



3D Computer Vision

**Efficient Methods
and Applications**

Christian Wöhler

X.media.publishing is an application-oriented series that specializes in the presentation and publication of multimedia as well as digital and print media.



X.media.publishing



Christian Wöhler

3D Computer Vision

Efficient Methods and Applications



Dr. Christian Wöhler
Daimler AG, Group Research
and Advanced Engineering
P. O. Box 2360
D-89013 Ulm
christian.woehler@daimler.com

ISSN 1612-1449
ISBN 978-3-642-01731-5 e-ISBN 978-3-642-01732-2
DOI 10.1007/978-3-642-01732-2
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009929715

© Springer-Verlag Berlin Heidelberg 2009

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover design: KünkelLopka GmbH

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Katja, Alexander, and Sebastian

Preface

This work provides an introduction to the foundations of three-dimensional computer vision and describes recent contributions to the field, which are of methodical and application-specific nature. Each chapter of this work provides an extensive overview of the corresponding state of the art, into which a detailed description of new methods or evaluation results in application-specific systems is embedded.

Geometric approaches to three-dimensional scene reconstruction (cf. Chapter 1) are primarily based on the concept of bundle adjustment, which has been developed more than 100 years ago in the domain of photogrammetry. The three-dimensional scene structure and the intrinsic and extrinsic camera parameters are determined such that the Euclidean backprojection error in the image plane is minimised, usually relying on a nonlinear optimisation procedure. In the field of computer vision, an alternative framework based on projective geometry has emerged during the last two decades, which allows to use linear algebra techniques for three-dimensional scene reconstruction and camera calibration purposes. With special emphasis on the problems of stereo image analysis and camera calibration, these fairly different approaches are related to each other in the presented work, and their advantages and drawbacks are stated. In this context, various state-of-the-art camera calibration and self-calibration methods as well as recent contributions towards automated camera calibration systems are described. An overview of classical and new feature-based, correlation-based, dense, and spatio-temporal methods for establishing point correspondences between pairs of stereo images is given. Furthermore, an analysis of traditional and newly introduced methods for the segmentation of point clouds and for the three-dimensional detection and pose estimation of rigid, articulated, and flexible objects in the scene is provided.

A different class of three-dimensional scene reconstruction methods is made up by photometric approaches (cf. Chapter 2), which evaluate the intensity distribution in the image to infer the three-dimensional scene structure. Basically, these methods can be divided into shape from shadow, photocalinometry and shape from shading, photometric stereo, and shape from polarisation. As long as sufficient information about the illumination conditions and the surface reflectance properties is available, these methods may provide dense depth maps of object surfaces.

In a third, fundamentally different class of approaches the behaviour of the point spread function of the optical system used for image acquisition is exploited in order to derive depth information about the scene (cf. Chapter 3). Depth from focus methods use as a reference the distance between the camera and the scene at which a minimum width of the point spread function is observed, relying on an appropriate calibration procedure. Depth from defocus methods determine the position-dependent point spread function, which in turn yields absolute depth values for the scene points. A semi-empirical framework for establishing a relation between the depth of a scene point and the observed width of the point spread function is introduced.

These three classes of approaches to three-dimensional scene reconstruction are characterised by complementary properties, such that it is favourable to integrate them into unified frameworks that yield more accurate and robust results than each of the approaches alone (cf. Chapter 4). Bundle adjustment and depth from defocus are combined to determine the absolute scale factor of the scene reconstruction result, which cannot be obtained by bundle adjustment alone if no a-priori information is available. Shading and shadow features are integrated into a self-consistent framework to reduce the inherent ambiguity and large-scale inaccuracy of the shape from shading technique by introducing regularisation terms that rely on depth differences inferred from shadow analysis. Another integrated approach combines photometric, polarimetric, and sparse depth information, yielding a three-dimensional reconstruction result which is equally accurate on large and on small scales. An extension of this method provides a framework for stereo image analysis of non-Lambertian surfaces, where traditional stereo methods tend to fail. In the context of monocular three-dimensional pose estimation, the integration of geometric, photopolarimetric, and defocus cues is demonstrated to behave more robustly and is shown to provide significantly more accurate results than techniques exclusively relying on geometric information.

