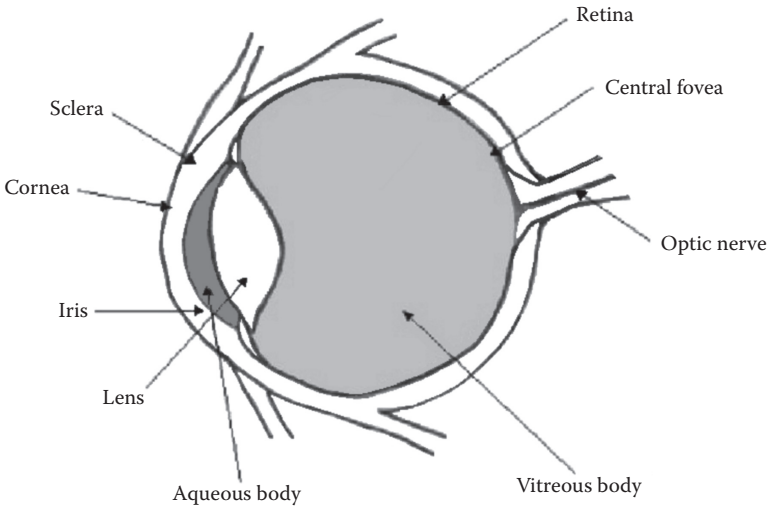


FIGURE 1.3: The human eye.

1.1.3 Pinhole Camera

Though we have introduced some of the basic ideas of resolution in our brief discussion of the human visual system, there are a number of complicating factors which make this system difficult to analyse, not the least of which is the non-linear response of the rods and cones to the intensity of the incoming light. As a consequence, and to direct attention to the design of physical instrumentation, we will look at simpler systems where the detector responds, to a good approximation, in a linear manner. In fact, in this book, we will mainly consider such systems since the primary notion of resolution which we wish to study depends upon this linearity.

A simpler imaging device than the human eye, which also gives insight into the problem of resolution, is the pinhole camera, well known to generations of school children. In its larger, room-sized, version, it is known as the camera obscura, from the Latin for darkened room, the origin of the modern term 'camera'. The name was coined in the early seventeenth century by Johannes Kepler in his book *Ad Vitellionem paralipomena, quibus astronomiae pars optica traditur* of 1604*, in which he described the inverse-square law governing the intensity of light, reflection by flat and curved mirrors and the principles of pinhole cameras.

The experimental origins of the camera obscura are lost in antiquity. There is written evidence that it was known by the Chinese philosopher Mo Zi (470–391 BCE), who most probably was the first to record, in the *Book of Mo Zi* (*Mo Ching*), the formation of an inverted image with a pinhole and screen. His descriptions (he called the chamber a 'locked treasure room.') can be found in volume four of Needham's book *Science and Civilization in China* [7].

Naturally occurring rudimentary pinhole cameras were mentioned by Aristotle (384–322 BCE) and Euclid (330–260 BCE). They wrote, for example, of light travelling through the slits of wicker baskets and the crossing of leaves to propose the rectilinear propagation of light.

In the book *Miscellaneous Morsels from Youyang* written in about 840 CE by Duan Chengshi (d. 863) during the Tang Dynasty (618–907), the author mentioned inverting the image of a Chinese pagoda tower beside a seashore. His explanation (that it was something to do with the sea) was corrected some two centuries later by the Song-Dynasty scientist Shen Kuo (1031–1095) who applied the correct interpretations in his book *The Dream Pool Essays* of 1088 CE. Yet, again, in the tenth century CE, another Chinese scientist, Yu Chao-Lung, used model pagodas to make pinhole images

* An English translation by William Donahue in 2000 is available: ISBN: 1-888009-12-8.

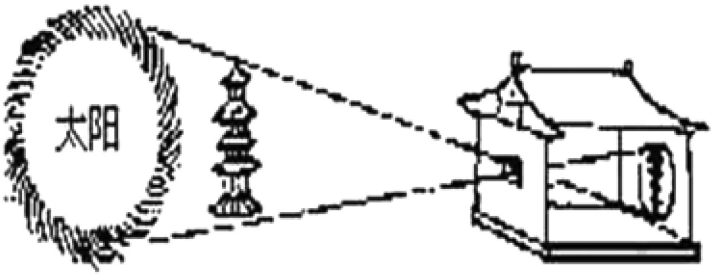


FIGURE 1.4: The pinhole camera.

on a screen. Needham [7], p. 98, states that 'inverted pagodas were being looked at (in China) at least as early as about +840'.

A depiction of a Chinese camera obscura, again imaging the canonical pagoda, is shown in Figure 1.4. This is from the book *Jing jing ling chi* (*Optical and Other Comments*) by Zheng Fu-Guang zhu (1780–1853) and has been used extensively since. For example, this figure was used by Wu et al. [40] in 2015. The Chinese characters translate as "sun."

The property of image inversion was noted by Al-Kindi in the sixth century CE and was again described in the eleventh century CE by Abū Ali al-Hasan ibn al-Haytham, latinised as Alhazen [29], and also three centuries later by his ardent disciple Kamal al-Din Hasan ibn Ali ibn Hasan al-Farisi. According to the historian Max Hertzberger [41] the first known unequivocal description of the camera obscura was by Alhazen. Alhazen overturned the euclidean theory of vision, in which the eye emitted light, in favour of the Aristotelian view of it entering from the outside. He also describes observations of the eclipse of the sun and noted that the smaller the pinhole, the clearer the picture, as did al-Farisi.

In the thirteenth century, the camera obscura was recommended for the safe observation of solar eclipses by Roger Bacon and by Leonardo da Vinci in his Codex Atlanticus written between 1478 and 1519.

Nowadays, camerae obscurae tend to have lenses, an innovation of Giambattista della Porta (1545–1615) (also known as John Baptista Porta) described with both concave and convex lenses in his book, *Magiae Naturalis, Libri XX*, published in 1599, and hence are no longer, strictly speaking, pinhole cameras.

The resolution of the pinhole camera, as judged by the human eye, depends upon the size of the aperture. If this is too large no image is formed. On reducing the aperture, the image is governed by geometrical optics and the resolution improves as the aperture gets smaller. As the aperture size is reduced, further diffraction takes over and the resolution then degrades with decreasing aperture. Petzval [42,43] was the first to attempt to determine the optimal pinhole size. Rayleigh [44] improved on his results and showed, using Huygens–Fresnel diffraction, that the best results are obtained when the aperture, as seen from the image plane, has a diameter of 0.9 times that of the first Fresnel zone. To be more precise, if r is the radius of the aperture, then

$$r^2 \frac{a+b}{ab} = 0.9\lambda,$$

where

λ is the wavelength of the radiation
a is the distance of the object to the pinhole
b is the distance of the pinhole from the image plane

The pinhole camera is discussed in more modern times by, for example, Hardy and Perrin [45], Wood [46], Goodman [47] and Sharma [48].

FIGURE 1.5: Photograph taken with a camera obscura at King's College London.

Figure 1.5 is a photograph taken with a camera obscura at King's College London*. The view is over the south bank of the River Thames. The scene was sunlit and an aperture of 5 mm diameter was used with a white image screen at 1 m distance. With less daylight, acceptable pictures could still be obtained with lensless apertures as large as 1 cm.

1.1.4 Coherent and Incoherent Imaging

In this section, we look at two fundamental forms of optical imaging, namely, coherent and incoherent imaging. Coherent imaging involves light from all parts of the illuminated object having the same phase relationships as time varies. In other words, the phase differences between parts of the object are independent of time. A typical scenario would be a semi-transparent object illuminated from behind by laser light. The amplitude and phase of the light would vary across the object but these variations would be constant in time. We will make the assumption when talking about coherent imaging that the detector responds linearly to the amplitude of the received waveform, rather than the intensity. This can be accomplished by optical homodyning or heterodyning.

In incoherent imaging, the phase relationships between light rays from different parts of the object vary with time. The object is then represented by an intensity distribution. We will make the assumption in this case that the detector responds linearly to the intensity of the received waveform.

Partially coherent imaging lies, as its name suggests, between coherent and incoherent imaging. It corresponds to the spatial coherence length being of comparable size to the width of the central lobe of the point-spread function. However, since the problem is neither entirely linear in intensity nor amplitude, it is easier to treat resolution in partially coherent imaging as a two-point resolution problem.

* The picture was taken by James French and set up by Luke Nicholls, Paco Rodriguez-Fortuno, Will Wardley and Diane Roth all of the Physics Department at King's College, London, UK.

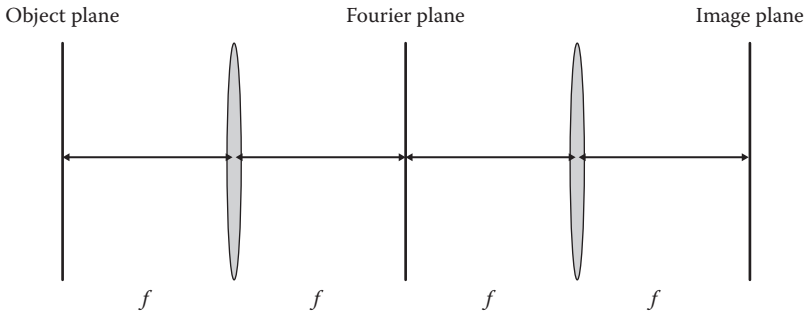


FIGURE 1.6: $4f$ system: object, pupil and image planes.

