AN INTRODUCTION TO 3D COMPUTER VISION TECHNIQUES AND ALGORITHMS

Bogusław Cyganek
Department of Electronics, AGH University of Science and Technology, Poland

J. Paul Siebert
Department of Computing Science, University of Glasgow, Scotland, UK

WILEY
A John Wiley and Sons, Ltd., Publication

AN INTRODUCTION TO 3D COMPUTER VISION TECHNIQUES AND ALGORITHMS

AN INTRODUCTION TO 3D COMPUTER VISION TECHNIQUES AND ALGORITHMS

Bogusław Cyganek
Department of Electronics, AGH University of Science and Technology, Poland

J. Paul Siebert
Department of Computing Science, University of Glasgow, Scotland, UK

WILEY
A John Wiley and Sons, Ltd., Publication

This edition first published 2009

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Cyganek, Boguslaw.
An introduction to 3D computer vision techniques and algorithms / by Boguslaw Cyganek and J. Paul Siebert.
p. cm.
Includes index.
ISBN 978-0-470-01704-3 (cloth)
1. Computer vision. 2. Three-dimensional imaging. 3. Computer algorithms. I. Siebert, J. Paul. II. Title
TA1634.C94 2008
006.3′7–dc22

2008032205

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

ISBN 978-0-470-01704-3

Set in 10/12pt Times by Aptara Inc., New Delhi, India.
Printed in Great Britain by CPI Antony Rowe, Chippenham, Wiltshire

To Magda, Nadia and Kamil
From Bogusław

To Sabina, Konrad and Gustav
From Paul

Contents

Preface

Recent decades have seen rapidly growing research in many areas of computer science, including computer vision. This comes from the natural interest of researchers as well as demands from industry and society for qualitatively new features to be afforded by computers. One especially desirable capability would be automatic reconstruction and analysis of the surrounding 3D environment and recognition of objects in that space. Effective 3D computer vision methods and implementations would open new possibilities such as automatic navigation of robots and vehicles, scene surveillance and monitoring (which allows automatic recognition of unexpected behaviour of people or other objects, such as cars in everyday traffic), medical reasoning, remote surgery and many, many more.

This book is a result of our long fascination with computers and vision algorithms. It started many years ago as a set of short notes with the only purpose 'to remember this or that' or to have a kind of 'short reference' just for ourselves. However, as this diary grew with the years we decided to make it available to other people. We hope that it was a good decision! It is our hope that this book facilitates access to this enthralling area, especially for students and young researchers. Our intention is to provide a very concise, though as far as possible complete, overview of the basic concepts of 2D and 3D computer vision. However, the best way to get into the field is to try it oneself! Therefore, in parallel with explaining basic concepts, we provide also a basic programming framework with the hope of making this process easier. We greatly encourage the reader to take the next step and try the techniques in practice.

Bogusław Cyganek, Kraków, Poland
J. Paul Siebert, Glasgow, UK

Acknowledgements

We would like to express our gratitude to all the people who helped in the preparation of this book!

In particular, we are indebted to the whole Wiley team who helped in the preparation of the manuscript. In this group special thanks go to Simone Taylor who believed in this project and made it happen. We would also like to express our gratitude to Sian Andrews, Laura Bell, Liz Benson, Emily Bone, Lucy Bryan, Kate Griffiths, Wendy Hunter, Alex King, Erica Peters, Kathryn Sharples, and Nicky Skinner.

We are also very grateful to the individuals and organizations who agreed to the use of their figures in the book. These are Professor Yuichi Ohta from Tsukuba University, as well as Professor Ryszard Szeliski from Microsoft Research. Likewise we would like to thank Dimensional Imaging Ltd. and Precision 3D Ltd. for use of their images. In this respect we would also like to express our gratitude to Springer Science and Business Media, IEEE Computer Society Press, the IET, Emerald Publishing, the ACM, Maney Publishing and Elsevier Science.

We would also like to thank numerous colleagues from the AGH University of Science and Technology in Kraków. We owe a special debt of gratitude to Professor Ryszard Tadeusiewicz and Professor Kazimierz Wiatr, as well as to Lidia Krawentek for their encouragement and continuous support.

We would also like to thank members of the former Turing Institute in Glasgow (Dr Tim Niblett, Joseph Jin, Dr Peter Mowforth, Dr Colin Urquhart and also Arthur van Hoff) as well as members of the Computer Vision and Graphics Group in the Department of Computing Science, University of Glasgow, for access to and use of their research material (Dr John Patterson, Dr Paul Cockshott, Dr Xiangyang Ju, Dr Yijun Xiao, Dr Zhili Mao, Dr Zhifang Mao (posthumously), Dr J.C Nebel, Dr Tim Boyling, Janet Bowman, Susanne Oehler, Stephen Marshall, Don Whiteford and Colin McLaren). Similarly we would like to thank our collaborators in the Glasgow Dental Hospital and School (Professor Khursheed Moos, Professor Ashraf Ayoub and Dr Balvinder Khambay), Canniesburn Plastic Surgery Unit (Mr Arup Ray), Glasgow, the Department of Statistics (Professor Adrian Bowman and Dr Mitchum Bock), Glasgow University, Professor Donald Hadley, Institute of Neurological Sciences, Southern General Hospital, Glasgow, and also those colleagues formerly at the Silsoe Research Institute (Dr Robin Tillett, Dr Nigel McFarlane and Dr Jerry Wu), Silsoe, UK.

Special thanks are due to Dr Sumitha Balasuriya for use of his Matlab codes and graphs. Particular thanks are due to Professor "Keith" van Rijsbergen and Professor Ray Welland without whose support much of the applied research we report would not have been possible.

We wish to express our special thanks and gratitude to Steve Brett from Pandora Inc. for granting rights to access their software platform.

Some parts of the research for which results are provided in this book were possible due to financial support of the European Commission under RACINE-S (IST-2001-37117) and IP-RACINE (IST-2-511316-IP) as well as Polish funds for scientific research in 2007–2008. Research described in these pages has also been funded by the UK DTI and the EPSRC & BBSRC funding councils, the Chief Scientist Office (Scotland), Wellcome Trust, Smith's Charity, the Cleft Lip and Palate Association, the National Lottery (UK) and the Scottish Office. Their support is greatly appreciated.

Finally, we would like to thank Magda and Sabina for their encouragement, patience and understanding over the three-year period it took to write this book.

Notation and Abbreviations

$I_k(x, y)$	Intensity value of a k-th image at a point with local image coordinates (x, y)
$\overline{I_k(x, y)}$	Average intensity value of a k-th image at a point with local image coordinates (x, y)
$\mathbf{I}$	Identity matrix; image treated as a matrix
$\mathbf{P}$	A vector (a point), matrix, tensor, etc.
$T[\mathbf{I}, \mathbf{P}]$	The Census transformation T for a pixel $\mathbf{P}$ in the image $\mathbf{I}$
i, j	Free coordinates
d_x, d_y	Displacements (offset) in the x and y directions
$D(\mathbf{p}_l, \mathbf{p}_r)$	Disparity between points $\mathbf{p}_l$ and $\mathbf{p}_r$
$\mathbf{D}$	Disparity map (a matrix)
$U(x, y)$	Local neighbourhood of pixels around a point (x, y)
$\mathbf{O}_c$	Optical centre point
$\mathbf{P}_c = [X_c, Y_c, Z_c]^T$	Coordinates of a 3D point in the camera coordinate system
$\mathbf{\Pi}$	Camera plane; a projective plane
$\mathbf{o} = (o_x, o_y)$	Central point of a camera plane
f	Focus length of a camera
b	Base line in a stereo system (a distance between cameras)
h_x, h_y	Physical horizontal and vertical dimensions of a pixel
$\mathbf{P} = [X, Y, Z]^T$	3D point and its coordinates
$\wp^n$	N-dimensional projective space
$\mathbf{P} = [X_h, Y_h, Z_h, 1]^T$	Homogenous coordinates of a point
$\mathbf{M}$	Camera matrix
$\mathbf{M}_\mathrm{i}$	Intrinsic parameters of a camera
$\mathbf{M}_\mathrm{e}$	Extrinsic parameters of a camera
$\mathbf{E}$	Essential matrix.
$\mathbf{F}$	Fundamental matrix.
$\mathbf{e}_i$	Epipole in an i-th image
SAD	Sum of absolute differences
SSD	Sum of squared differences
ZSAD	Zero-mean sum of absolute differences
ZSSD	Zero-mean sum of squared differences
ZSSD-N	Zero-mean sum of squared differences, normalized

SCP	Sum of cross products
SCP-N	Sum of cross products, normalized
RMS	Root mean square
RMSE	Root mean square error
<Lxx, Lyy>	Code lines from a line *Lxx* to *Lyy*
HVS	Human Visual System
SDK	Software Development Kit
$\wedge$	logical 'and'
$\vee$	logical 'or'
LRC	Left-right checking (cross-checking)
OCC	Occlusion constraint
ORD	Point ordering constraint
BMD	Bimodality rule
MGJ	Match goodness jumps
NM	Null method
GT RMS	Ground-truth RMS
WTA	Winner-takes-all
*	Convolution operator

Part I

1

Introduction

The purpose of this text on stereo-based imaging is twofold: it is to give students of computer vision a thorough grounding in the image analysis and projective geometry techniques relevant to the task of recovering three-dimensional (3D) surfaces from stereo-pair images; and to provide a complete reference text for professional researchers in the field of computer vision that encompasses the fundamental mathematics and algorithms that have been applied and developed to allow 3D vision systems to be constructed.