The developed three-dimensional scene reconstruction methods are examined in different application scenarios. A comparison to state-of-the-art systems is provided where possible. In the context of industrial quality inspection (cf. Chapter 5), the performance of pose estimation is evaluated for rigid objects (plastic caps, electric plugs) as well as flexible objects (tubes, cables). The integrated surface reconstruction methods are applied to the inspection of different kinds of metallic surfaces, where the achieved accuracies are found to be comparable to those of general-purpose active scanning devices which, however, require a much higher instrumental effort.

The developed techniques for object detection and tracking in three-dimensional point clouds and for pose estimation of articulated objects are evaluated in the context of partially automated industrial production scenarios requiring a safe interaction between humans and industrial robots (cf. Chapter 6). An overview of existing vision-based robotic safety systems is given, and it is worked out how the developed three-dimensional detection and pose estimation techniques are related to state-of-the-art gesture recognition methods in human–robot interaction scenarios.

The third addressed application scenario is completely different and regards remote sensing of the lunar surface by preparing elevation maps (cf. Chapter 7). While the spatial scales taken into account differ by many orders of magnitude from those encountered in the industrial quality inspection domain, the underlying physical processes are fairly similar. An introductory outline of state-of-the-art geometric, photometric, and combined approaches to topographic mapping of solar system bodies is given. Especially the estimation of impact crater depths and shapes is an issue of high geological relevance. Generally, such measurements are based on the determination of shadow lengths and do not yield detailed elevation maps. It is demonstrated for lunar craters that three-dimensional surface reconstruction based on shadow, reflectance, and geometric information yields topographic maps of high resolution, which are useful for a reliable crater classification. Another geologically relevant field is the three-dimensional reconstruction of lunar volcanic edifices, especially lunar domes. These structures are so low that most of them do not appear in the existing lunar topographic maps. Based on the described photometric three-dimensional reconstruction methods, the first catalogue to date containing heights and edifice volumes for a statistically significant number of lunar domes has been prepared. It is outlined briefly why the determined three-dimensional morphometric data are essential for deriving basic geophysical parameters of lunar domes, such as lava viscosity and effusion rate, and how they may help to reveal their origin and mode of formation.

Finally (cf. Chapter 8), the main results of the presented work and the most important conclusions are summarised, and possible directions of future research are outlined.

Heroldstatt, May 2009

Christian Wöhler

Acknowledgements

First of all, I wish to express my gratitude to my wife Khadija Katja and my sons Adnan Alexander Émile and Sebastian Marc Amin for their patience and continuous encouragement.

I am grateful to Prof. Dr. Gerhard Sagerer (Technical Faculty, Bielefeld University), Prof. Dr. Reinhard Klette (Computer Science Department, University of Auckland), and Prof. Dr. Rainer Ott (Faculty of Computer Science, Electrical Engineering, and Information Technology, Stuttgart University) for providing the reviews for this work.

Moreover, I wish to thank Prof. Dr. Gerhard Sagerer, Prof. Dr. Franz Kummert, Joachim Schmidt, and Niklas Beuter from Bielefeld University for the fruitful collaboration. I gratefully acknowledge to be given the opportunity to become a visiting lecturer at the Technical Faculty and thus to stay in touch with the university environment. I also wish to thank Prof. Dr. Horst-Michael Groß from the Technical University of Ilmenau for the long-lasting cooperation.

Special thanks go to my colleagues in the Environment Perception department at Daimler Group Research and Advanced Engineering in Ulm for providing a lively and inspiring scientific environment, especially to Dr. Lars Krüger (to whom I am extraordinarily indebted for his critical reading of the manuscript), Prof. Dr. Rainer Ott, Dr. Ulrich Kreßel, Frank Lindner, and Kia Hafezi, to our (former and current) PhD students Dr. Pablo d'Angelo, Dr. Marc Ellenrieder, Björn Barrois, Markus Hahn, and Christoph Hermes, and Diplom students Annika Kuhl, Tobias Gövert, and Melanie Krauß. I also wish to thank Claus Lörcher and his team colleagues, Werner Progscha, Dr. Rolf Finkele, and Mike Böpple for their continuous support.