Consider first the case of coherent imaging through a circular aperture. The essence of this is described by the $4f$ system depicted in Figure 1.6, where the two lenses are identical, with focal length f. Assume we have a semi-transparent object, illuminated from behind by a laser, placed in the front focal plane of the first lens. Assume also that this object affects the amplitude and phase of light passing through it in a time-independent manner. For a circular aperture of radius a and a single point source, one can write the amplitude pattern at the detector location (i.e. the point-spread function or impulse response) as

$$\psi(\zeta) = 2\pi a^2 \frac{J_1(2\pi a\zeta/\lambda)}{2\pi a\zeta/\lambda}, \tag{1.1}$$

where

ζ is the angle of deviation from the axis in the diffraction pattern
J_1 is a Bessel function of the first kind

The function ψ is the well-known Airy pattern. For incoherent imaging, the corresponding point-spread function is given by ψ^2.

It is sometimes necessary to consider the image of a line, rather than a point. For incoherent imaging with a circular pupil, the resulting function is a Struve function of the first order (see Williams and Becklund [49]).

Now, for simplicity, let us consider 1D coherent imaging and look at the physics behind the point-spread function. This is essentially a double Fraunhofer diffraction process. Hence, we are working within the paraxial approximation of Huygens–Fresnel diffraction, as well as the additional assumptions of Fraunhofer diffraction. Assume we have the same experimental set-up as that in Figure 1.6, except that the lenses are now cylindrical. The effect of the lenses is to bring the Fraunhofer patterns to focus at finite distances. The object, also assumed to be constant in the third dimension, generates a Fraunhofer diffraction pattern in the common Fourier plane of the two lenses. This pattern, truncated by a stop in the Fourier plane to reflect the band-limit of the system, then undergoes a further Fraunhofer diffraction to generate an image in the image plane. The stop cuts out all the angular spatial frequencies above a certain maximum Ω. The image formed in the back focal plane of the second lens then has the form

$$g(y) = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} e^{i\omega y}F(\omega)d\omega = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} e^{i\omega y}\left[\int_{-\infty}^{\infty} f(x)e^{i\omega x}dx\right]d\omega,$$

$$= \int_{-\infty}^{\infty} \frac{\sin\Omega\,(y-x)}{\pi(y-x)} f(x)dx. \tag{1.2}$$

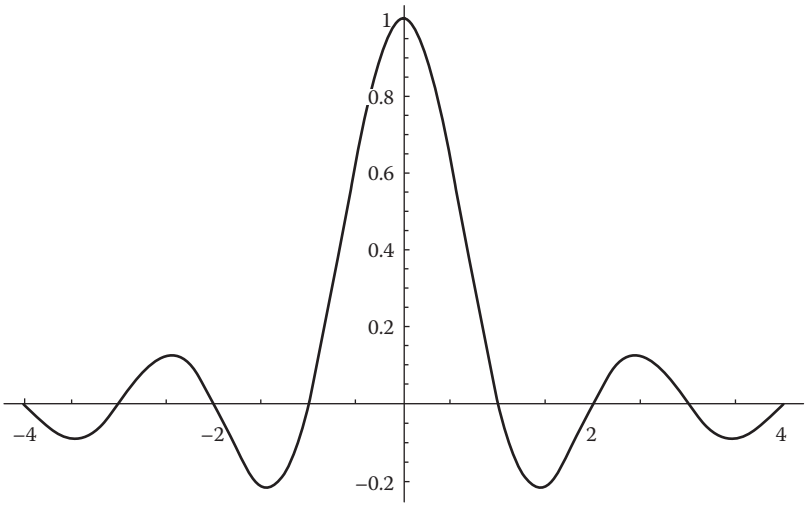


FIGURE 1.7: The sinc function.

The first factor within the integral in (1.2) is the point-spread function. We will see that there is a strong parallel between (1.2) and an equation arising in band-limited communication theory, which we will develop further later.

The kernel in (1.2) is related, by a simple transformation of variables, to the function sinc(x) given by

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

This function is plotted in Figure 1.7.

It is well known that this function occurs frequently in sampling theory and the theory of the cardinal series. This has led people to speculate and even to assert that Woodward, who coined the term in 1952 [50], added the 'c' in sinc as a shortened form of the Latin 'sinus cardinalis'. This is not correct*.

To set the record straight, we have the following statement [51] from Woodward himself, written for us to include here†:

> I very clearly remember finding myself in the impasse of how to say 'the (sin *pi x*)/(*pi x*) function'. For a start, it is not a function but an expression. Had I known it at the time, Church's lambda notation would have come to the rescue with '$\lambda x\cdot$(sin *pi x*)/(*pi x*)', but this solution would have been impractical for more than one reason. I needed a new function name, and decided to make one up for myself. It ought to start sin and be modified with an extra letter. I can still remember going through the alphabet to find one that 'felt' right and was pronounceable. I chose the letter 'c' because it made me think of a cosine, which, like sinc, is an even function and has value 1 at argument 0. Nothing whatever to do with cardinal!

* The complete statement in this paper with I.L. Davies is 'where sinc is an abbreviation for the function (sin πx)/πx. This function occurs so often in Fourier analysis and its applications that it does seem to merit some notation of its own'.

† One of the authors of this book, ERP, was fortunate enough to have Philip Woodward as his supervisor in his first job at the Royal Radar Establishment in the United Kingdom.

In the case of 1D incoherent imaging, the intensity distributions of the object, f_I, and image, g_I, are related by

$$g_I(y) = \int_{-\infty}^{\infty} \frac{\sin^2 \Omega (y-x)}{\pi^2 (y-x)^2} f_I(x) dx. \tag{1.3}$$

For the aforementioned examples, the image intensity pattern of a single point source consists of a central main lobe together with sidelobes. The Rayleigh resolution criterion for incoherent imaging then states that the two point sources are just resolved when the central maximum of the image intensity pattern of one source is situated over the first minimum of the image intensity pattern of the other source. Instead of two-point sources in an image, the same resolution criteria can be applied, for example, to the separation of spectral lines in a spectrometer.

In an image, if the two sources are coherent, the resolution criterion needs careful consideration. In the case of fully coherent sources, the image amplitude pattern can be written as the sum of the individual image amplitude patterns. The image intensity pattern will then be the modulus squared of the total image amplitude pattern. The image intensity pattern will thus depend on the phase relationship between the sources. For further details, see Lipson and Lipson [52].

The Rayleigh resolution criterion for the 1D incoherent imaging problem specifies that two points are resolvable if the angular distance between them is greater than the distance between the central peak of $\sin^2 \Omega (x-y)/\pi^2 (x-y)^2$ and its first zero. Let this distance be d. Then

$$\Omega d = \pi,$$

so that

$$d = \frac{\pi}{\Omega}. \tag{1.4}$$

It should be noted that the Rayleigh resolution criterion is purely a rule of thumb based on perception by the human visual system. This was stated clearly by Rayleigh [53]:

> This rule is convenient on account of its simplicity and it is sufficiently accurate in view of the necessary uncertainty as to what exactly is meant by resolution.

It is curious that a large literature should have sprung up involving claims to do better than the Rayleigh limit, when it was never claimed in the first place that this was a hard limit.

1.1.5 Abbe Theory of the Coherently Illuminated Microscope

As a prelude to a fuller discussion of resolution, we will discuss the Abbe* theory of resolution since, historically, this was one of the first ways of quantifying resolution and has persisted since (see [54–58]). Abbe submitted his 1881 paper and his two-part 1882 paper in English (he was an Honorary Fellow of the Royal Microscopic Society). The famed 1873 paper comprised 63 pages in German without a single equation or diagram. Nevertheless, it was perfectly adequate to establish his diffraction limit, as we shall discuss in the following text.

The theory applies to microscopes where the illumination is monochromatic and a semi-transparent object is illuminated from behind. The coherence in the illumination is achieved using a pinhole source and a variable condenser aperture. Abbe was led to consider imaging optical gratings

* We just note here that Abbe's name is quite frequently mispronounced; the 'e' is not acute but pronounced as 'uh'.

of varying spatial period, d (spatial frequency $\frac{1}{d}$). He considered in the first case illumination by a normally incident coherent beam generating forward diffraction orders: $0, \pm 1, \pm 2, \ldots, \pm m$, at angles $\pm\theta_m$ to the grating normal. These angles are given by the grating equation

$$nd\sin\theta_m = m\lambda, \tag{1.5}$$

where

λ is the optical wavelength in vacuum
n is the refractive index of the medium

He also considered the important case of illumination by an oblique incident beam. These two cases give rise to two different resolution limits, the latter being twice as fine as the former. We call the first the on-axis limit and the second the oblique illumination limit. We consider first the on-axis case. Abbe explains that the high-frequency part of the image is formed by the interference of these diffraction orders overlapping in the image plane. The on-axis situation is depicted in Figures 1.8 through 1.10 for three cases which represent a coarse grating, a fine grating and a grating at the Abbe resolution limit, respectively. The zero-order diffracted beam is not shown but plays a key role in image formation by adding a positive constant-amplitude term, which guarantees a non-negative amplitude pattern. Negative amplitudes, if squared, would lead to spurious double-frequency intensities in the image.

The first-order diffraction angle, θ_1, naturally increases as the grating spacing decreases. This can be seen to reduce the region of overlap of the ± 1 diffraction orders until, at the Abbe limit, the region of overlap vanishes.