Prior to reviewing the contents of this text, we shall set the context of this book in terms of the underlying objectives and the explanation and design of 3D vision systems. We shall also consider briefly the historical context of optics and vision research that has led to our contemporary understanding of 3D vision.

Here we are specifically considering 3D vision systems that base their operation on acquiring stereo-pair images of a scene and then decoding the depth information *implicitly* captured within the stereo-pair as parallaxes, i.e. relative displacements of the contents of one of the images of the stereo-pair with respect to the other image. This process is termed *stereo-photogrammetry*, i.e. measurement from stereo-pair images. For readers with normal functional binocular vision, the everyday experience of observing the world with both of our eyes results in the perception of the relative distance (depth) to points on the surfaces of objects that enter our field of view. For over a hundred years it has been possible to configure a stereo-pair of cameras to capture stereo-pair images, in a manner analogous to mammalian binocular vision, and thereafter view the developed photographs to observe a miniature 3D scene by means of a stereoscope device (used to present the left and right images of the captured stereo-pair of photographs to the appropriate eye). However, in this scenario it is the brain of the observer that must decode the depth information locked within the stereo-pair and thereby experience the perception of depth. In contrast, in this book we shall present underlying mechanisms by which a computer program can be devised to analyse digitally formatted images captured by a stereo-pair of cameras and thereby recover an *explicit* measurement of distances to points *sampling* surfaces in the imaged field of view. Only by explicitly recovering depth estimates does it become possible to undertake useful tasks such as 3D measurement or reverse engineering of object surfaces as elaborated below. While the science of stereo-photogrammetry is a well-established field and it has indeed been possible to undertake 3D

An Introduction to 3D Computer Vision Techniques and Algorithms Bogusław Cyganek and J. Paul Siebert

measurement by means of stereo-pair images using a manually operated measurement device (the stereo-comparator) since the beginning of the twentieth century, we present fully automatic approaches for 3D imaging and measurement in this text.

1.1 Stereo-pair Images and Depth Perception

To appreciate the structure of 3D vision systems based on processing stereo-pair images, it is first necessary to grasp, at least in outline, the most basic principles involved in the formation of stereo-pair images and their subsequent analysis. As outlined above, when we observe a scene with both eyes, an image of the scene is formed on the retina of each eye. However, since our eyes are horizontally displaced with respect to each other, the images thus formed are not identical. In fact this stereo-pair of retinal images contains slight displacements between the relative locations of local parts of the image of the scene with respect to each image of the pair, depending upon how close these local scene components are to the point of *fixation* of the observer's eyes. Accordingly, it is possible to reverse this process and deduce how far away scene components were from the observer according to the magnitude and direction of the parallaxes within the stereo-pairs when they were captured. In order to accomplish this task two things must be determined: firstly, those local parts of one image of the stereo-pair that match the corresponding parts in the other image of the stereo-pair, in order to find the local parallaxes; secondly, the precise geometric properties and configuration of the eyes, or cameras. Accordingly, a process of *calibration* is required to discover the requisite geometric information to allow the imaging process to be inverted and relative distances to surfaces observed in the stereo-pair to be recovered.

1.2 3D Vision Systems

By definition, a stereo-photogrammetry-based 3D vision system will require stereo-pair image acquisition hardware, usually connected to a computer hosting software that automates acquisition control. Multiple stereo-pairs of cameras might be employed to allow all-round coverage of an object or person, e.g. in the context of whole-body scanners. Alternatively, the object to be imaged could be mounted on a computer-controlled turntable and overlapping stereo-pairs captured from a fixed viewpoint for different turntable positions. Accordingly, sequencing capture and image download from multiple cameras can be a complex process, and hence the need for a computer to automate this process.

The stereo-pair acquisition process falls into two categories, active illumination and passive illumination. Active illumination implies that some form of pattern is projected on to the scene to facilitate finding and disambiguating parallaxes (also termed *correspondences* or *disparities*) between the stereo-pair images. Projected patterns often comprise grids or stripes and sometimes these are even colour coded. In an alternative approach, a random speckle texture pattern is projected on to the scene in order to augment the texture already present on imaged surfaces. Speckle projection can also guarantee that that imaged surfaces appear to be randomly textured and are therefore locally uniquely distinguishable and hence able to be matched successfully using certain classes of image matching algorithm. With the advent of 'high-resolution' digital cameras the need for pattern projection has been reduced, since the surface texture naturally present on materials, having even a matte finish, can serve to facilitate

matching stereo-pairs. For example, stereo-pair images of the human face and body can be matched successfully using ordinary studio flash illumination when the pixel sampling density is sufficient to resolve the natural texture of the skin, e.g. skin-pores. A camera resolution of approximately 8–13M pixels is adequate for stereo-pair capture of an area corresponding to the adult face or half-torso.

The acquisition computer may also host the principal 3D vision software components:

- An image matching algorithm to find correspondences between the stereo-pairs.
- Photogrammetry software that will perform system calibration to recover the geometric configuration of the acquisition cameras and perform 3D point reconstruction in world coordinates.
- 3D surface reconstruction software that builds complete manifolds from 3D *point-clouds* captured by each imaging stereo-pair.

3D visualisation facilities are usually also provided to allow the reconstructed surfaces to be displayed, often *draped* with an image to provide a *photorealistic* surface model. At this stage the 3D shape and surface appearance of the imaged object or scene has been captured in explicit digital metric form, ready to feed some subsequent application as described below.

1.3 3D Vision Applications

This book has been motivated in part by the need to provide a manual of techniques to serve the needs of the computer vision practitioner who wishes to construct 3D imaging systems configured to meet the needs of practical applications. A wide variety of applications are now emerging which rely on the fast, efficient and low-cost capture of 3D surface information. The traditional role for image-based 3D surface measurement has been the reserve of *close-range* photogrammetry systems, capable of recovering surface measurements from objects in the range of a few tens of millimetres to a few metres in size. A typical example of a classical close-range photogrammetry task might comprise surface measurement for manufacturing quality control, applied to high-precision engineered products such as aircraft wings.

Close-range video-based photogrammetry, having a lower spatial resolution than traditional plate-camera film-based systems, initially found a niche in imaging the human face and body for clinical and creative media applications. 3D clinical photographs have the potential to provide quantitative measurements that reduce subjectivity in assessing the surface anatomy of a patient (or animal) before and after surgical intervention by providing numeric, possibly automated, scores for the shape, symmetry and longitudinal change of anatomic structures. Creative media applications include whole-body 3D imaging to support creation of human avatars of specific individuals, for 3D gaming and cine special effects requiring virtual actors. Clothing applications include body or foot scanning for the production of custom clothing and shoes or as a means of sizing customers accurately. An innovative commercial application comprises a ‘virtual catwalk’ to allow customers to visualize themselves in clothing prior to purchasing such goods on-line via the Internet.

There are very many more emerging uses for 3D imaging beyond the above and commercial ‘reverse engineering’ of premanufactured goods. 3D vision systems have the potential to revolutionize autonomous vehicles and the capabilities of robot vision systems. Stereo-pair cameras could be mounted on a vehicle to facilitate autonomous navigation or configured

within a robot workcell to endow a 'blind' pick-and-place robot, both object recognition capabilities based on 3D cues and simultaneously 3D spatial quantification of object locations in the workspace.

1.4 Contents Overview: The 3D Vision Task in Stages

The organization of this book reflects the underlying principles, structural components and uses of 3D vision systems as outlined above, starting with a brief historical view of vision research in Chapter 2. We deal with the basic existence proof that binocular 3D vision is possible, in an overview of the human visual system in Chapter 3. The basic projective geometry techniques that underpin 3D vision systems are also covered here, including the geometry of monocular and binocular image formation which relates how binocular parallaxes are produced in stereo-pair images as a result of imaging scenes containing variation in depth. Camera calibration techniques are also presented in Chapter 3, completing the introduction of the role of image formation and geometry in the context of 3D vision systems.

We deal with fundamental 2D image analysis techniques required to undertake image filtering and feature detection and localization in Chapter 4. These topics serve as a precursor to perform image matching, the process of detecting and quantifying parallaxes between stereo-pair images, a prerequisite to recovering depth information. In Chapter 5 the issue of spatial scale in images is explored, namely how to structure algorithms capable of efficiently processing images containing structures of varying scales which are unknown in advance. Here the concept of an image *scale-space* and the *multi-resolution image pyramid* data structure is presented, analysed and explored as a precursor to developing matching algorithms capable of operating over a wide range of visual scales. The core algorithmic issues associated with stereo-pair image matching are contained in Chapter 6 dealing with distance measures for comparing image patches, the associated parametric issues for matching and an in-depth analysis of area-based matching over scale-space within a practical matching algorithm. Feature-based approaches to matching are also considered and their combination with area-based approaches. Then two solutions to the stereo problem are discussed: the first, based on the *dynamic programming*, and the second one based on the *graph cuts* method. The chapter ends with discussion of the *optical flow* methods which allow estimation of local displacements in a sequence of images.

Having dealt with the recovery of disparities between stereo-pairs, we progress logically to the recovery of 3D surface information in Chapter 7. We consider the process of *triangulation* whereby 3D points in world coordinates are computed from the disparities recovered in the previous chapter. These 3D points can then be organized into surfaces represented by *polygonal meshes* and the *3D point-clouds* recovered from *multi-view* systems acquiring more than one stereo-pair of the scene can be fused into a coherent surface model either directly or via volumetric techniques such as *marching cubes*. In Chapter 8 we conclude the progression from theory to practice, with a number of case examples of 3D vision applications covering areas such as face and body imaging for clinical, veterinary and creative media applications and also 3D vision as a visual prosthetic. An application based only on image matching is also presented that utilizes motion-induced inter-frame disparities within a cine sequence to synthesize missing or damaged frames, or sets of frames, in digitized historic archive footage.