Furthermore, I am grateful to the members of the Geologic Lunar Research Group, especially Dr. Raffaello Lena, Dr. Charles A. Wood, Paolo Lazzarotti, Dr. Jim Phillips, Michael Wirths, K. C. Pau, Maria Teresa Bregante, and Richard Evans, for sharing their experience in many projects concerning lunar observation and geology.

My thanks are extended to the Springer editorial staff, especially Hermann Engesser, Dorothea Glaunsinger, and Gabi Fischer, for their advice and cooperation.

Contents

Part I Methods of 3D Computer Vision

1 Geometric Approaches to Three-dimensional Scene Reconstruction	3
1.1 The Pinhole Camera Model	3
1.2 Bundle Adjustment Methods	7
1.3 Geometric Aspects of Stereo Image Analysis	9
1.3.1 Euclidean Formulation of Stereo Image Analysis	9
1.3.2 Stereo Image Analysis in Terms of Projective Geometry	12
1.4 Geometric Calibration of Single and Multiple Cameras	17
1.4.1 Methods for Intrinsic Camera Calibration	17
1.4.2 The Direct Linear Transform (DLT) Method	18
1.4.3 The Camera Calibration Method by Tsai (1987)	21
1.4.4 The Camera Calibration Method by Zhang (1999a)	25
1.4.5 The Camera Calibration Method by Bouguet (2007)	27
1.4.6 Self-calibration of Camera Systems from Multiple Views of a Static Scene	28
1.4.7 Semi-automatic Calibration of Multiocular Camera Systems	41
1.4.8 Accurate Localisation of Chequerboard Corners	51
1.5 Stereo Image Analysis in Standard Geometry	62
1.5.1 Image Rectification According to Standard Geometry	62
1.5.2 The Determination of Corresponding Points	66
1.6 Three-dimensional Pose Estimation and Segmentation Methods	87
1.6.1 Pose Estimation of Rigid Objects	88
1.6.2 Pose Estimation of Non-rigid and Articulated Objects	95
1.6.3 Point Cloud Segmentation Approaches	113
2 Photometric Approaches to Three-dimensional Scene Reconstruction 127	
2.1 Shape from Shadow	127
2.1.1 Extraction of Shadows from Image Pairs	128
2.1.2 Shadow-based Surface Reconstruction from Dense Sets of Images	130

2.2	Shape from Shading	132
2.2.1	The Bidirectional Reflectance Distribution Function (BRDF)	132
2.2.2	Determination of Surface Gradients	137
2.2.3	Reconstruction of Height from Gradients	142
2.2.4	Surface Reconstruction Based on Eikonal Equations	144
2.3	Photometric Stereo	146
2.3.1	Classical Photometric Stereo Approaches	147
2.3.2	Photometric Stereo Approaches Based on Ratio Images	148
2.4	Shape from Polarisation	151
2.4.1	Surface Orientation from Dielectric Polarisation Models	151
2.4.2	Determination of Polarimetric Properties of Rough Metallic Surfaces for Three-dimensional Reconstruction Purposes	154
3	Real-aperture Approaches to Three-dimensional Scene Reconstruction	159
3.1	Depth from Focus	161
3.2	Depth from Defocus	162
3.2.1	Basic Principles	162
3.2.2	Determination of Small Depth Differences	167
3.2.3	Determination of Absolute Depth Across Broad Ranges	170
4	Integrated Frameworks for Three-dimensional Scene Reconstruction	181
4.1	Monocular Three-dimensional Scene Reconstruction at Absolute Scale	182
4.1.1	Combining Motion, Structure, and Defocus	183
4.1.2	Online Version of the Algorithm	184
4.1.3	Experimental Evaluation Based on Tabletop Scenes	185
4.1.4	Discussion	195
4.2	Self-consistent Combination of Shadow and Shading Features	196
4.2.1	Selection of a Shape from Shading Solution Based on Shadow Analysis	197
4.2.2	Accounting for the Detailed Shadow Structure in the Shape from Shading Formalism	200
4.2.3	Initialisation of the Shape from Shading Algorithm Based on Shadow Analysis	202
4.2.4	Experimental Evaluation Based on Synthetic Data	204
4.2.5	Discussion	205
4.3	Shape from Photopolarimetric Reflectance and Depth	206
4.3.1	Shape from Photopolarimetric Reflectance	207
4.3.2	Estimation of the Surface Albedo	211
4.3.3	Integration of Depth Information	212
4.3.4	Experimental Evaluation Based on Synthetic Data	217
4.3.5	Discussion	222
4.4	Stereo Image Analysis of Non-Lambertian Surfaces	223