As may be seen in Figure 1.10 and as noticed, for example, by Walton [59], using the imaging formula for a simple lens in the Abbe limiting case, FDC, FOB and EOB are similar triangles, independent of magnification. From this, it is seen that the lens aperture subtends an angle of 2θ at the grating. Thus, its numerical aperture (NA) is $n \sin\theta$. Using the grating equation (1.5) for $m = 1$, we arrive immediately at Abbe's on-axis formula:

$$d = \frac{\lambda}{n\sin\theta} = \frac{\lambda}{NA}. \tag{1.6}$$

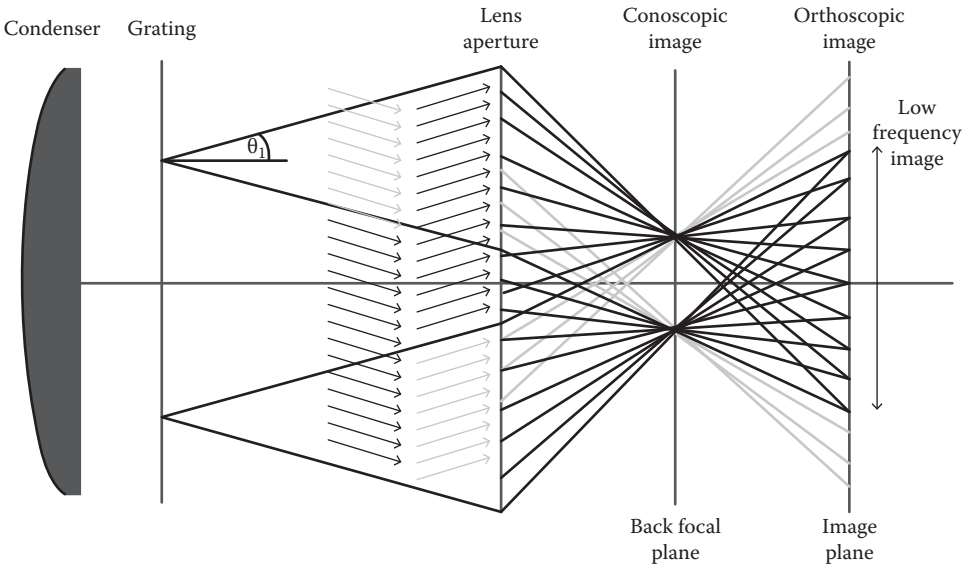


FIGURE 1.8: Illustration of the Abbe on-axis resolution limit, a coarse grating.

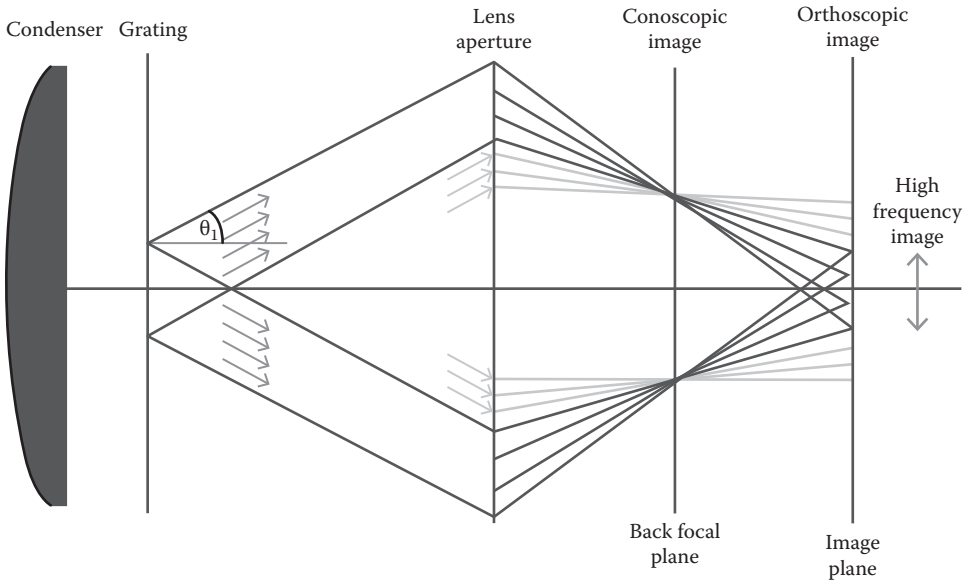


FIGURE 1.9: Illustration of the Abbe on-axis resolution limit: a finer grating.

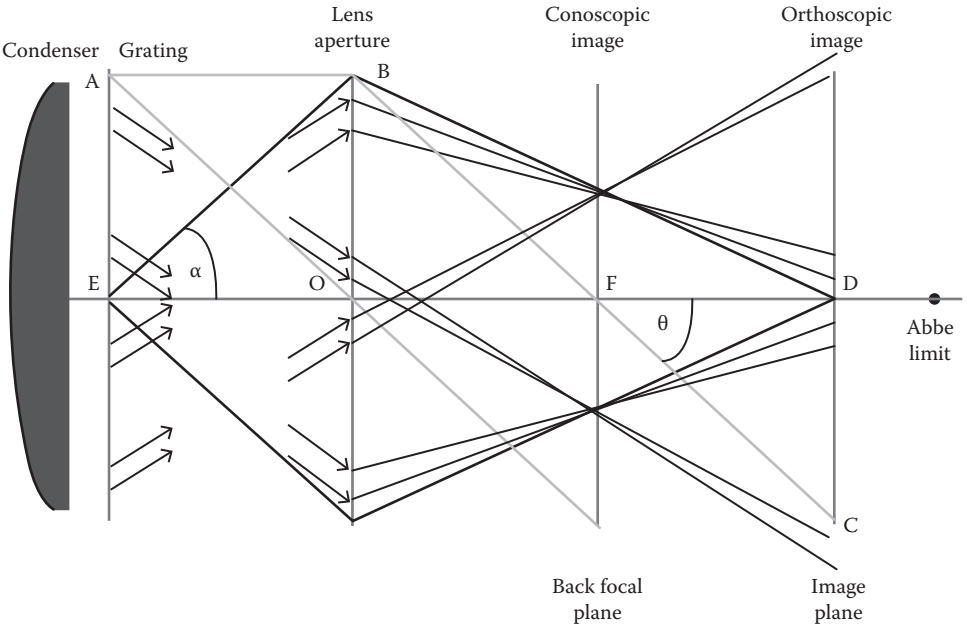


FIGURE 1.10: Illustration of the Abbe on-axis resolution limit, a grating at the Abbe on-axis limit.

In the oblique-illumination case Abbe noted that an illumination beam entering the system at the positive first-order angle of a given grating spacing could provide a diffracted beam at the negative first-order angle for a grating of half this spacing which would also enter the objective and thus double the resolution. Equation 1.6 then becomes $d = \frac{\lambda}{2n \sin \theta} = \frac{\lambda}{2NA}$ which is nowadays taken to be the 'Abbe diffraction limit.' A related strategy is to use a fully opened condenser to provide

high-input-angle illumination. We might note that the Abbe limit would only hold in an infinitesimal area in the object at the optical axis of the lens; only from such an area would the two diffraction orders pass through the extremities of the lens and meet in the image plane. A cylindrical lens would allow this to be extended to an infinitessimal band in the direction of the striations. At the limit, of course, the intensity of the image would be zero.

One should note that sometimes one finds in the literature the statement that the resolution limit is due to the fact that diffraction from higher spatial frequencies cannot enter the lens. As can be seen in Figure 1.10, this is not necessarily true and does not play a part in Abbe's argument. He was quite clear that the reason is as given earlier.

The reader should note that Figures 1.8 through 1.10 represent Gaussian optics constructions. Hence, there is an assumption that the lens is sufficiently large (no diffractive effects at the lens aperture) which is approximately true.

The Abbe theory is often described as a double-diffraction process with Fraunhofer diffraction from object to focal plane and then from focal plane to image. It is analysed as such in Born and Wolf [60], but this description relies on the rôle of an aperture in the focal plane at the expense of those at the entrance of the objective, emphasised by Abbe, and at the ocular, which, as we shall see in a later section, was emphasised by Helmholtz.

1.1.6 Digression on the Sine Condition

Abbe states that the microscope objective lens is assumed to obey the 'sine condition' (although not here named as such), one form of which was defined in this same 1873 paper, namely:

> When a system of lenses is perfectly aplanatic for one of its focal planes, every ray proceeding from that focus strikes the plane of its conjugate focus at a point whose linear distance from the axis is equal to the product of the equivalent focal length of the system and the sine of the angle which that ray forms with the axis.

This statement is said by Smith [61] to be equivalent to the large-angle sine condition, which we call later the general case or 'full' sine condition. However, in this form, the sine condition applies only to an object at infinity, as described in Born and Wolf [60]. Fortunately, for Abbe, the object and image may be interchanged in their argument which is the case when considering a microscope objective delivering essentially parallel output rays to its ocular.

More generally, the sine condition requires that a small planar object in the neighbourhood of the optical axis is imaged sharply at full aperture. The general form is attributed to Clausius [62] and to Helmholtz [63] by Born and Wolf [60], section 4.5.1, who unfortunately omit the 'infinite-image' case given in the Abbe reference quoted earlier, but give later references to Abbe [64], which they say considers the general case. Helmholtz's extension is $y\mu \sin\theta = y'\mu' \sin\theta'$, where the unprimed and primed quantities apply to each side of the lens, respectively.

It is stated by Volkmann from the Carl Zeiss company [65], in an authoritative paper on 'Ernst Abbe and his work' that Abbe, guided by his own experimental results, derived the sine condition to correct for both spherical aberration and coma and thus to achieve perfectly aplanatic imaging. Abbe was thus the first to associate the sine condition with the correction of coma. *A propos* of our further discussion on Abbe's 1873 paper to follow later, Volkmann also mentions the careful guarding of commercially valuable results by each workshop as trade secrets.

The full sine condition is also equivalent to the conservation of étendue, the product of source area and system-entrance pupil solid angle, each implying the other. The optical power transmitted through a non-absorbing 'lunette' is the product of its étendue ($m^2 \times sr$) and the source radiance ($W/(m^2 \times sr)$).

In Figure 1.12, we demonstrate simply how the sine condition arises in the aplanatic imaging of small patches perpendicular to the axis of a lens system; the optical path length (eikonal) must be the same between the end points of the object and image as that between the points on the axis. In fact, it

is the same for all conjugate points of object and image. If the refractive indices differ on either side of the lens, the optical path differences must be multiplied by their respective values. In the paraxial (Gaussian) imaging case, the sines may be replaced by the angles themselves, and in this form, it is known as the Smith–Helmholtz formula.