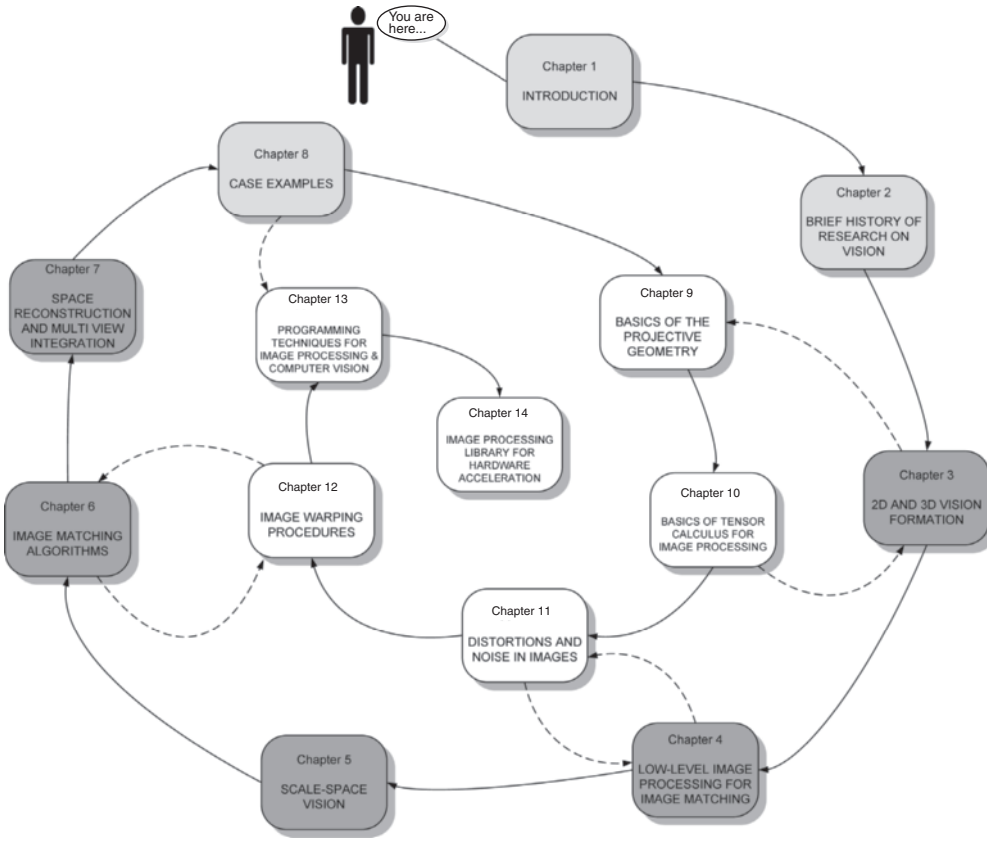

Figure 1.1 Organization of the book

The remaining chapters provide a series of detailed technical tutorials on projective geometry, tensor calculus, image warping procedures and image noise. A chapter on programming techniques for image processing provides practical hints and advice for persons who wish to develop their own computer vision applications. Methods of object oriented programming, such as design patterns, but also proper organization and verification of the code are discussed. Chapter 14 outlines the software presented in the book and provides the link to the recent version of the code.

Figure 1.1 depicts possible order of reading the book. All chapters can be read in number order or selectively as references to specific topics. There are five main chapters (Chapters 3–7), three auxiliary chapters (Chapters 1, 2 and 8) as well as five technical tutorials (Chapters 9–13). The latter are intended to aid understanding of specific topics and can be read in conjunction with the related main chapters, as indicated by the dashed lines in Figure 1.1.

2

Brief History of Research on Vision

2.1 Abstract

This chapter is a brief retrospective on vision in art and science. 3D vision and perspective phenomena were first studied by the architects and artists of Ancient Greece. From this region and time comes *The Elements* by Euclid, a treatise that paved the way for geometry and mathematics. Perspective techniques were later applied by many painters to produce the illusion of depth in flat paintings. However, called an 'evil trick', it was denounced by the Inquisition in medieval times. The blooming of art and science came in the Renaissance, an era of Leonardo da Vinci, perhaps the most ingenious artist, scientist and engineer of all times. He is attributed with the invention of the camera obscura, a prototype of modern cameras, which helped to acquire images of a 3D scene on a flat plane. Then, on the 'shoulders of giants' came another 'giant', Sir Isaac Newton, whose *Opticks* laid the foundation for modern physics and also the science of vision. These and other events from the history of research on vision are briefly discussed in this chapter.

2.2 Retrospective of Vision Research

The first people known to have investigated the phenomenon of depth perception were the Ancient Greeks [201]. Probably the first writing on the subject of disparity comes from Aristotle (380 BC) who observed that, if during a prolonged observation of an object one of the eyeballs is pressed with a finger, the object is experienced in double vision.

The earliest known book on optics is a work by Euclid entitled *The Thirteen Books of the Elements* written in Alexandria in about 300 BC [116]. Most of the definitions and postulates of his work constitute the foundations of mathematics since his time. Euclid's works paved the way for further progress in optics and physiology, as well as inspiring many researchers over the following centuries. At about the same time as Euclid was writing, the anatomical structure of human organs, including the eyes, was examined by Herofilus from Alexandria. Subsequently Ptolemy, who lived four centuries after Euclid, continued to work on optics.

Many centuries later Galen (AD 180) who had been influenced by Herofilus' works, published his own work on human sight. For the first time he formulated the notion of the *cyclopean* eye, which 'sees' or visualizes the world from a common point of intersection within the

An Introduction to 3D Computer Vision Techniques and Algorithms Bogusław Cyganek and J. Paul Siebert

optical nervous pathway that originates from each of the eyeballs and is located perceptually at an intermediate position between the eyes. He also introduced the notion of parallax and described the process of creating a single view of an object constructed from the binocular views originating from the eyes.

The works of Euclid and Galen contributed significantly to progress in the area of optics and human sight. Their research was continued by the Arabic scientist Alhazen, who lived around AD 1000 in the lands of contemporary Egypt. He investigated the phenomena of light reflection and refraction, now fundamental concepts in modern geometrical optics.

Based on Galen's investigations into anatomy, Alhazen compared an eye to a dark chamber into which light enters via a tiny hole, thereby creating an inverted image on an opposite wall. This is the first reported description of the *camera obscura*, or the pin-hole camera model, an invention usually attributed to Roger Bacon or Leonardo da Vinci. A device called the camera obscura found application in painting, starting from Giovanni Battista della Porta in the sixteenth century, and was used by many masters such as Antonio Canal (known as Canaletto) or Bernaldo Bellotto. A painting by Canaletto, entitled *Perspective*, is shown in Figure 2.1. Indeed, his great knowledge of basic physical properties of light and projective

Figure 2.1 *Perspective* by Antonio Canal (Plate 1). (1765, oil on canvas, Gallerie dell'Accademia, Venice)

Figure 2.2 Painting by Bernardo Bellotto entitled *View of Warsaw from the Royal Palace* (Plate 2). (1773, oil on canvas, National Museum, Warsaw)

geometry allowed him to reach mastery in paintings. His paintings are very realistic which was a very desirable skill of a painter, since we have to remember that these were times when people did not yet know of photography.

Figure 2.2 shows a view of eighteenth-century Warsaw, the capital of Poland, painted by Bernaldo Bellotto in 1773. Just after, due to invasion of the three neighbouring countries, Poland disappeared from maps for over a century.

Albrecht Dürer was one of the first non-Italian artists who used principles of geometrical perspective in his art. His famous drawing *Draughtsman Drawing a Recumbent Woman* is shown in Figure 2.3.

However, the contribution of Leonardo da Vinci cannot be overestimated. One of his famous observations is that a light passing through a small hole in the camera obscura allows the

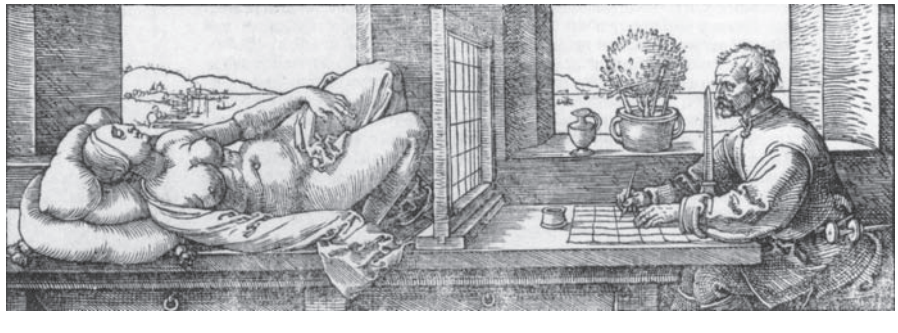

Figure 2.3 A drawing by Albrecht Dürer entitled *Draughtsman Drawing a Recumbent Woman*. (1525, woodcut, Graphische Sammlung Albertina, Vienna)