4.4.1	Iterative Scheme for Disparity Estimation	225
4.4.2	Qualitative Behaviour of the Specular Stereo Algorithm	229
4.5	Three-dimensional Pose Estimation Based on Combinations of Monocular Cues	230
4.5.1	Appearance-based Pose Estimation	
Relying on Multiple Monocular Cues	231	
4.5.2	Contour-based Pose Estimation Using Depth from Defocus .	236
Part II Application Scenarios		
5	Applications to Industrial Quality Inspection	243
5.1	Inspection of Rigid Parts	244
5.1.1	Object Detection by Pose Estimation	244
5.1.2	Pose Refinement	248
5.2	Inspection of Non-rigid Parts	253
5.3	Inspection of Metallic Surfaces	256
5.3.1	Inspection Based on Integration of Shadow and Shading Features	256
5.3.2	Inspection of Surfaces with Non-uniform Albedo	257
5.3.3	Inspection Based on SfPR and SfPRD	259
5.3.4	Inspection Based on Specular Stereo	266
5.3.5	Discussion	273
6	Applications to Safe Human–Robot Interaction	277
6.1	Vision-based Human–Robot Interaction	277
6.1.1	The Role of Gestures in Human–Robot Interaction	278
6.1.2	Safe Human–Robot Interaction	279
6.1.3	Pose Estimation of Articulated Objects in the Context of Human–Robot Interaction	282
6.2	Object Detection and Tracking in Three-dimensional Point Clouds .	291
6.3	Detection and Spatio-temporal Pose Estimation of Human Body Parts	293
6.4	Three-dimensional Tracking of Human Body Parts	296
7	Applications to Lunar Remote Sensing	303
7.1	Three-dimensional Surface Reconstruction Methods for Planetary Remote Sensing	304
7.1.1	Topographic Mapping of Solar System Bodies	304
7.1.2	Reflectance Behaviour of Planetary Regolith Surfaces .	307
7.2	Three-dimensional Reconstruction of Lunar Impact Craters	311
7.2.1	Shadow-based Measurement of Crater Depth	311
7.2.2	Three-dimensional Reconstruction of Lunar Impact Craters at High Resolution	314
7.3	Three-dimensional Reconstruction of Lunar Wrinkle Ridges and Faults	322

7.4	Three-dimensional Reconstruction of Lunar Domes	325
7.4.1	General Overview of Lunar Mare Domes	325
7.4.2	Observations of Lunar Mare Domes	328
7.4.3	Image-based Determination of Morphometric Data	331
7.4.4	Geophysical Insights Gained from Topographic Data	343
8	Conclusion	351
	References	359

Chapter 1

Geometric Approaches to Three-dimensional Scene Reconstruction

Reconstruction of three-dimensional scene structure from images was an important topic already in the early history of photography, which was invented by Niepce and Daguerre in 1839. The first photogrammetric methods were developed in the middle of the 19th century by Laussedat and Meydenbauer for mapping purposes and reconstruction of buildings (Luhmann, 2003). These photogrammetric methods were based on geometric modelling of the image formation process, exploiting the perspective projection of a three-dimensional scene into a two-dimensional image plane. Image formation by perspective projection corresponds to the pinhole camera model. There are different image formation models, describing optical devices such as fisheye lenses or omnidirectional lenses. In this work, however, we will restrict ourselves to the pinhole model since it represents the most common image acquisition devices.

1.1 The Pinhole Camera Model

In the pinhole camera model, the camera lens is represented by its optical centre, corresponding to a point situated between the three-dimensional scene and the two-dimensional image plane, and the optical axis, which is perpendicular to the plane defined by the lens and passes through the optical centre (Fig. 1.1). The intersection point between the image plane and the optical axis is termed principal point in the computer vision literature (Faugeras, 1993). The distance between the optical centre and the principal point is termed principal distance and is denoted by b . For real lenses, the principal distance b is always larger than the focal length f of the lens, and the value of b approaches f if the object distance Z is much larger than b . This issue will be further examined in Chapter 3.