Robert Smith (1689–1768)* was the son of John Smith who was educated at Trinity College, Cambridge. Coached by his father, he entered Trinity himself in 1708. His mother, Hannah Smith, was the aunt of Roger Cotes (1682–1716), who also, after coaching by his uncle, went up to Trinity in 1699. Cotes was an able mathematician, said to be the second only to Isaac Newton in his time. He introduced the Euler formula, $\cos\theta + i\sin\theta = e^{i\theta}$, and is known for collaborative work with Newton, also at Trinity, resulting in, for example, the Newton–Cotes quadrature formula. He also helped to edit Newton's *Principia Mathematica*. While Robert was an undergraduate, he lodged with his cousin Roger, who, with the support of Newton, had been made the first Plumian professor of Astronomy in 1707 and was provided work as his assistant.

Thomas Plume, archdeacon of Rochester, had bequeathed nearly £2000 to the college to maintain a professor and erect an astronomical observatory over the great gate. Cotes constructed instrumentation for the observatory and made some significant astronomical observations, for example, of meteors and a solar eclipse.† On his death in 1716, Smith was elected to succeed him,‡ lecturing on optics and hydrostatics, and held the chair until 1760.

Cotes and Smith were both pioneers in optics. Unfortunately, Cotes only published one paper in his lifetime. This was on finding rational approximations as convergents of continued fractions. However, on Cotes' death, many of his mathematical papers were edited by Smith and published in 1772 in a book, *In Harmonia mensurarum et alia opuscula Mathematica*, [68]. Cotes's additional works were later published in Thomas Simpson's (of Simpson's rule) *The Doctrine and Application of Fluxions* (1776) [69]. Smith himself published *A Compleat System of Optics* in 1738 [70]. This publication was arguably the most influential optical textbook of the eighteenth century [67]. The sections on telescope design and fabrication were the most important English language manual for eighteenth century century telescope makers. It was also published in Dutch in 1753, in German in 1755 and in two different French translations in 1767. In 1739, Voltaire wrote to Smith:

> I have perus'd yr book of optics, I cannot be so mightily pleased with a book, without loving the author.

And later he praised it in his 1741 edition of *Elemens de la Philosophie de Newton* [71]. Smith later became master of the college and also vice chancellor of the university.

Lord Rayleigh (1842–1919), who also occupied the Plumian Chair at Trinity and also was chancellor of the university, wrote in 1886 [72]:

> It is little to the credit of English science that the fundamental optical theorems of Cotes and Smith should have passed almost into oblivion, until rediscovered in a somewhat different form by Lagrange, Kirchhoff and von Helmholtz.

Rayleigh continued:

> ...I was struck with the utility of Smith's phrase 'apparent distance'... and was thus induced to read his ch.5 book ii.

In Section 10 of the preface to his book, Smith writes:

> The causes that suggest our ideas of distance, and the determination of the apparent distance of an object seen in glasses, is another famous inquiry of no small difficulty, of which

* See [66,67] for historical references to Smith and this period.

† The observatory was later resited on the outskirts of Cambridge and among many illustrious succeeding holders of the Plumian chair were Sir George Airy and Sir Arthur Eddington.

‡ He also became master of mechanics and professor of astronomy to King George II.

> much has been written, but with little certainty and satisfaction to the curious. I have therefore considered this point in a very particular manner, and have settled it on such a foundation of reason and experience, as, I hope, will admit of no doubt or dispute for the future. And upon the Principle by which I introduce the consideration of apparent distance into geometry, I have not only determined it in vision with any number of glasses, but by the help of geometrical places, have shewn is regular variations, while the eye, object or system of glasses are moving forwards or backwards; and have found the variations so determined to be agreeable to experience.

He follows in the next section with

> By the help of the said principle and of an admirable Dioptrick theorem invented by Mr Cotes, I was also enabled to give very general and yet very easy determinations of the apparent distance, magnitude, situation, distinctness, brightness, the greatest angle of vision and visible area that is, of all the appearances of an object seen by rays coming from any number of speculums, lenses or mediums having plane or spherical surfaces; and in corollaries from them to deduce the known properties of telescopes and microscopes of all sorts, which however are independently demonstrated in other places of the book and remarks.

This is paraphrased by Courtney [73]:

> In one significant result Smith shows that a certain relationship between the magnification and location of object and image for one lens remains invariant for a system of lenses. This result, later again used by Smith and discovered independently by both Lagrange and Helmholtz, is now sometimes referred to as the Smith–Helmholtz formula.

Rayleigh also comments that

> Smith's splendid work ...founded upon Cotes 'noble and beautiful theorem*'...was evidently unknown to Helmholtz.

Unfortunately, Rayleigh himself seemed ignorant of the extraordinary later work of Leonhard Euler published in 1759 [74], who, together with Cotes, was acknowledged by Lagrange in the following opening to his paper [75]

> Deux grands Géometres, feu M. Cotes et M. Euler, ont entrepris de ramener La Théorie des lunettes à des formules générales. Le premier a donné le beau Théorème qu'on lit dans le Chapitre V du Seconde Livre de l'Optique de Smith, et qui sert à déterminer la route d'un rayon qui traverse autant de lentilles que l'on veut, disposées sur le même axe.
>
> Two great Geometers, as were M. Cotes and M. Euler, undertook to frame the theory of lens systems in general terms. The first gave us the beautiful theorem which one finds in chapter V of the second book of optics of Smith, and which serves to determine the path of a ray which traverses as many lenses as one wishes, aligned along the same axis.

It is also the case that recognition of Smith as well as Cotes would have been more appropriate for Lagrange's 'loi général d'optique', which considered 'lunettes' of coaxial optical elements essentially identical to those of his own.† Cotes' beautiful theorem can be found in Chapter 5 of the second book of Smith [70].

A figure of a 'lunette' (as termed by Lagrange) from Euler's paper of 1759 [74] is shown in Figure 1.11.

---

* To all appearances, therefore, it seems that Cotes was the first to understand and to use the concept now known as 'étendue'.

† Historical dates for what we might now call this 'étendue sequence' were Cotes, (1715); Smith, [70]; Euler [74], Lagrange [75], Abbe [54], Helmholtz [63] and Rayleigh [72]. We could perhaps add Gabor, [131], as a latecomer in using the Smith–Helmholtz formula as a starting point for his work on information theory and logons to be discussed later in this chapter.

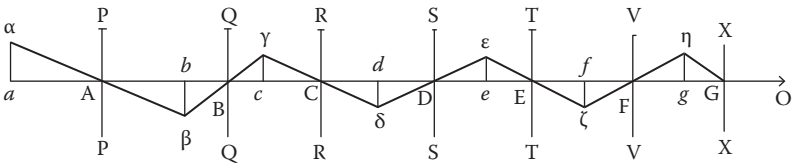


FIGURE 1.11: Euler's lunette.

For Abbe's diffraction limit to be achieved fully in practice, it is essential that lenses fulfilling the sine condition be used. It would technically also need the microscope to use cylindrical lenses!

1.1.7 Further Discussion on Abbe's Work

It is interesting to note that, although Abbe promised several times that he would publish a theoretical explanation of his results and indeed explained it to his University students in his own way, he is generally thought never to have done so. We must state, however, that his famous resolution formula does appear belatedly but explicitly in his 1882 paper [56] cited earlier.

It is worth mentioning here that Lummer and Reiche in 1910 [76] published an eagerly awaited, but in the event disappointing, work. It was professedly a reproduction of Abbe's theories as propounded by Abbe himself. Lummer attended a series of lectures by Abbe in the winter of 1887 in Jena, where Abbe had been appointed professor in 1870. The work was said to be founded solely on the carefully preserved notes of these lectures. According to an anonymous reviewer of this book in *Nature* [77], however, this was not the case, as the following quotation from the review indicates

> ... this is not the only feature which tends to produce a sense of uncertainty as to how far the account given can be regarded as a direct reproduction of Abbe's presentation of the subject.

In Fripp's translation [78] of Abbe's paper referenced earlier, he comments in his foreword:

> I may here state that the mathematical demonstrations on which Dr. Abbe builds his theory, and the detail of experimental method pursued by him in the practical portion of this enquiry, are not communicated in the present article, which is simply a general statement of results.

Lord Rayleigh in 1896 [79] also comments that

> In the earliest paper by Professor Abbe, which somewhat preceded that of Helmholtz, similar conclusions were reached; but the demonstrations were deferred, and, indeed, they do not appear ever to have been set forth in a systematic manner.

In 1906, also Porter [58] stated that

> …the complete mathematical development has never been published.

A further doubter was Dr. J.P. Gordon in a paper from an optical conference in 1912 in London [80], which includes the following two paragraphs:

> The theory of diffraction, as developed in works on optics at the present time is in a strangely incondite condition, so much so, indeed, that the imperfections of the accepted textbooks in this respect afford ground for a serious reproach to our 20th-century science. Two classes of diffraction phenomena are described, and are called by the names of Fresnel and Fraunhofer respectively; the latter exhibited in focal planes, the former in what Sir Almroth Wright has aptly called the apertural planes of optical instruments of the telescope,

> for example. But these two classes of phenomena are usually treated apart and investigated on different lines, so that the connection between them is obscure.
>
> The only successful attempt to bring them into clear relationship with one another known to the present writer is that made by Professor von Helmholtz in his paper in Poggendorff's Annalen of 1874 on the theoretical limits of the resolving power of the microscope*. In that paper the principle which systematises the whole body of phenomena is rather implied than explained and, perhaps for that reason, it has not obtained such currency in scientific literature as its appearance under such conditions would seem calculated to secure. In fact, Fresnel's method of investigation is practically the only method in use for explaining diffraction phenomena, and Fresnel's method, for a reason to be presently mentioned, is wholly unsuited to the investigation of Fraunhofer phenomena. As a consequence of this inadequate discussion of the subject the phenomena discussed are very imperfectly dealt with.

Gordon thus not only avoids even mention of Abbe but also denigrates Helmholtz's publication a year later of work on the limits of microscope resolution, which deferred to the priority of Abbe (see Section 1.1.8).