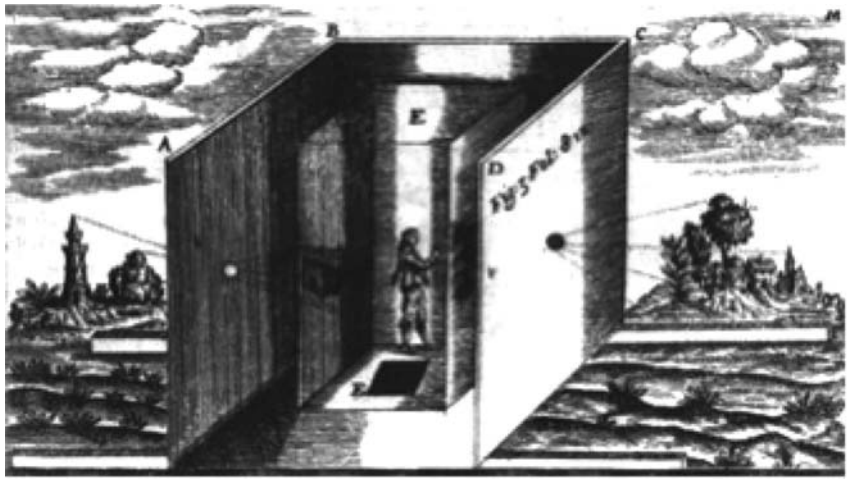

Figure 2.4 Drawing of the camera obscura from the work of the Jesuit Athanasius Kircher, around 1646

observation of all surrounding objects. From this he concluded that light rays passing through different objects cross each other in any point from which they are visible. This observation suggests also the wave nature of light, rather than light comprising a flow of separate particles as was believed by the Ancient Greeks. Da Vinci's unquestionable accomplishment in the area of stereoscopic vision is his analysis of partial and total occlusions, presented in his treatise entitled *Trattato della Pittura*. Today we know that these phenomena play an important role in the human visual system (HVS), facilitating correct perception of depth [7] (section 3.2).

Other accomplishments were made in Europe by da Vinci's contemporaries. For instance in 1270 Vitello, who lived in Poland, published a treatise on optics entitled *Perspectiva*, which was the first of its kind. Interestingly, from almost the same time comes a note on the first binoculars, manufactured probably in the glassworks of Pisa.

Figure 2.4 depicts a drawing of a camera obscura by the Jesuit Athanasius Kircher, who lived in the seventeenth century.

In the seventeenth century, based on the work of Euclid and Alhazen, Kepler and Descartes made further discoveries during their research on the HVS. In particular, they made great contributions towards understanding of the role of the retina and the optic nerve in the HVS.

More or less at the same time, i.e. the end of the sixteenth and beginning of the seventeenth centuries, the Jesuit Francois D'Aguillon made a remarkable synthesis of contemporary knowledge on optics and the works of Euclid, Alhazen, Vitello and Bacon. In the published treatise *Opticorum Libri Sex*, consisting of six books, D'Aguillon analysed visual phenomena and in particular the role of the two eyes in this process. After defining the locale of visual convergence of the two eyeballs, which he called the horopter, D'Aguillon came close to formulating the principles of stereovision which we still use today.

A real breakthrough in science can be attributed to Sir Isaac Newton who, at the beginning of the eighteenth century, published his work entitled *Opticks* [329]. As first, he correctly described a way of information passing from the eyes to the brain. He discovered that visual

sensations from the "inner" hemifields of the retina (the mammalian visual field is split along the vertical meridian in each retina), closest to the nose, are sent through the optic nerves directly to the corresponding cerebral hemispheres (cortical lobes), whereas sensations coming from the "outer" hemifields, closest to the temples, are crossed and sent to the opposite hemispheres. (The right eye, right hemifield and left eye, left hemifield cross, while the left eye, right hemifield and the right eye, left hemifield do not cross.) Further discoveries in this area were made in the nineteenth century not only thanks to researchers such as Heinrich Müller and Bernhard von Gudden, but also thanks to the invention of the microscope and developments in the field of medicine, especially physiology.

In 1818 Vieth made a precise explanation of the horopter, being a spherical placement of objects which cause a focused image on the retina, a concept that was already familiar to D'Aguillon. At the same time this observation was reported by Johannes Müller, and therefore the horopter is termed the Vieth–Müller circle.

In 1828 a professor of physics of the Royal Academy in London, Sir Charles Wheatstone, formulated the principles underlying stereoscopic vision. He also presented a device called a *stereoscope* for depth perception from two images. This launched further observations and discoveries; for instance, if the observed images are reversed, then the perception of depth is also reversed. Inspired by Wheatstone's stereoscope, in 1849 Sir David Brewster built his version of the stereoscope based on a prism (Figure 2.5), and in 1856 he published his work on the principles of stereoscopy [56].

The inventions of Wheatstone and Brewster sparked an increased interest in three-dimensional display methods, which continues with even greater intensity today due to the invention of the random dot autostereograms, as well as the rapid development of personal computers. Random dot stereograms were analysed by Bela Julesz who in 1960 showed that

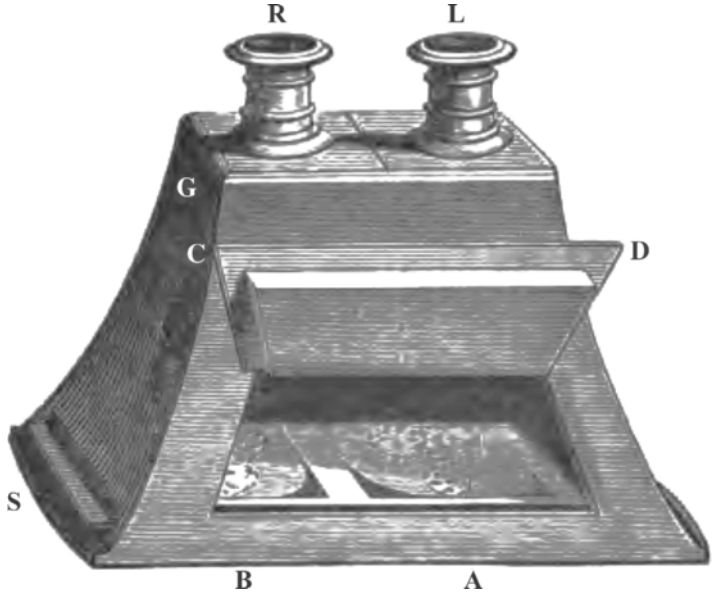

Figure 2.5 Brewster's stereoscope (from [56])

depth can be perceived by humans from stereo-pairs of images comprising only random dots (the dots being located with relative shifts between the images forming the stereo-pair) and no other visible features such as corners or edges.

Recent work reported by the neurophysiologists Bishop and Pettigrew showed that in primates special cells, which react to disparity signals built from images formed on two retinas of the eyes, are already present in the input layer (visual area 1, V1) of the visual cortex. This indicates that depth information is processed even earlier in the visual pathway than had been thought.

2.3 Closure

In this chapter we have presented a very short overview of the history of studies on vision in art and science. It is a very wide subject which could have merited a separate book by itself. Nevertheless, we have tried to point out those, in our opinion, important events that paved the way for contemporary knowledge on vision research, which also inspired us to write this book. Throughout the centuries, art and science were interspersed and influenced each other. An example of this is the camera obscura which, first devised by artists, after centuries became a prototype of modern cameras. These are used to acquire digital images, then processed with vision algorithms to infer knowledge on the surrounding environment, for instance. Further information on these fascinating issues can be found in many publications, some of which we mention in the next section.

2.3.1 Further Reading

There are many sources of information on the history of vision research and photography. For instance the Bright Bytes Studio web page [204] provides much information on camera obscuras, stereo photography and history. The Web Gallery of Art [214] provides an enormous number of paintings by masters from past centuries. The book by Brewster mentioned earlier in the chapter can also be obtained from the Internet [56]. Finally, Wikipedia [215] offers a wealth of information in many different languages on most of the subjects, including paintings, computer vision and photography.

Part II

3

2D and 3D Vision Formation

3.1 Abstract

This chapter is devoted mainly to answering the question: "What is the difference between having one image of a scene, compared to having two images of the same scene taken from different viewpoints?" It appears that in the second case the difference is a fundamental one: with two (or more) views of the same scene, taken however at different camera positions, we can infer depth information by means of geometry: three-dimensional (3D) information can be recovered through a process known as *triangulation*. This is why having two eyes makes a difference.

We start with a brief overview of what we know about the human visual system which is an excellent example of precision and versatility. Then we discuss the image acquisition process using a single camera. The main concept here is the simple pin-hole camera model which is used to explain the transformation from 3D world-space to the 2D imaging-plane as performed by a camera. The so-called extrinsic and intrinsic parameters of a camera are introduced next. When images of a scene are captured using two cameras simultaneously, these cameras are termed a *stereo-pair* and produce stereo-pairs of images. The properties of cameras so configured are determined by their *epipolar geometry*, which tells us the relationship between world points observed in their fields of view and the images impinging on their respective sensing planes. The image-plane locations of each world point, as sensed by the camera pair, are called corresponding or matched points. Corresponding points within stereo-pair images are connected by the fundamental matrix. If known, it provides fundamental information on the epipolar geometry of the stereo-pair setup. However, finding corresponding points between images is not a trivial task. There are many factors which can confound this process, such as occlusions, limited image resolution and quantization, distortions, noise and many others. Technically, matching is said to be *under constrained*: there is not sufficient information available within the compared images to guarantee finding a unique match. However, matching *can* be made easier by applying a set of rules known as *stereo constraints*, of which the most important is the *epipolar constraint*, and this implies that corresponding points always lie on corresponding epipolar lines. The epipolar constraint limits the search for corresponding points from the entire 2D space to a 1D space of epipolar lines. Although the positions of the epipolar lines are not known in advance, in the special case when stereo-pair cameras are

An Introduction to 3D Computer Vision Techniques and Algorithms Bogusław Cyganek and J. Paul Siebert

configured with parallel optical axes – called the canonical, fronto-parallel, or standard stereo system – the epipolar lines follow the image (horizontal) scan-lines. The problem of finding corresponding points is therefore one of the essential tasks of computer vision.