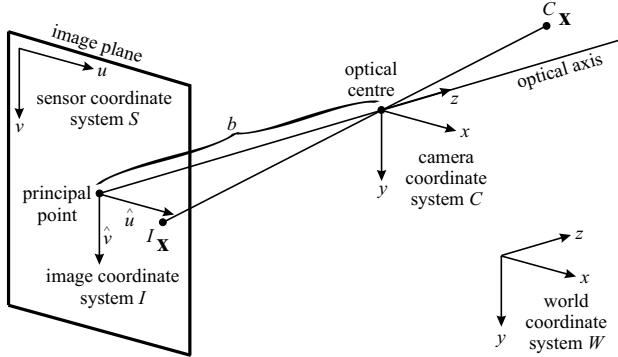


Fig. 1.1 The pinhole camera model. A scene point $C\mathbf{x}$ defined in the camera coordinate system is projected into the image point $I\mathbf{x}$ located in the image plane.

Euclidean Formulation

In this work we will utilise a notation similar to the one by Craig (1989) for points, coordinate systems, and transformation matrices. Accordingly, a point \mathbf{x} in the camera coordinate system C is denoted by $C\mathbf{x}$, where the origin of C corresponds to the principal point. Similarly, a transformation of a point in the world coordinate system W into the camera coordinate system C is denoted by a transformation ${}_W^C T$, where the lower index defines the original coordinate system and the upper index the coordinate system into which the point is transformed. The transformation ${}_W^C T$ corresponds to an arbitrary rotation and translation. In this notation, the transformation is given by $C\mathbf{x} = {}_W^C T {}^W \mathbf{x}$. A scene point $C\mathbf{x} = (x, y, z)^T$ defined in the camera coordinate system C is projected on the image plane into the point $I\mathbf{x}$, defined in the image coordinate system I , such that the scene point $C\mathbf{x}$, the optical centre, and the image point $I\mathbf{x}$ are connected by a straight line in three-dimensional space (Fig. 1.1). Obviously, all scene points situated on this straight line are projected into the same point in the image plane, such that the original depth information z gets lost. Elementary geometrical considerations yield for the point $I\mathbf{x} = (\hat{u}, \hat{v})$ in the image coordinate system:

$$\begin{aligned}\hat{u} &= -b \frac{x}{z} \\ \hat{v} &= -b \frac{y}{z}.\end{aligned}\tag{1.1}$$

The coordinates \hat{u} and \hat{v} in the image plane are measured in the same metric units as x , y , z , and b . The principal point is given in the image plane by $\hat{u} = \hat{v} = 0$. In contrast, pixel coordinates in the coordinate system of the camera sensor are denoted by u and v .

While it may be useful to regard the camera coordinate system C as identical to the world coordinate system W for a single camera, it is favourable to explicitly

define a world coordinate system as soon as multiple cameras are involved. The orientation and translation of each camera i with respect to this world coordinate system is then expressed by ${}_{\mathcal{W}}^{\mathcal{C}_i}T$, transforming a point ${}^W\mathbf{x}$ from the world coordinate system \mathcal{W} into the camera coordinate system \mathcal{C}_i . The transformation ${}_{\mathcal{W}}^{\mathcal{C}_i}T$ is composed of a rotational part R_i , corresponding to an orthonormal matrix of size 3×3 determined by three independent parameters, e.g. the Euler rotation angles (Craig, 1989), and a translation vector \mathbf{t}_i denoting the offset between the coordinate systems. This decomposition yields

$${}_{\mathcal{C}_i}\mathbf{x} = {}_{\mathcal{W}}^{\mathcal{C}_i}T({}^W\mathbf{x}) = R_i {}^W\mathbf{x} + \mathbf{t}_i. \quad (1.2)$$

Furthermore, the image formation process is determined by the intrinsic parameters $\{c_j\}_i$ of each camera i , some of which are lens-specific while others are sensor-specific. For a pinhole camera equipped with a digital sensor, these parameters comprise the principal distance b , the effective number of pixels per unit length k_u and k_v along the horizontal and the vertical image axis, respectively, the pixel skew angle θ , and the coordinates u_0 and v_0 of the principal point in the image plane. For most modern camera sensors, the skew angle amounts to $\theta = 90^\circ$ and the pixels are of quadratic shape with $k_u = k_v$.