This apparent lack of explicit theory has also been quoted by a number of later authors. It has even been hypothesised by Fripp, in his translation of Abbe already quoted, that commercial confidence may have been a factor. This followed an article by Lardner [81], in which he stated

> Now the solution of this problem presented ...difficulties so great as to have been regarded by some of the highest scientific authorities ...as absolutely insurmountable ...we must admit that its solution has mainly been the work of practising opticians.

Fripp writes:

> We still look for an adequate scientific theory of the microscope in our present micrographic literature. And the rules and methods of construction now employed in such optical combinations as the microscope objective, are known only to those who have made personal sacrifices of time, study and money to attain it. In a word, the most successful and important achievement of optical science is a trade secret. It is scarcely possible to urge a stronger proof of the value.

In our view, the problem with Abbe's 1873 paper has been greatly overstated, in that a careful reading of the almost impenetrable lengthy and florid German text reveals that, in spite of the complaints of Rayleigh and others, all the facts of the matter are indeed present and there would be nothing of any significance added by rewriting it in the form of mathematical equations. These can be fully constructed by reading the text. It might even be speculated that equations were deliberately avoided to make the work more acceptable to its intended audience. Even today, researchers in some less quantitative fields are known to be frightened of them! In fact, Abbe explicitly states in the paper that it a condensed summary:

> in the hope that it will be acceptable to many practical microscopists.

The full paper which he then promised would follow was delayed by illness and unfortunately never appeared. However, although Abbe states that the paper

> in no wise claims to be a full development or establishment of the facts to be set forth

as the following translated quotations show, there is no doubt that his verbal arguments describe perfectly the details of his diffraction limit of resolution as it is known today.

---

* This paper will be discussed shortly later.

First, we have the Abbe statement that

> ... all minute structures whose elements lie so close together as to occasion noticeable diffraction phenomena will not be geometrically imaged, that is to say, the image will not be formed, point for point, as usually described by the re-union in a focal point (or plane) of pencils of light which, starting from the object, undergo various changes of direction in their entrance and passage through the objective; for even when the dioptric conditions requisite for such a process are fulfilled, the image so formed shows none of the finer structural detail, unless at least two of the diffraction-pencils which are caused by the splitting up of rectilinear rays are re-united.

He goes on to say that

> the conclusions here deduced from facts won by direct observation, are fully substantiated by the theory of undulation of light, which shows not only why microscopic structural detail is not imaged according to dioptric law, but also how a different process of image formation is actually brought about. It can be shown that the images of the illuminating surface, which appear in the upper focal plane of the objective, (the direct image and the diffraction images) must each represent, at the point of correspondence, equal oscillation phases when each single colour is examined separately.

Even more specifically we quote further:

> The proof that an objective can resolve very minute striae on a diatom or Norbert's test plate,* attests, strictly speaking, nothing more than that its angular aperture answers to the calculable angle of diffraction of the interlinear distance of the striae on the test, and that it is not so badly constructed that a sufficient correction of its outer zone is impossible.

We must surely assume that the 'calculable angle' was indeed able to be calculated by Abbe himself. Also, as discussed earlier, Abbe spells out explicitly in words the associated 'infinite image' sine condition, although note that there are some small misprints in Fripp's published translation of this condition†; a correct translation is given shortly in the following text.

Apart from these misprints, the paper was translated extremely well by Fripp and read at a meeting of the Bristol Microscopical Society in the United Kingdom on 16 December 1874. It was then published in English with a translator's preface [78].

Without impugning his priority in formulating the famous diffraction limit of microscopic resolution, Abbe's views in this first paper can, however, be corrected in the light of later developments with which he himself concurred. This is recounted in the seventh edition of 1891, updating the sixth edition, of the popular and influential book of the time on Microscopy by William Carpenter, MD, FRS [82]. Carpenter, a physiologist and registrar of the University of London for 25 years, oversaw the first six editions from 1856 to 1881 (we know that over 10,000 copies were sold worldwide even up to the fourth edition) but he was eventually obliged to hand over the preparation of the seventh edition to the Rev. W.H. Dallinger, FRS. Under Carpenter, it had been too early to feature Abbe's work at all and therefore Dallinger rewrote the first five chapters and replaced them with seven essentially new ones. He had asked Abbe to summarise, for this edition,

> the results of his twenty years of unremitting and marvellously productive labour.

Unfortunately, the state of Abbe's health and his many obligations did not permit this, but Abbe did consent to examine the results and commented, as quoted in its preface:

> I feel great satisfaction in seeing my views represented in the book so extensively and intensively.

* Friedrich Norbert of Barth in North Germany used a circle divider to produce parallel rulings down to 110 nm.

† The first occurrence of the word 'sum' should be 'product' and the second should be 'sine'.

In his original papers, Abbe had decisively divided microscopic imaging into two parts which he called, respectively, 'defining' and 'resolving'. The former took place at low aperture and conformed with dioptric geometric conventions, in which a recognisable representation of the absorption and scattering of the object was visible directly in the image. The latter occurred as a separate phenomenon when the object itself additionally contained sufficiently fine structures or minute particles to diffract light at high angles through the outer edge of the objective aperture. He writes (in translation):

> it appears that the production of microscopic images is closely connected with a peculiar and hitherto neglected physical process, which has its seat in and depends on the nature of the object itself.

In this case, no direct similarity of object and image was seen but 'undulatory theory' was employed to interpret the image and resolving power as we have discussed earlier. Our present-day understanding of imaging as a single Fraunhofer diffraction mechanism involving the far-field Fourier transform of the entrance aperture was to evolve much later.

In fact, on page 64 of the seventh edition of the aforementioned book, Dallinger makes clear that Abbe no longer held his original views on two different sorts of imaging, thus falling in line with the modern view.

At a later stage in his work and with previous contributions by G.B. Amici in Italy in 1840, who introduced water-immersion lenses, F.H. Wenham, and J.W. Stephenson in the United Kingdom, E. Hartnack in Paris and R.B. Tolles and C.A. Spencer in the United States (see, for example [83]), Abbe in 1877 determined that oil of cedar wood was an optimum match for crown glass in an immersion (or homogeneous) system and was able to increase the NA from the maximum possible value of 1.0 for a dry objective to 2.5 and over in a range of Zeiss objectives. In fact, the idea of using an immersion fluid in microscopy goes back all the way to the use of water immersion in the seventeenth century by perhaps the most renowned of all microscopists, Robert Hooke.

Of course, in view of the originality and importance of Abbe's original work, it has had countless reviews and references over the years to the present day, which we can in no way attempt to list. We hope, however, to have summarised here the essence of this literature and will move on to consider the parallel work of his contemporary, Hermann von Helmholtz.

1.1.8 Work of Helmholtz

Another confirmation of the priority of Abbe in the question of the resolution limit is afforded by the fact that in the year following the publication of his paper, Helmholtz arrived at the Abbe formula explicitly in a paper submitted to the Annalen der Physik und Chemie [63]. At the point of submission, Helmholtz became aware of the paper of Abbe and immediately conceded Abbe's priority; so much so that he considered withdrawing his own. That he did not was due to the fact that it was scheduled to appear in a prestigious 'jubilee' edition of the Annals celebrating 50 continuous years of editorship by Poggendorf. He felt that he could not withdraw on this special occasion. Instead, the following statement was included in an explanatory postscript:

> Die besondere festliche Veranlassung, zu welcher dieser Band der Annalen veröffentlicht wird, verbietet mir, meine arbeit zurückzuhalten oder ganz zurückziehen. Da sie die von Herrn Abbe noch zurückgehaltenen Beweis der von uns beiden gebrauchten Theoreme und einige einfache Versuche zur Erläuterung der theoretischen Betrachtungen enthält mag ihre Veröffentlichung auch vom wissenschaftlichen Standpunkte aus entschuldigt werden.

The translation of this explanation we have already summarised earlier. In this section, we shall see that it was a wise choice to go ahead and publish since, although the same formula is reached, in spite of the statement of Fripp in our following paragraph, the viewpoint and treatment of diffraction is quite different in the two papers.

Fripp also translated this paper of Helmholtz [63], and in comparing it with the work of Abbe, stated that

> The theoretical grounds taken by these two authors are identical, and their results, so far as the researches were directed to the same points, also agree. But in each essay the mode of treatment is thoroughly independent, and the experimental proof of the conclusions respectively obtained is conducted by each writer in a separate and original method. The mathematical demonstrations omitted in Professor Abbe's article are fortunately supplied by Professor Helmholtz, and the two essays are confirmatory and supplementary to each other in several other respects, whilst in both we recognise that clearness of thought and precise knowledge of the subject treated, which justifies entire confidence in the conclusions.

We note, however, that Fripp's enthusiasm for the work of Helmholtz was not shared later by Gordon who, as stated in our earlier reference to him in Section 1.1.5 of this chapter, felt that in Helmholtz's paper:

> the whole body of phenomena is rather implied than explained.

In this paper, Helmholtz jumps straight into the effects of diffraction. He states that

> If, perhaps, occasional allusion has been made to diffraction as a cause of deterioration of the microscopic image, I have yet nowhere found any methodical investigation into the nature and amount of its influence, but such an investigation shows, as will here appear, that diffraction necessarily and inevitably increases with the increase of magnifying power, and at length presents an impassable limit to the further extension of microscopic vision. That diffraction and consequent obscurity of microscopic image must necessarily increase with increasing amplifications of the image, and this quite independently of any particular construction of the instrument, rests as a fact upon a general law which applies to all optical apparatus, and which was first formularised by Lagrange.

From this quoted work of Lagrange of 1803 [84], it is known that the magnification of any system of coaxial lenses (termed as 'lunette' since the eighteenth century), which, from a microscopic source on axis as input, delivers an almost cylindrical output beam parallel to the common axis, has a magnification given by the ratio of the diameters of the input cone at the objective and the cylindrical exit beam at the ocular input. This latter is usually of small diameter since it is optimally adjusted to the area of the pupil of the eye (normal amplification).