It appears that by means of point correspondences the extrinsic and intrinsic parameters of a camera can be determined. This is called camera calibration and is also discussed in this chapter. We conclude with a discussion of a practical implementation of the presented concepts, with data structures to represent images and some C++ code examples which come from the image library provided with this book.

3.2 Human Visual System

Millions of years of evolution have formed the human visual system (HVS) and within it the most exquisite, unattainable and mysterious stereoscopic depth perception engine on planet Earth. The vision process starts in the eye, a diagram of which is depicted in Figure 3.1.

Incident light at first passes through the pupil which controls the amount of light passing to the lens of the eye. The size of the pupil aperture is controlled by the iris pupilliary sphincter muscles. The larger this aperture becomes, the larger the spherical aberration and smaller the depth of focus of the eye. The visual axis joins a point of fixation and the fovea. Although an eye is not rotationally symmetric, an approximate optical axis can be defined as a line joining the centre of curvature of the cornea and centre of the lens. The angle between the two axes is about 5°. It should be noted that the eye itself is not a separate organ but a 150 mm extension of the brain. In the context of computer vision, the most important part of the eye is the retina which is the place of exchange that converts an incoming stream of photons into corresponding neural excitations.

In the context of binocular vision and stereoscopic perception of depth, it is important that the eyes are brought into convergence such that the same scene region is projected onto the respective foveae of each eye. Figure 3.2 presents a model of binocular vision: an image of a certain point H is created in the two eyes, exactly in the centres of their foveae.

On each retina images of the surrounding 3D points are also created. We mark the distance of those images in respect to the corresponding fovea. Under this assumption, the two image points on each of the retinas are *corresponding* when their *distances* to their corresponding

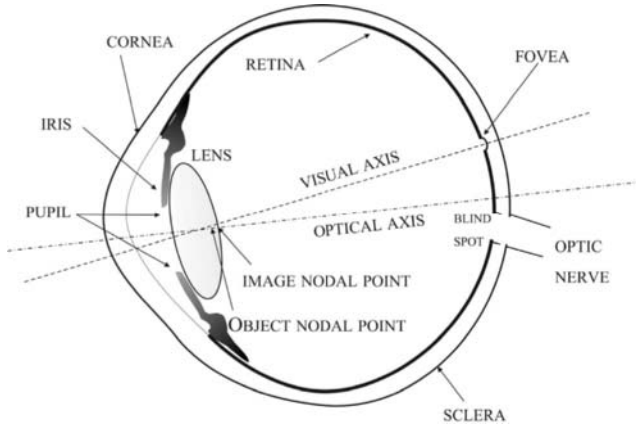

Figure 3.1 Schematic of a human eye

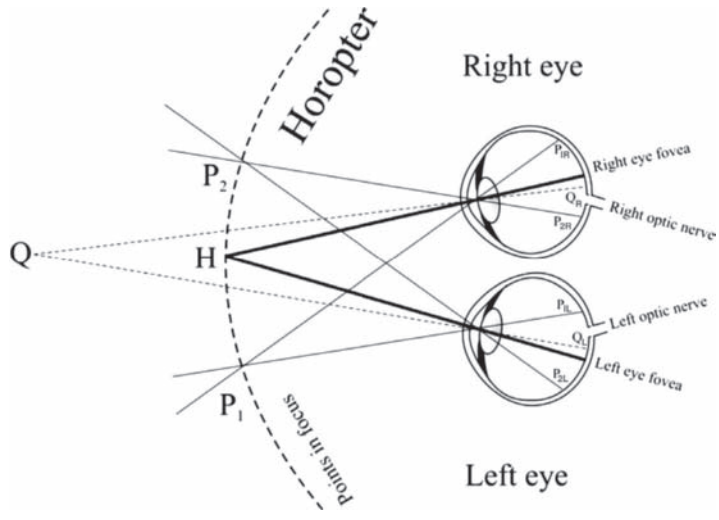

Figure 3.2 Disparity on the retina of an eye. The horopter is denoted by a broken line. H is a point of fixation

foveae are the *same*. In Figure 3.2 this condition is fulfilled for the points P_1 and P_2, but not for Q. That is, the distances P_{1R} and P_{1L} are the same. This holds also for P_{2R} and P_{2L} but not for the Q_R and Q_L which are in opposite directions from the foveae. However, the latter property allows the HVS to conclude that Q is further from the horopter. Conducting now the reverse reasoning, i.e. looking for 3D points such that their retinal images are the same distance from the two foveae, we find the 3D region known as the *horopter*. Retinal images of all points other than those belonging to the horopter are said to be non-corresponding. The relative difference in distance from the fovea for of each these non-corresponding points is termed *retinal disparity* [201, 442]. It is evident now that the horopter points have zero retinal disparity. The retinal disparity is used by the HVS to assess distance to 3D locations in the world.

The signals induced on the fovea are transferred to the input of the primary visual cortex of the brain, labelled by neuro-anatomists as Visual Area 1 (V1). This area of the visual cortex is the first location in the entire structure where individual neurons receive binocular input. It was also discovered that some neurons in V1 respond exclusively to mutual excitations from the two eyes. Those neurons, called disparity detectors, are sensitive to stereoscopic stimuli [442].

In addition, the relationship between the *firing rates* of these disparity detecting neurons, measured in units of *impulses per second*, and input retinal disparity is called the disparity-tuning function. It has an evident maximum for zero retinal disparity (i.e. it is "tuned" to respond best to zero disparity), that is for 3D points lying on the horopter [201].

Many experiments have been conducted to achieve a better understanding of the stereoscopic processes in the HVS. A phenomenon first noticed during such research was the influence of luminance variation on the process of associating corresponding visual stimuli from each eye, i.e. disparity detection. In the simplest case this concerns the detection, i.e. correlation, of corresponding image edges in each retina, while correlation of corresponding textured areas is more complex. In 1979 Marr and Poggio [299] put forward a theory that stereoscopic matching relies on the correlation of retinal image locations in which the second derivative of the luminance signal is crossing a zero value; these are the so-called zero-crossings.

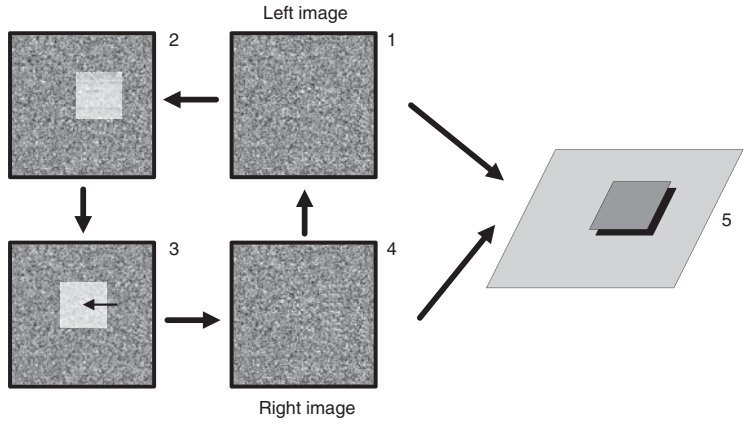

Figure 3.3 Construction of a random dot stereogram: (1) left image; (2) extracted region in the left image; (3) shift of this region; (4) right image; (5) depth effect when observed by two eyes

Zero-crossings corresponds to those regions in an image that exhibit the greatest change in the signal instead of the greatest absolute value of the signal itself. Further research undertaken by Mayhew and Frisby [302] showed that stereoscopic correlation in the HVS does not depend exclusively on the zero-crossings but on a more generalised matching mechanism applied to the spectral components of the two-dimensional luminance signal. Mallot *et al.* [291] revealed the possibility of a secondary correlation mechanism being invoked when the luminance signal is changing very slowly. Based on these results it can be stated that the HVS prefers to correlate more general features, if available, in the image. This relates correlation based on zero-crossings and also correlation based on signal value maxima. However, correlation based on matching spectral components of the luminance signal dominates when these are the most distinctive features found in the images. When there are neither significant zero-crossings nor other signal differences, the HVS is capable of estimating disparity values based on correlating the maximal values of the low-pass components of the luminance signal.

A qualitatively new development was reported by Julesz in 1960 [235] when he demonstrated the so-called random dot stereogram.[1] A random dot stereogram comprises a stereo-pair of images in which the first image of the pair is created by generating a field of random points. The second image of the stereo-pair is generated by copying the first image and then selecting and displacing by a small amount a specific region within the copy. Figure 3.3 outlines steps of this construction. Table 3.4 (page 62) contains another example of a random dot stereogram. When constructing random stereograms the random dots can be substituted by random lines [201].

When observed by two eyes, the random dot stereogram allows perception of depth, as seen in Figure 3.3 in a form of a rectangle closer to the observer. Further research on this subject has shown that the stereoscopic effect is attained even if one of the random images

[1] This type of stereogram was already known, however, among artists.

is disturbed, e.g. by adding some spurious dots or by low-pass filtering. On the other hand, a change of luminance polarity (i.e. light and dark regions are exchanged in one image of the stereogram) leads to a loss of the stereo effect.

Research on depth perception based exclusively on a perception of colours has shown that colour information also affects this process to a limited degree [201].