For a real lens system, however, the observed image coordinates of scene points may deviate from those given by Eq. (1.1) due to the effect of lens distortion. In this work we employ the lens distortion model by Brown (1966, 1971) which has been extended by Heikkilä and Silvén (1997) and by Bouguet (1999). The distorted coordinates ${}^I\mathbf{x}_d$ of a point in the image plane are obtained from the undistorted coordinates ${}^I\mathbf{x}$ according to

$${}^I\mathbf{x}_d = (1 + k_1 r^2 + k_3 r^4 + k_5 r^6) {}^I\mathbf{x} + \mathbf{d}_t, \quad (1.3)$$

where ${}^I\mathbf{x} = (\hat{u}, \hat{v})^T$ and $r^2 = \hat{u}^2 + \hat{v}^2$. If radial distortion is present, straight lines in the object space crossing the optical axis still appear straight in the image, but the observed distance of a point in the image from the principal point deviates from the distance expected according to Eq. (1.1). The vector

$$\mathbf{d}_t = \begin{pmatrix} 2k_2\hat{u}\hat{v} + k_4(r^2 + 2\hat{u}^2) \\ k_2(r^2 + 2\hat{v}^2) + 2k_4\hat{u}\hat{v} \end{pmatrix} \quad (1.4)$$

is termed tangential distortion. The occurrence of tangential distortion implies that straight lines in the object space crossing the optical axis appear bent in some directions in the image.

When a film is used as an imaging sensor, \hat{u} and \hat{v} directly denote metric distances on the film with respect to the principal point, which has to be determined by an appropriate calibration procedure (cf. Section 1.4). When a digital camera sensor is used, the transformation

$${}^S\mathbf{x} = {}^I T({}^I\mathbf{x}) \quad (1.5)$$

from the image coordinate system into the sensor coordinate system is defined in the general case by an affine transformation ${}_I^S T$ (as long as the sensor has no “exotic” architecture such as a hexagonal pixel raster, where the transformation would be still more complex). The corresponding coordinates ${}^S \mathbf{x} = (u, v)^T$ are measured in pixels.

At this point it is useful to define a projection function $\mathcal{P}\left({}^{C_i}_W T, \{c_j\}_i, {}^W \mathbf{x}\right)$ which projects a point ${}^W \mathbf{x}$ defined in the world coordinate system into the sensor coordinate system of camera i by means of a perspective projection as defined in Eq. (1.1) with

$${}_{S_i} \mathbf{x} = \mathcal{P}\left({}^{C_i}_W T, \{c_j\}_i, {}^W \mathbf{x}\right). \quad (1.6)$$

Since Eq. (1.1) is based on Euclidean geometry, it is nonlinear in z , implying that the function \mathcal{P} is nonlinear as well. It depends on the extrinsic camera parameters defined by the transformation ${}^{C_i}_W T$ and on the lens-specific and sensor-specific intrinsic camera parameters $\{c_j\}_i$.

Formulation in Terms of Projective Geometry

To circumvent the nonlinear formulation of perspective projection in Euclidean geometry, it is advantageous to express the image formation process in the more general mathematical framework of projective geometry (Faugeras, 1993; Birchfield, 1998). A point $\mathbf{x} = (x, y, z)^T$ in three-dimensional Euclidean space is represented in three-dimensional projective space by the homogeneous coordinates $\tilde{\mathbf{x}} = (X, Y, Z, W)^T = (x, y, z, 1)^T$. Overall scaling is unimportant, such that $(X, Y, Z, W)^T$ is equivalent to $(\alpha X, \alpha Y, \alpha Z, \alpha W)^T$ for any nonzero value of α . To recover the Euclidean coordinates from a point given in three-dimensional projective space, the first three coordinates X , Y , and Z are divided by the fourth coordinate W according to $\mathbf{x} = (X/W, Y/W, Z/W)^T$. The general transformation in three-dimensional projective space is a matrix multiplication by a 4×4 matrix. For the projection from a three-dimensional world into a two-dimensional image plane a matrix of size 3×4 is sufficient. Hence, analogous to Eq. (1.1), in projective geometry the projection of a scene point ${}^{C_i} \tilde{\mathbf{x}}$ defined in the camera coordinate system C_i into the image coordinate system I_i is given by the linear relation

$${}_{I_i} \tilde{\mathbf{x}} = \begin{bmatrix} -b & 0 & 0 & 0 \\ 0 & -b & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} {}^{C_i} \tilde{\mathbf{x}}. \quad (1.7)$$