Bearing in mind the results of Abbe, unknown at the time by Helmholtz, we thus have something of a confusion, if not a contradiction, well spotted by Gordon, in Helmholtz's description earlier of the origin of diffraction in the microscope.

Clearly, if higher magnification is achieved by increasing the input NA, this only collects higher-angle diffracted light and does not cause more or less diffraction, which is wholly determined by the structure of the object itself and the illuminating beam. In fact, as clearly understood by Abbe, the more diffracted light that is collected, the more 'informative' will be the image, which will be composed of more diffraction orders. Contrary to the implication of Helmholtz's statement, the higher the magnification, for a given ocular diameter, the more should be the 'extension of microscopic vision'.

Helmholtz's arguments against this viewpoint may be seen later in the paper where he states that

> The theory of diffraction of rays in the microscope leads, as will be shown in the following pages, to the conclusion, that any single point of light in the object must, when viewed through the microscope, appear exactly as if an actual luminous point, situated in the place of the object, were observed through an aperture corresponding in size and position to the ocular images (at the so-called eye spot) of the respective narrowest diaphragm aperture.

Diffraction effects of fine structures in the object itself, the overlapping of which in the image to manifest these details was the primary concern of Abbe, are not mentioned by Helmholtz at this point in his discussion.

Note that in the aforementioned penultimate paragraph, we have considered the ocular field–lens diameter to be fixed so that increased magnification would be achieved by increasing the NA of the microscope objective. If increased magnification instead gives rise to a reduced beam diameter at the ocular, then indeed diffraction by this aperture of a point of light at the object will give rise to diffraction rings or stripes which will increasingly blur the object the higher the magnification. To exhibit such diffraction effects, Helmholtz estimates the ocular input aperture for light of 550 nm wavelength to be less than around 1.8 mm diameter.

How is it then that both of these researchers eventually arrive at the same formula for the limit of resolution for the microscope from two apparently orthogonal calculations? One might also ask how Fripp's comments in his translation of the Helmholtz paper concluded that Abbe and Helmholtz were tackling the same problem. His statement that

> The theoretical grounds taken by these two authors are identical

seems to be just incorrect. His quote from book 1 of Virgil's Aeneid about the state of the literature in the field in 1876 'tantaene animus celestibus irae?' (in Dryden's translation 'Can heav'nly minds such high resentment show?') is more to the point.

One can also sympathise with Gordon's choice of the adjective 'incondite' for the state of the field as late as 1912, even though this somewhat understates the problem. Helmholtz himself must be guilty of not understanding completely the work of Abbe. It seems that historical confusion still reigns even to the present day.

Helmholtz goes on in his paper to prove the following theorem, first for the case of small angles, where the sine or tangent of the divergence angles within the instrument are approximately equal to the angles themselves, and later in the paper for arbitrary values of the divergence angles, where they are all to be replaced by their sines; all images are stipulated to be in planes at right angles to the optic axis.

> In a centred system of spherical refracting or reflecting surfaces the product of the divergence-angle of any ray, the refraction index of the medium through which that ray passes, and the magnitude of the image to which the rays passing through that medium belong, remains unchanged by every refraction, provided always that the conditions of production of an accurate image are duly preserved. This product will therefore have the same value after emergence of the rays as it had before they entered the system of lenses.

We have met this condition, the sine condition, already in this chapter as a generalisation, illustrated in Figure 1.12, of Abbe's 'infinite-image' sine condition of 1873.

An immediate implication of the sine condition is that for a given 'lunette', the larger the input NA observing a microscopic field the narrower the divergence angle of the emerging rays at the ocular. This is the rule that connects the diffraction effects considered by Helmholtz with high magnification. The smallest diffraction ring from a circular aperture of diameter d emerges at a visual angle of $\sin^{-1}(1.22\lambda/d)$, and outer rings have width $\sin^{-1}(\lambda/d)$, the same as the angular diffraction orders of a line grating of spacing d.

1.1.9 Filters, Signals and Fourier Analysis

We have seen in both Fraunhofer diffraction and the Abbe theory that decomposition of the object in terms of sinusoids is important. Another subject where such a decomposition is routinely used is communication theory and we will see that there are very strong parallels between band-limited communication channels and 1D coherent imaging. In the former case, there is a finite maximum frequency of sinusoid which can be transmitted down the channel, whereas in the latter case, there

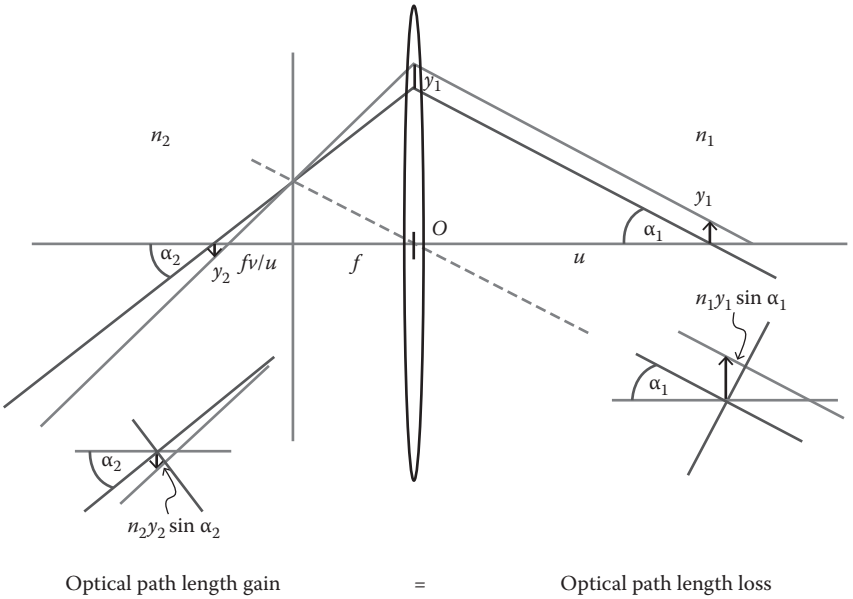


FIGURE 1.12: Illustrating the sine condition; for small y_1 and arbitrary α_1; the optical path lengths between all corresponding points on the object of height y_1 and its image of height y_2 are the same, that is, $n_1 y_1 \sin \alpha_1 = n_2 y_2 \sin \alpha_2$. The lens is then aspheric and coma free.

is a finite maximum spatial frequency which can contribute to the image. To go between the two problems, we merely have to change the time variable into a space variable and the frequency variable into the spatial frequency. In the rest of this chapter and in the following one, we will pursue this analogy, though we remind the reader that for it to hold, there are significant approximations in the optics problem.

We will also see that we can draw on the parallels between communication theory and imaging to apply the notion of information, which is central to communication theory, to imaging.

It is a reasonable question to ask why Fourier analysis should have been so widely used in signal processing. To answer the question, we will give an argument put forward by Slepian [85]. Signals in communications and other areas of electrical engineering are often processed using what are termed linear time-invariant (shift invariant) filters or channels. Given an input signal $s_{in}(t)$, a filter F and an output signal $s_{out}(t)$, we write

$$s_{out}(t) = F[s_{in}(t)].$$

The filter F is linear if for two input signals $s_{1in}(t)$, $s_{2in}(t)$, and two coefficients c_1 and c_2 one can write

$$c_1 s_{1out}(t) + c_2 s_{2out}(t) = F[c_1 s_{1in}(t) + c_2 s_{2in}(t)].$$

In particular, for such an F,

$$c s_{out}(t) = F(c s_{in}(t)). \tag{1.7}$$

F is time invariant if

$$s_{out}(t - \tau) = F[s_{in}(t - \tau)], \quad \forall \tau \in \mathbb{R}.$$

Consider the response of a linear, time-invariant filter to a complex sinusoidal input signal $s_{in}(t) = e^{i2\pi ft}$. Denoting the output signal by s_{out}, we have

$$s_{out}(t-\tau) = F(e^{i2\pi f(t-\tau)}).$$

Now noting the property of exponential functions

$$e^a e^b = e^{a+b}$$

and using (1.7) with $c = e^{-i2\pi f\tau}$, we have

$$s_{out}(t-\tau) = e^{-i2\pi f\tau} F(e^{i2\pi ft}). \tag{1.8}$$

Taking the exponential over to the other side, we have

$$e^{i2\pi f\tau} s_{out}(t-\tau) = F(e^{i2\pi ft}) = s_{out}(t).$$

Noting that our choice of τ was arbitrary, for the second step to be true, s_{out} must itself be a complex exponential times some factor d:

$$s_{out}(t) = de^{i2\pi ft}.$$

To determine this factor d, put $t = 0$ in (1.8) which yields

$$d = F(1).$$

From this, it follows that the output signal is the input signal multiplied by a complex factor $F(1)$. In other words, viewing the filter as an operator, its eigenfunctions are the complex sinusoids, with corresponding eigenvalues $F(1)$. In general, $F(1)$ will vary from frequency to frequency. To recognise this, let us denote $F(1)$ by $H(f)$. The function $H(f)$ is known as the transfer function of the filter F. Since signals are typically filtered by linear time-invariant filters and the complex sinusoids are a natural set of functions for analysing such filters, it became *de rigueur* to expand the signals themselves in terms of complex sinusoids. As a consequence, Fourier analysis has been a standard tool throughout the development of electrical engineering.

We might remark that a more mathematical argument to justify the use of complex sinusoids, which we will not elaborate here, is that when dealing with continuous-time, stationary signals, the inherent time invariance can be described by the additive group of translations on the real line. The characters of this group are the exponentials e^{iyx}, for real x and y.

It should be borne in mind that in real-life time-invariance is an approximation which will be accurate only when dealing with oscillatory signals of sufficient length compared with their periods. Very often this is true; however, Fourier analysis should not be automatically applied without careful consideration.