It has been discovered that the stereo correlation process depends also on other factors, leading to a theory that predicts that those compared locations which conform in size, shape, colour, and motion are more privileged during stereo matching. It would also explain why it takes more time for the HVS to match random dot stereograms which do not possess such features. This theory can also be interpreted in the domain of computational stereo matching methods: if a certain local operator can gather enough information in a given neighbourhood of pixels, such as local frequency, orientation or phase, then subsequent matching can be performed more reliably and possibly faster than would otherwise be possible when such information is missing. This rather heuristic rule can be justified by experiment. An example of a tensor operator that quantifies local image structure is presented in section 4.6.

Another known stereo matching constraint adopted by the HVS is so-called most related image matching. It implies that if there is a choice, an image or an image sub-region is considered to be 'matched' if it gives the highest number of meaningful matches. Otherwise the preferred image is one which contains the highest number of space point projections. Due to this strategy, the HVS favours those images, or their sub-regions, that are potentially the most interesting to an observer, since they are closest to him or her.

Yet another constraint discovered by Julesz [235], is the disparity gradient limit. This concept, explained in more detail in section 3.5, is very often used in computer image matching.

Other constraints are based on experience acquired from daily observations of the surrounding space. One of which is that the daily environment usually is moderately 'dense', since we have to move in it somehow. A similar observation indicates that surrounding objects are not transparent either. From these observations we can draw other matching constraints based on: surface continuity, figural smoothness, matching point ordering and matching point uniqueness (section 3.5). Their function in and influence on the HVS, although indicated by many experiments, have not yet been completely explained.

Yet another phenomenon plays an important role in both human and machine stereovision, namely that of occlusions which are explained in Figure 3.4.

How partial occlusions of observed objects influence their binocular perception was investigated by Leonardo Da Vinci [93]. Recent work by Anderson indicates that the occlusion phenomenon has a major influence on the stereovision perception process [7]. The area visible exclusively to the left eye is called the left visible area. Similarly for the right eye we get the right visible area. In Figure 3.4 these areas are marked in light grey. The area observable to both eyes simultaneously can be perceived in full stereo vision. In contrast, the dark area to the left of the object in Figure 3.4 presents a totally occluded location to both eyes. Far beyond the object there is again an area visible to both eyes, so effectively an object does not occlude the whole space behind it, only a part. It is also known and easily verified that the half-occluded regions seen by the right eye falls close to the right edge of the occluding object. Similarly, the half-occluded regions seen by the left eye fall near the left edge of such an occluding object. This situation is portrayed in Figure 3.4.

The effect of partial occlusions is inevitably connected with a break in the smoothness (continuity) of a perceived surface in depth. Thus, due to the presence of partial occlusions,

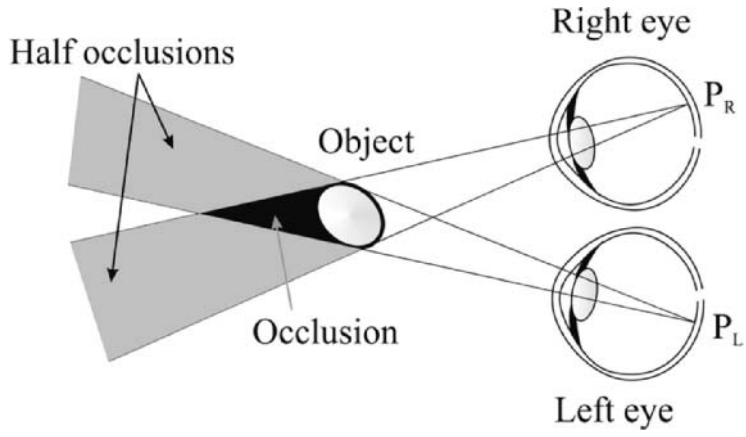

Figure 3.4 Phenomenon of occlusions. Partial occlusions are lighter. The dark area remains totally occluded by an object

it is possible to distinguish depth discontinuities from gradually changing surfaces which, in turn, are limited by the maximum allowable disparity gradient. These and other facts show that the HVS actively decomposes vertical and horizontal image parallaxes into disparities and half-occlusions [7]. They form two complementary sources of visual information. Retinal disparities provide information about the relative depth of observed surfaces visible to both eyes simultaneously. On the other hand, partial-occlusions which are visible to each eye separately, give sufficient data for segmentation of the observable scene into coherent objects at object boundaries.

It is interesting to mention that also the gradient of the *vertical* disparity can be used to infer distance from observed objects, as has been shown by Mayhew and Longuet-Higgins [303] and discussed also by Brenner *et al*. [55]. However, recent psychophysical experiments have indicated that such information is not used by the HVS. Indeed, vertical image differences are not always vertical parallaxes. Sometimes they are caused by half-occlusions. Based on these observations and psychophysical experiments, Anderson [7] suggests that interocular differences in vertical position can influence stereoscopic depth perceived by the HVS by signalling the presence of occluding contours.

Depth perception by the HVS is not only induced purely by stereovision mechanisms, it is also supported by the phenomena of head and eye movements, as well as by motion parallax.

Many psychophysical experiments lead to the observation that there is continuous rivalry between the different vision cues that impinge on the HVS. Then the HVS detects such objects that arise from maxima in the density of goodmatches, when simultaneously in agreement with daily experience.

Depth information acquired by the HVS, as well as other visual cues such as information on colour, edges, shadows and occlusions are only ingredients gathered by the brain to generate inferences about the world. How these visual inferences are then integrated and interpreted into a unified percept is still not known, although hypotheses and models have been proposed by researchers. Knowledge of the function of the visual system has been garnered indirectly by means of observations of two different sets of phenomena known from medicine

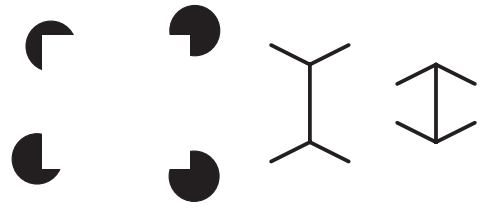

Figure 3.5 Visual illusions. (a) An artificial rectangle is clearly visible although not drawn directly. (b) The two vertical lines are exactly the same length (which can be verified with a ruler), although the left one is perceived to be longer

and psychophysiology. The first set of phenomena are described in case studies that record sight diseases and their subsequent cures. It was clinically observed that those patients who were visually impaired from birth and then had their ability to perceive visual sensations restored, had difficulties learning how to perceive objects and how to interpret scenes, although they can easily detect basic features [201, 442]. Indirectly this provides us with some insight into the conceptual stages and complexity of the seeing mechanisms of our brains.

Visual illusions comprise a second set of phenomena that help us understand how the visual pathways translate retinal images into the perception of objects. There are many illusions that trick our visual system by providing visual cues that do not agree with the physics of the 3D world learned by daily experience [125, 161, 360]. Two simple illusions apparently related to the human perception of depth are presented in Figure 3.5. The first example (Figure 3.5(a)) illustrates the role of occlusions in visual perception. Our acquired experience on transparency of objects makes us perceive an illusory figure whose existence is only cued (i.e. made apparent) by the presence of occluding contours overlaid on other visible objects in the image.

The second example (Figure 3.5(b)) shows two lines of *exactly the same* length, which terminate with an arrow-head at each line end. However, the arrow head pairs for correspondng line ends point in opposite directions. None the less, the first line gives an impression of being longer. This phenomenon can be explained by daily experience. The left configuration in Figure 3.5(b) suggests that the central line is further from the observer compared to the right hand line configuration. This makes us believe that the left line has to be longer in the 3D world.

What seems a common observation about such illusions in 2D images is that we experience some false interpretation of the 'flat' patterns because our visual system always tries to interpret image data as if it were views of real 3D objects [442].

In other words the heuristics we have evolved for visual perception are grounded in the assumption that we observe scenes embedded in 3D space. An understanding of these heuristics may provide the potential means by which we can craft binocular depth recovery algorithms that perform as robustly as those depth perception mechanisms of the HVS.

3.3 Geometry and Acquisition of a Single Image

In this section we provide an introduction to the geometry and image acquisition of a single camera. More specifically, we start with an explanation of the projective transformation with basic mathematics describing this process. Then, the so-called pin-hole model of a camera

is presented. Finally, we discuss the extrinsic and intrinsic parameters of acquisition with a single camera.

3.3.1 Projective Transformation

Every image acquisition system, either the human or machine visual system, by its nature performs some kind of transformation of real 3D space into 2D local space. Finding the parameters of such a transformation is fundamental to describing the acquisition system.

For most cameras a model that describes the space transformation they perform is based either on the parallel or central perspective projections. The linear parallel projection is the simplest approach. However, it only roughly approximates what we observe in real cameras [185]. Therefore the parallel projection, although linear, can be justified only if the observed objects are very close to the camera.

A better approach to describing the behaviour of real optical systems can be obtained using the perspective projective transformation which can be described by a linear equation, in a higher dimensional space of so-called homogeneous coordinates [95, 119, 122, 180]. Additionally, when describing real optical elements a simple projective transformation has to be augmented with nonlinear terms to take into account physical parameters of these [113, 185].

3.3.2 Simple Camera System: the Pin-hole Model

The simplest form of real camera comprises a pinhole and an imaging screen (or plane). Because the pinhole lies between the imaging screen and the observed 3D world scene, any ray of light that is emitted or reflected from a surface patch in the scene is constrained to travel through the pinhole before reaching the imaging screen. Therefore there is correspondence between each 2D area on the imaging screen and the area in the 3D world, as observed "through the pinhole" from the imaging screen. It is the solid angle of rays that is subtended by the pinhole that relates the field of view of each region on the imaging screen to the corresponding region imaged in the world. By this mechanism an image is built up, or projected (derived from the Latin projicere from *pro* "forward" and *jacere* "to throw") from world space to imaging space. A mathematical model of the simple pin-hole camera is illustrated in Figure 3.6. Notice that the imaging screen is now in front of the pin-hole. This formulation simplifies the concept of projection to that of magnification. In order to understand how points in the real world are related mathematically to points on the imaging screen two coordinate systems are of particular interest:

1. The external coordinate system (denoted here with a subscript 'W' for 'world') which is independent of placement and parameters of the camera.
2. The camera coordinate system (denoted by 'C', for 'camera').

The two coordinate systems are related by a translation, expressed by matrix $\mathbf{T}$, and rotation, represented by matrix $\mathbf{R}$.

The point $\mathbf{O_c}$, called *a central* or *a focal point*, together with the axes X_c, Y_c and Z_c determine the coordinate system of the camera. An important part of the camera model is the image plane $\mathbf{\Pi}$. We can observe in Figure 3.6 that this plane $\mathbf{\Pi}$ has been tessellated into rectangular elements, i.e. tiled, and that within an electronic camera implementation these tiles will form discrete photosensing locations that sample any image projected onto the plane. Each tile is

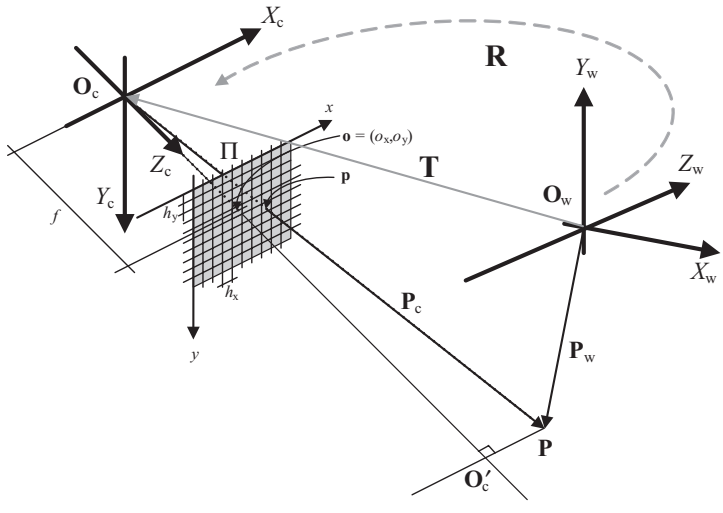

Figure 3.6 Pin-hole model of the perspective camera with two coordinate systems: external W and internal C

called a *pixel*, i.e. picture element, and is indexed by a pair of coordinates expressed by natural numbers. Figure 3.6 depicts the plane $\mathbf{\Pi}$ with a discrete grid of pixels. The projection of the point $\mathbf{O_c}$ on the plane $\mathbf{\Pi}$ in the direction of Z_c determines *the principal point* of local coordinates (o_x, o_y). The principal axis is a line between points $\mathbf{O_c}$ and $\mathbf{O'_c}$. The distance from the image plane to the principal point is known as *the focal length*. Lastly, the values h_x and h_y determine physical dimensions of a single pixel.

Placement of a given point $\mathbf{P}$ from the 3D space depends on the chosen coordinate system: in the camera coordinate system it is a column vector $\mathbf{P}_c$; in the external coordinate system it is a column vector $\mathbf{P}_w$.

Point $\mathbf{p}$ is an image of point $\mathbf{P}$ under the projection with a centre in point $\mathbf{O_c}$ on to the plane $\mathbf{\Pi}$. Coordinates of the points $\mathbf{p}$ and $\mathbf{P}$ in the camera coordinate system are denoted as follows:[2]

$$\begin{aligned} \mathbf{P} &= [X, Y, Z]^T \\ \mathbf{p} &= [x, y, z]^T . \end{aligned} \tag{3.1}$$

Since the optical axis is perpendicular to the image plane, then taking into account that the triangles $\triangle \mathbf{O_c p o}$ and $\triangle \mathbf{O_c P O'}_c$ are similar and placing $z = f$, we obtain immediately

$$x = f\frac{X}{Z}, \quad y = f\frac{Y}{Z}, \quad z = f. \tag{3.2}$$

Equation (3.2) constitutes a foundation of the pin-hole camera model.

[2]Points are denoted by letters in bold, such as $\mathbf{p}$. Their coordinates are represented either by the same letter in italic and indexed starting from 1, such as $\mathbf{p} = (p_1, p_2, p_3, p_4)$, or as $\mathbf{p} = (x, y)$ and $\mathbf{p} = (x, y, z)$ for 2D or 3D points, respectively. When necessary, points are assumed to be column vectors, such as $\mathbf{p} = [x, y, z]^T$.

The pin-hole camera model can be defined by providing two sets of parameters.

1. The extrinsic parameters.
2. The intrinsic parameters.

In the next sections we discuss these two sets in more detail.

3.3.2.1 Extrinsic Parameters

The mathematical description of a given scene depends on the chosen coordinate system. With respect to the chosen coordinate system and based solely on placement of the image plane we determine an exact placement of the camera. Thereafter, it is often practical to select just the camera coordinate system as a reference. The situation becomes yet more complicated, however, if we have more than one camera since the exact (relative) position of each camera must be determined.

A change from the camera coordinate system 'C' to the external world coordinate system 'W' can be accomplished providing a translation $\mathbf{T}$ and a rotation $\mathbf{R}$ (Figure 3.6). The translation vector $\mathbf{T}$ describes a change in position of the coordinate centres $\mathbf{O_c}$ and $\mathbf{O_w}$. The rotation, in turn, changes the corresponding axes of each system. This change is described by the orthogonal[3] matrix $\mathbf{R}$ of dimensions 3×3 [132, 430].

For a given point $\mathbf{P}$, its coordinates related to the camera 'C' and external coordinates related to the external world 'W' are connected by the following formula:

$$\mathbf{P}_\mathrm{c} = \mathbf{R}(\mathbf{P}_\mathrm{w} - \mathbf{T}), \tag{3.3}$$

where $\mathbf{P}_\mathrm{c}$ expresses placement of a point $\mathbf{P}$ in the camera coordinate system, $\mathbf{P}_\mathrm{w}$ is its placement in the external coordinate system, $\mathbf{R}$ stands for the rotation matrix and $\mathbf{T}$ is the translation matrix between origins of those two coordinate systems. The matrices $\mathbf{R}$ and $\mathbf{T}$ can be specified as follows:

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \\ \mathbf{R}_3 \end{bmatrix}_{3\times 3} = \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix}_{3\times 3}, \quad \mathbf{T} = \mathbf{O_w} - \mathbf{O_c} = \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix}_{3\times 1}, \tag{3.4}$$

where $\mathbf{R}_i$ denotes an i-th row of the rotation matrix $\mathbf{R}$, i.e. $\mathbf{R} = [R_{i1}, R_{i2}, R_{i3}]_{1\times 3}$.

Summarizing, we say that *the extrinsic parameters of the perspective camera* are all the necessary geometric parameters that allow a change from the camera coordinate system to the external coordinate system and vice versa. Thus, the extrinsic parameters of a camera are just introduced matrices $\mathbf{R}$ and $\mathbf{T}$.

[3] That is, $\mathbf{R}\mathbf{R}^\mathrm{T} = 1$.

3.3.2.2 Intrinsic Parameters

The intrinsic camera parameters can be summarized as follows.

1. The parameters of the projective transformation itself: For the pin-hole camera model, this is given by the focal length f.
2. The parameters that map the camera coordinate system into the image coordinate system: Assuming that the origin of the image constitutes a point $\boldsymbol{o} = (o_x, o_y)$ (i.e. a central point) and that the physical dimensions of pixels (usually expressed in μm) on a camera plane in the two directions are constant and given by h_x and h_y, a relation between image coordinates x_u and y_u and camera coordinates x and y can be stated as follows (see Figure 3.6):

$$\begin{aligned} x &= (x_u - o_x)h_x \\ y &= (y_u - o_y)h_y, \end{aligned} \tag{3.5}$$

   where a point (x, y) is related to the camera coordinate system 'C', whereas (x_u, y_u) and (o_x, o_y) to the system of a local camera plane. It is customary to assume that $x_u \geq 0$ and $y_u \geq 0$. For instance, the point of origin of the camera plane $(x_u,y_u) = (0, 0)$ transforms to the point $(-o_xh_x,-o_yh_y)$ of the system 'C'. More often than not it is assumed also that $h_x = h_y = 1$. A value of h_y/h_x is called an aspect ratio. Under this assumption a point from our example is simply $(-o_x,-o_y)$ in the 'C' coordinates, which can be easily verified analysing Figure 3.6.
3. Geometric distortions that arise due to the physical parameters of the optical elements of the camera: Distortions encountered in real optical systems arise mostly from the nonlinearity of these elements, as well as from the dependence of the optical parameters on the wavelength of the incident light [185, 343, 382]. In the first case we talk about spherical aberration, coma, astigmatism, curvature of the view field and distortions. The second case is related to the chromatic aberration [50, 185, 382]. In the majority of practical situations, we can model these phenomena as radial distortions, the values of which increase for points more distant from the image centre. The radial distortions can be modelled by providing a nonlinear correction (offset) to the real coordinates of a given image point. This can be accomplished by adding even-order polynomial terms, as follows:

$$x_v = \frac{x_u}{1 + k_1r^2 + k_2r^4}, \quad y_v = \frac{y_u}{1 + k_1r^2 + k_2r^4}, \tag{3.6}$$

   where $r^2 = x_v^2 + y_v^2$, k_1 and k_2 are the new intrinsic parameters of the perspective camera that model the influence of the radial distortions of the optical system; x_u and y_u are the ideal (i.e. as if there were no distortions) coordinates of a given image point; and x_v and y_v are modified coordinates reflecting the radial distortions.