1.1.10 Optical Transfer Functions and Modulation Transfer Functions

The Abbe theory suggests that Fourier methods are likely to be useful in optics. It was recognised by Frieser [86], and Selwyn [87], that the image of a sinusoidal object is also sinusoidal. The Fourier approach to optics was developed by Duffieux [88], and Luneberg [89], among others.

In his review of the second edition of Duffieux's book, Welford [90] states that the introduction of Fourier methods into optics should strictly be ascribed to Abbe, Rayleigh and Michelson but that its revival and growth in the 1940s, 1950s and 1960s was due to a great extent to Prof. Duffieux who

was the first to see the concept clearly. This work, together with his paper with Lansraux of 1945 [91], started a major change in ideas on optical image formation. Similar observations on the rôle of Duffieux were made by Hopkins in his Thomas Young oration of 1962 [92]. The subject is covered well by Goodman [47].

The resolving power of an optical imaging system is often analysed using the optical transfer function. This is a direct analogue of the transfer function in signal processing which we have just discussed. A full discussion may be found in Williams and Becklund [49]. There are two different types, depending on whether one is dealing with coherent or incoherent imaging.

The more commonly encountered transfer function is that associated with incoherent imaging so we will start with this. Suppose we have a sine wave modulating a positive constant function as our object intensity. The value of the constant is chosen so that the resulting function is nonnegative, as indeed, it must be if it is an intensity.

After passing through the system, provided the system transforms sine waves to sine waves, the output intensity distribution at the detector will also be a sine wave modulating a constant function, though the sine wave may have experienced some phase shift.

Let us assume the intensity distribution of the object is given by

$$I(x) = a + b(\omega)\sin \omega x.$$

Let the intensity distribution of the image be

$$I'(x) = c + d(\omega)\sin(\omega x + \phi(\omega)).$$

The modulation (or contrast) of the object or image is given by $b(\omega)/a$ or $d(\omega)/c$, respectively.

We define the modulation (or contrast) factor of the imaging system $D(\omega)$ to be the ratio of the modulation of the image to the modulation of the object:

$$D(\omega) = \frac{d(\omega)a}{b(\omega)c}.$$

We then define the optical transfer function $T(\omega)$ by

$$T(\omega) = D(\omega)e^{i\phi(\omega)}.$$

This is essentially (up to a normalisation constant) the Fourier transform of the system point-spread function. For coherent imaging, the optical transfer function is just defined to be the Fourier transform of the amplitude point-spread function.

Recalling (1.2) and (1.3) we note that the point-spread function in (1.3) is the square of the point-spread function in (1.2). Using the convolution theorem and the knowledge that the point-spread function in (1.2) is real, we then have that the transfer function for the incoherent problem is the autocorrelation function of that for the coherent problem. This is more generally true for any pair of point-spread functions but one should note that if the point-spread function for the coherent problem is complex, then that for the incoherent problem is the modulus squared of the former one. For the 1D problem, the transfer function for (1.2) is just the top-hat function, that is, it is unity over the aperture and zero elsewhere. The transfer function for the incoherent problem is then a triangular function of twice the width of that for the coherent problem.

The diffraction limit for both coherent and incoherent imaging is defined to be the sinusoid at the cut-off frequency. Since the cut-off frequency for the incoherent problem is twice that of the coherent one, one could argue that the solution should have twice the resolution. However, this is over-simplistic. First of all, one is not comparing like with like; incoherent imaging involves intensities, whereas coherent imaging involves amplitudes. Second, even if one overlooks this,

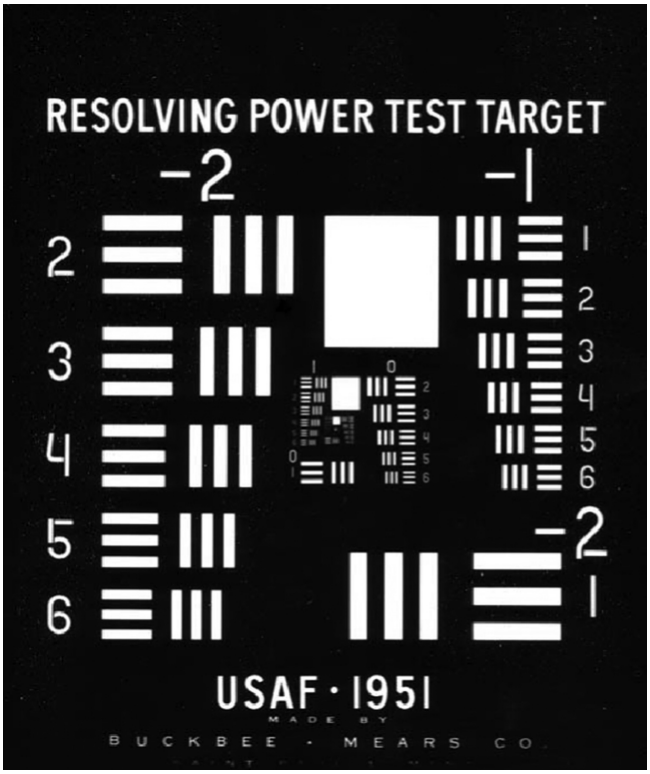


FIGURE 1.13: U.S. Air Force resolution chart of 1951.

Goodman [47], p. 156, gives an example of an object which is resolved better in coherent imaging. In practice, the resolution is defined at the point where the modulation transfer factor is a few percent of its maximum height. The Rayleigh criterion for incoherent imaging corresponds to a height of about 9%.

The foregoing suggests that a good way of measuring the resolution of an optical system is to use sinusoidal objects of varying frequency. A more popular way for incoherent imaging has been to use a bar target. One way of constructing this is to use a long line of bars resembling a picket fence. The width of the bars is varied (by having different spacings on a resolution chart) until they can just be seen, and this is defined to be the resolution limit. For reasons of compactness, the number of bars at each spacing is sometimes restricted to three, as in the U.S. Air Force resolution chart of 1951, shown in Figure 1.13. Typically, the resolution is quoted in terms of line pairs per unit of length, where a line pair consists of a dark line and a light one. The resolution of the human eye is better than 1 minute of arc per line pair in the fovea, falling off away from the fovea.

Figure 1.14 shows a spoke target, also called a Siemens star. Use of this type of resolution chart gives information on how the resolution changes with angle. Modern charts for testing the resolution of digital cameras such as the ISO 12233 test chart are considerably more sophisticated.

It will be perceived that there are two obvious problems with using resolution charts. The first of these is that they use square waves rather than sinusoids. If one carries out the Fourier series decomposition of a square wave, it will involve multiple frequencies. Each of these is attenuated differently by the system and so the square wave loses shape as it is transmitted. The second problem is that only short portions of the square waves are used on the chart. These represent poor approximations to sine waves. Nevertheless, one can Fourier analyse bar charts (see Williams and Becklund [49] for

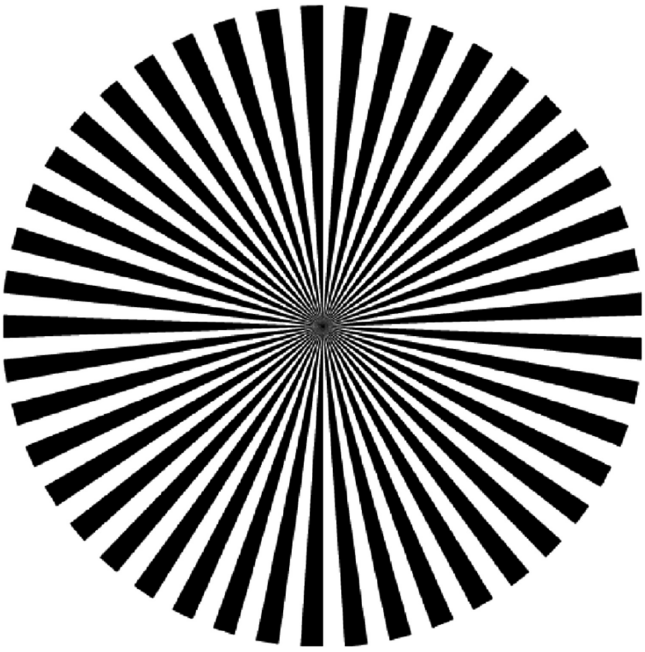


FIGURE 1.14: A typical spoke target.

an analysis of the three-bar pattern) and from this, determine how different frequencies are affected by the system.

The performance of the human eye has been studied using transfer-function methods and some results can be found in Barten [93].

1.1.11 Some Observations on the Term Spectrum

The word spectrum occurs in different contexts throughout this book and it is perhaps worth pointing out some of the meanings here. We have already encountered the analysis of signals and objects in optics in terms of sine waves. Fourier analysis of this sort could be said to consist of determining the amplitude spectrum of the signal or object. Engineers think of spectral analysis as the determination of the power-spectral density (or spectrum) of a time series from a given portion of that time series. Mathematicians, when dealing with linear integral operators, refer to eigenvalue spectra and singular-value spectra. Physicists often use the term spectrum to represent the range of colours which light is broken down into when it passes through a prism. Within the context of this book, the word spectrum will be qualified by an additional word such as 'amplitude', 'power', 'eigenvalue' or 'singular value' to avoid confusion.

It is interesting to look into the etymology of the word spectrum since its meaning has changed from its original one. Guerlac [94] gave an account of its origins. He stated that Newton, in his work *Opticks* of 1704, referred to a solar image projected onto a screen as a spectrum. By this, he meant something which could be made to appear on the screen. This could then be referred to as an apparition, ghost, phantom or spectre, or in Latin, spectrum. When the solar light was passed through a prism, it was broken down into a range of colours. Up until the early nineteenth century, the resulting apparition on the screen was referred to as a prismatic spectrum. Since then, the word 'prismatic' has been dropped so that a word which originally meant a phantom now means a range of some property such as colour.