An iterative algorithm for finding x_v and y_v is provided by Klette *et al.* [246]. Trucco and Verri suggest that for most real optical systems with a CCD sensor of around 500×500 image elements, setting k_2 to 0 does not introduce any significant change to the quality of the camera model [430].

3.3.3 Projective Transformation of the Pin-hole Camera

Substituting (3.3) and (3.5) into (3.2) and disregarding distortions (3.6) we obtain the linear equation of the pin-hole camera:[4]

$$\mathbf{p} = \mathbf{MP}, \tag{3.7}$$

where $\mathbf{p}$ is an image of the point $\mathbf{P}$ under transformation $\mathbf{M}$ performed by the pin-hole camera. Linearity in (3.7) is due to the homogeneous[5] transformation of the point coordinates.

The matrix $\mathbf{M}$ in (3.7), called a projection matrix, can be partitioned into the following product of two matrices:

$$\mathbf{M} = \mathbf{M}_\mathrm{i}\mathbf{M}_\mathrm{e}, \tag{3.8}$$

where

$$\mathbf{M_i} = \begin{bmatrix} \frac{f}{h_x} & 0 & o_x \\ 0 & \frac{f}{h_y} & o_y \\ 0 & 0 & 1 \end{bmatrix}_{3\times 3}, \quad \mathbf{M_e} = \begin{bmatrix} \mathbf{R}_1 & -\mathbf{R}_1\mathbf{T} \\ \mathbf{R}_2 & -\mathbf{R}_2\mathbf{T} \\ \mathbf{R}_3 & -\mathbf{R}_3\mathbf{T} \end{bmatrix}_{3\times 4}. \tag{3.9}$$

The matrices $\mathbf{R}$ and $\mathbf{T}$ are given in (3.4). $\mathbf{M}_\mathrm{i}$ defines the intrinsic parameters of the pin-hole camera, that is, the distance of the camera plane to the centre of the camera's coordinate system, as well as placement of the central point o and physical dimensions of the pixels on the camera plane – these are discussed in section 3.3.2.2. The matrix $\mathbf{M}_\mathrm{e}$ contains the extrinsic parameters of the pin-hole camera and relates the camera and the external 'world' coordinate systems (section 3.3.2.1).

The three equations above can be joined together as follows:

$$\mathbf{p} = \begin{bmatrix} x_{uh} \\ y_{uh} \\ z_{uh} \end{bmatrix} = \underbrace{\begin{bmatrix} \frac{f}{h_x} & 0 & o_x \\ 0 & \frac{f}{h_y} & o_y \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{M}_\mathrm{i}} \underbrace{\begin{bmatrix} \mathbf{R}_1 & -\mathbf{R}_1\mathbf{T} \\ \mathbf{R}_2 & -\mathbf{R}_2\mathbf{T} \\ \mathbf{R}_3 & -\mathbf{R}_3\mathbf{T} \end{bmatrix}}_{\mathbf{M}_\mathrm{e}} \mathbf{P}, \tag{3.10}$$

where $\mathbf{P} = [\mathbf{P}_\mathrm{w}\ 1]^\mathrm{T}$ is a point $\mathbf{P}_\mathrm{w}$ expressed in the homogeneous coordinates.

Let us observe that

$$x_u = \frac{x_{uh}}{z_{uh}}, \quad y_u = \frac{y_{uh}}{z_{uh}}. \tag{3.11}$$

[4]Derivation of the equations for the projective transformation of a camera can be found in section 3.8.

[5]Before further study, readers not familiar with the concept of homogeneous coordinates are asked to read section 10.1.

As already alluded to, it is often assumed that $(o_x,o_y) = (0,0)$, and also $h_x = h_y = 1$. With these assumptions (3.10) takes on a simpler form

$$\mathbf{p} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{R}_1 & -\mathbf{R}_1\mathbf{T} \\ \mathbf{R}_2 & -\mathbf{R}_2\mathbf{T} \\ \mathbf{R}_3 & -\mathbf{R}_3\mathbf{T} \end{bmatrix} \mathbf{P}. \tag{3.12}$$

Equation (3.7) defines a transformation of the projective space $\wp^3$ into the projective plane $\wp^2$. However, note that this transformation changes each point of a line into exactly one and the same image point of the image plane. This line is given by the central point O_c and any other point from the projective space. Therefore the projective transformation (3.7) assigns exactly the same image point to *all* the points belonging to the mentioned line. This fact can be embedded into (3.7) by introduction of an additional scaling parameter, as follows:

$$\gamma\mathbf{p} = \mathbf{MP}, \tag{3.13}$$

where γ is a scalar. Equations (3.7) and (3.13) differ only by the scalar γ. It can also be said that (3.7) is a version of (3.13) after dividing both sides by a nonzero scalar γ. Thus, without loss of generality we will assume henceforth that (3.7) holds, where the matrix $\mathbf{M}$ is defined only up to a certain multiplicative parameter γ.

3.3.4 Special Camera Setups

For some camera setups it is possible to assume that distances among observed objects are significantly smaller than the average distance $\bar{z}$ from those objects to the centre of projection. Under this assumption we obtain a simplified camera model; termed *weak perspective* [314, 322, 430]. In this model the perspective projection simplifies to the parallel projection by the scaled magnification factor $f/\bar{z}$. Equations (3.2) transform then to the following set of equations:

$$x = \frac{f}{\overline{Z}}X, \quad y = \frac{f}{\overline{Z}}Y, \quad z = f, \tag{3.14}$$

where $\overline{Z}$ is assumed to be much larger than f and constant for the particular setup of a camera and a scene. This simplification makes (3.14) independent of the current depth of an observed point $\mathbf{P}_w$. Thus, in the case of a camera with a simplified perspective the element at indices 3×1 of the matrix $\mathbf{M}_e$ in (3.9) changes to $\mathbf{0}$, and the element 3×2 of this matrix changes to $\overline{Z}$ (section 3.8). The latter, in turn, can be defined selecting an arbitrary point $\mathbf{A}_w$, which is the same for acquisition of the whole scene

$$\overline{Z} = \mathbf{R}_3(\mathbf{A}_w - \mathbf{T}). \tag{3.15}$$

The mathematical extension to this simplification is a model of an *affine camera* in which proportions of distances measured alongside parallel directions are invariant [122, 314, 322, 430]. There are also other camera models that take into consideration parameters of real lenses, e.g. see Kolb *et al*. [251]. Finally, more information on design of real lenses can be found in [113, 382].

3.3.5 Parameters of Real Camera Systems

The quality of the images obtained by real acquisition systems depends also on many other factors beyond those already discussed. These are related to the physical and technological phenomena which influence the acquisition process. In this section we briefly discuss such factors.

1. *Limited dynamics of the system.* The basic photo-transducer element within a modern digital camera converts the number of photons collected over a specific time interval (the integration interval of the device, analogous to the exposure time in a film camera) within each pixel within the sensor array into a voltage. While this voltage is linearly proportional to the intensity of the input photon flux arriving at a given pixel, the following analog-to-digital converter circuitry is limited to a finite number of bits of precision with which to represent the incoming voltage. Therefore, in order to extend the allowable input signal range nonlinear limiting circuits are introduced prior to digitisation. One such limiter is the pre-knee circuit [246] whose circuit characteristic causes a small degree of saturation for higher values of the input signal. As a result, the input range of the system is increased but at a cost of a slight nonlinearity.
2. *Resolution of the CCD element and aliasing*. In agreement with the sampling theory, to avoid aliasing, a device converting continuous signals into a discrete representation must fulfil the Nyquist sampling criterion (i.e. the sampling frequency has to be at least twice the value of the highest frequency component of the sampled signal). In the rest of this book we assume that this is the case and that aliasing is not present [312, 336]. In real imaging systems there are two factors that can help to alleviate the problem of aliasing. The first consists of the application of low-pass filters at the input circuitry. The second is the natural low-pass filter effect due to the lens itself, manifest as a point spread function (PSF) or modulation transfer function (MTF) which naturally limits the high spatial frequencies present prior to digitisation [66, 430].
3. *Noise*. Each image acquisition channel contains many sources of noise. In the CCD device there is a source of noise in the form of cross-talk. This is the phenomenon of charge leakage between neighbouring photoreceptors in each row of the CCD. Another source of noise comes from the filters and the analogue-to-digital converter. The latter adds so-called quantization noise which is a result of the finite length of bit streams representing analogue signals. The most frequently encountered types of image noise can be represented by Poison and Gaussian distributions. Schott Noise is by far the most significant source of noise in a modern imaging sensor. This noise source results from the statistical variation of the photon arrival rate from any illumination source. In any fixed time interval the standard deviation of the photon flux rate is proportional to the square of the illumination intensity. Other sources of noise are now becoming less significant than Schott noise, hence this fundamental limit of physics now tends to dominate image capture performance. The interested reader is referred to the ample literature [95, 158, 172, 183, 224, 226, 247, 346, 430]. Different types of noise are also discussed in Chapter 11.
4. *Signal saturation*. The phenomenon of signal saturation results from an excessive signal level being applied to the input of the acquisition channel. Such a signal is nonlinearly attenuated and cannot be accurately converted by A/C converters due to their limited dynamic range. Where there is insufficient scene illumination, as can be caused by shadows, the