1.2 Resolution and Prior Information

A perusal of the literature might lead one to the conclusion that the concept of resolution is surprisingly vague. This vagueness is often due to prior information about the problem in hand not being made explicit. A useful historical survey of some different concepts of optical resolution is that of den Dekker and van den Bos [95]. For the problems we are largely preoccupied within this book, we assume that the object we seek is more or less smooth. However, there are situations where this is not appropriate. If one knows that the object consists of one or two point sources, the resolution problem simplifies dramatically.

1.2.1 One- and Two-Point Resolution

Some of the early work on resolution was concerned with the separation of spectral lines by a spectrometer. Even though spectral lines have non-zero width, due to various mechanisms, it can still be appropriate to model them as lines of zero width. Under these circumstances, two-point resolution is a meaningful concept, where the points in question are the positions of two neighbouring lines. The textbook problem is to decide whether one is looking at one or two points. The Rayleigh resolution criterion addresses this problem. There are various other similar two-point resolution criteria, in particular, those of Dawes, Schuster, Sparrow and Houston.

Two-point resolution for the human eye is known as visual acuity. According to Born and Wolf [60], the angular separation of two-point sources which can be resolved by the eye is about 1 minute of arc, corresponding to a separation of 0.0045 mm on the retina. Given that the smallest cones are about 0.0015 mm across, we see that this resolution corresponds to about three cones width.

There is also a concept of one-point resolution where the problem is to determine the position of the point source. If one adopts this notion of resolution, then every scientific experiment which aims to determine some parameter has a resolution problem at its core.

These one- and two-point problems can be thought of as parameter estimation problems, though, in the two-point problem, elements of decision theory are also needed. Two-point resolution theory should only be used where one has very strong prior information that one is looking at isolated point sources.

1.2.2 Different Two-Point Resolution Criteria

In this section, we mention several two-point resolution criteria which are reasonably well known. Clearly, there is potentially an uncountable infinity of such rules of thumb. For 2D problems where the point-spread function is not circularly symmetric, we consider the appropriate slices through the point-spread functions. The following criteria correspond to incoherent imaging.

The Dawes criterion, proposed by Dawes [96], applies to the separation of closely spaced double stars using a telescope. The angular separation in radians is given by $1.02\lambda/D$ where D is the aperture diameter and λ is the wavelength of light. It is close to the full width at half maximum of the point-spread function and corresponds to a 5% dip between the maxima when the point-spread function is the Airy pattern.

Schuster [97] proposed the criterion that two point sources are resolved if the main lobes of their point-spread functions do not overlap. This is equivalent to twice the Rayleigh-criterion spacing.

Houston [98] suggested comparing the distance between the central maxima of the composite intensity pattern with the full width at half maximum of the individual point-spread functions. If the former is greater than the latter, the sources are said to be resolved. This is thus very close to the Dawes criterion, if one applies it to telescopy.

Sparrow's criterion [99] corresponds to when the dip in between the two central maxima ceases to exist. Given that one knows that one is dealing with either one- or two-point sources, the extended shape of the image indicates that two sources are present. The Sparrow limit is roughly half the Rayleigh limit.

Buxton's criterion [100] deals with the amplitude diffraction patterns and defines two-point sources to be resolved when the closest points of inflexion of the diffraction patterns coincide.

Two-point resolution criteria for the case of partial coherence between the point sources are discussed in den Dekker and van den Bos [95].

It is an obvious question as to what happens if the point sources are of unequal strength. This opens a new Pandora's box and is discussed in Treanor [101] for the case of double stars. He came to the conclusion that if the fainter star coincided with the first minimum of the diffraction pattern of the brighter one and the peak of the fainter star was greater than the first sidelobe of the brighter one, the stars were resolvable.

1.3 Communication Channels and Information

We have already mentioned that there is a close connection between resolution in coherent optics and the ability of a communication channel to transmit information. We can think of a scientist imaging an object using a microscope as receiving information about the object transmitted by Mother Nature; the better the resolving power of the microscope, the more information about the object is received. There is hence an intuitive connection between the level of resolved detail and amount of information. We will now consider this in more detail. From our perspective, there are four main ingredients to communication theory. These are the bandlimit of the channel, the sampling theorem, the $2WT$ theorem and the use of entropy in quantifying information flow down the channel. In this section, we will look at the origins of these ingredients. They will then be brought together in the following sections.

Until the last half of the last century, the quantification of resolution for communication channels relied heavily upon the use of Fourier analysis. In this approach, any 'signal' to be resolved is broken up into sine and cosine waves of increasing frequency, up to a limit beyond which the system under consideration is unable to transmit. The permissible frequency range may also be limited from below and the system is said to have a 'bandpass filter' property. The system then distorts signals which have frequency components outside the band; in particular, it spreads out sharp signals so that, if they are too closely spaced, they may not be able to be separated.

The so-called 2WT theorem gives the number of independent signals of duration T which can be transmitted down a communication channel of bandwidth W as $2WT$. This theorem, though strictly incorrect, is approximately valid for large enough values of T.

When the duration of real signals transmitted down an inevitably band-limited communication channel is not sufficiently great that the 2WT theorem is valid, it is of interest to pose the question as to how many linearly independent band-limited signals of a certain time duration can be used. This can be viewed as a resolution problem with the number of such signals corresponding to the number of resolvable elements (number of degrees of freedom), and we will discuss this 'information capacity' further in the following text in connection with the work of Nyquist, Hartley and Shannon. We will see in the next chapter that there are better basis functions than sinusoids for analysing this problem (see [102]) but for the current chapter, which is a historical overview, we will stick with sinusoids.

1.3.1 Early Steps towards the 2WT Theorem

Nyquist and Küpfmüller independently arrived at the conclusion that the number of telegraph signals which can be transmitted down a line per second (the signalling speed) is proportional to the bandwidth of the line. Though it is often quoted that Nyquist discovered this in his 1924 paper [103] it is only stated explicitly in the 1928 paper [104]. Küpfmüller [105] does arrive at the proportionality between signalling speed and bandwidth.

Hartley [106] made the next step towards the 2WT theorem by concluding that the maximum rate at which information may be transmitted down a band-limited communication channel is proportional to the band limit. He further concluded that the total amount of information which can be transmitted down the channel is proportional to the band limit and the time which is available for transmission.

1.3.2 Nyquist Rate

In his 1928 paper [104], Nyquist proved that the minimum bandwidth, in Hz, required for reconstructing the code elements unambiguously was half the number of code elements transmitted per second. Conversely, for a given band limit, W, $2W$ is the maximum number of code elements per second that can be transmitted and received unambiguously. This maximum rate at which the code elements are sent is referred to as the Nyquist rate. The time interval $1/2W$ is referred to as a Nyquist interval. We reproduce here the reasoning in [104] since it is illuminating.

Consider a signal $E(t)$ of total duration T seconds which takes the form of N rectangular pulses of height a_i, $i = 1, \ldots, N$. Assume the time origin is such that

$$E(t) = a_h, \quad \frac{(h-1)T}{N} < t < \frac{hT}{N}, \quad h = 1, \ldots, N.$$

Assume that this signal is periodically extended so that it repeats every T seconds. It may then be expanded in a Fourier series:

$$E(t) = \frac{A_0}{2} + \sum_{n=1}^{\infty} (A_n \cos npt + B_n \sin npt),$$

where $p = 2\pi/T$. Projecting $E(t)$ against $\cos kpt$ and $\sin kpt$, respectively, to find the coefficients A_k and B_k, one finds

$$A_k = \frac{8 \sin \omega_k/4s}{\omega_k/s} \frac{1}{N} \sum_{h=1}^{N} a_h \cos \frac{2\pi k}{N} \left(h - \frac{1}{2}\right),$$

$$B_k = \frac{8 \sin \omega_k/4s}{\omega_k/s} \frac{1}{N} \sum_{h=1}^{N} a_h \sin \frac{2\pi k}{N} \left(h - \frac{1}{2}\right),$$

where

$$\omega_k = \frac{2\pi k}{T}$$

and

$$s = \frac{N}{2T}.$$

The parameter s is referred to as the speed of signalling. Following Nyquist, let us define a quantity

$$F(\omega_k) = \frac{8 \sin \omega_k/(4s)}{\omega_k/s},$$

so that we may write

$$A_k - iB_k = F(\omega_k)(C_k - iS_k),$$

where

$$C_k = \frac{1}{N} \sum_{h=1}^{N} a_h \cos \frac{2\pi k}{N} \left(h - \frac{1}{2}\right)$$

and

$$S_k = \frac{1}{N} \sum_{h=1}^{N} a_h \sin \frac{2\pi k}{N} \left(h - \frac{1}{2}\right).$$

Since the information in the signal is entirely encoded in the a_h, it follows that this is contained in the terms C_k and S_k. The function $F(\omega_k)$ depends on the shape of the pulse used (in our case a rectangular pulse).

Now consider how the expressions C_k and S_k vary with k. It is not difficult to show that

$$C_{nN+k} = -C_k, \quad S_{nN+k} = -S_k,$$

for n odd and

$$C_{2nN+k} = C_k, \quad S_{2nN+k} = S_k,$$

for n even. It is also true that

$$C_{N-n} = -C_n, \quad S_{N-n} = S_n.$$

This implies that there is no new information contained in the values of C_k and S_k for $k > N/2$. This translates to a condition on the maximum frequency needed to transmit the information. Since $\omega_k = 2\pi k/T$, we have $\omega_{max} = \pi N/T$. This is the Nyquist frequency.

1.3.3 Hartley's Information Capacity

Nyquist [103] wrote down a formula for the speed of transmission of intelligence down a communication channel, S, in terms of the number of current values (i.e. signal levels) m:

$$S = K \log m,$$

where K is a constant. This is derived in the appendix of his paper. He makes the assumption that the transmitted code consists of characters of the same duration and that each character can take m values.