This formulation of perspective projection is widely used in the fields of computer vision (Faugeras, 1993) and computer graphics (Foley et al., 1993). An important class of projective transforms is defined by the essential matrix, containing the extrinsic parameters of two pinhole cameras observing a scene from two different viewpoints. The fundamental matrix is a generalisation of the essential matrix and contains as additional information the intrinsic camera parameters (Birchfield, 1998). A more detailed explanation of the essential and the fundamental matrix will

be given in Section 1.3 in the context of the epipolar constraint of stereo image analysis.

In the formulation of projective geometry, the transformation from the world coordinate system W into the camera coordinate system C_i is defined by the 3×4 matrix

$$[R_i \mid \mathbf{t}_i]. \quad (1.8)$$

The projection from the coordinate system C_i of camera i into the sensor coordinate system S_i is given by the matrix

$$A_i = \begin{bmatrix} \alpha_u & \alpha_u \cot \theta & u_0 \\ 0 & \alpha_v / \sin \theta & v_0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (1.9)$$

with α_u , α_v , θ , u_0 , and v_0 as the intrinsic parameters of the pinhole camera i . In Eq. (1.9), the scale parameters α_u and α_v are defined according to $\alpha_u = -bk_u$ and $\alpha_v = -bk_v$. The complete image formation process can then be described in terms of the projective 3×4 matrix P_i which is composed of a perspective projection along with the intrinsic and extrinsic camera parameters according to

$${}^{S_i}\tilde{\mathbf{x}} = P_i {}^W\tilde{\mathbf{x}} = A_i [R_i \mid \mathbf{t}_i] {}^W\tilde{\mathbf{x}}, \quad (1.10)$$

such that $P_i = A_i [R_i \mid \mathbf{t}_i]$. For each camera i , the linear projective transformation P_i describes the image formation process in projective space.

1.2 Bundle Adjustment Methods

Most geometric methods for three-dimensional scene reconstruction from multiple images are based on establishing corresponding points in the images. For a scene point ${}^W\mathbf{x}$ observed in N images, the corresponding image points ${}^{S_i}\mathbf{x}$ in each image i , where $i = 1, \dots, N$, can be determined manually or by automatic correspondence search methods. Given the extrinsic and intrinsic camera parameters, each image point ${}^{S_i}\mathbf{x}$ defines a ray in three-dimensional space, and in the absence of measurement errors all N rays intersect in the scene point ${}^W\mathbf{x}$.

First general scene reconstruction methods based on images acquired from different views were developed e.g. by Kruppa (1913) and Finsterwalder (1899). Overviews of these early methods are given by Aström (1996) and Luhmann (2003). They aim for a determination of intrinsic and extrinsic camera parameters and the three-dimensional coordinates of the scene points. Kruppa (1913) presents an analytical solution for the scene structure and extrinsic camera parameters from a minimal set of five corresponding image points.

Classical bundle adjustment methods (Brown, 1958; Luhmann, 2003; Lourakis and Argyros, 2004) jointly recover scene points and camera parameters from a set of K corresponding image points. The measured image coordinates of the scene

points in the images of the N cameras are denoted by the sensor coordinates ${}^{S_i}\mathbf{x}_k$, where $i = 1, \dots, N$ and $k = 1, \dots, K$. The image coordinates inferred from the extrinsic camera parameters ${}^C_i T$, the intrinsic camera parameters $\{c_j\}_i$, and the K scene point coordinates ${}^W \mathbf{x}_k$ are given by Eq. (1.6). Bundle adjustment corresponds to a minimisation of the reprojection error

$$E_B = \sum_{i=1}^N \sum_{k=1}^K \left\| {}^{I_i} T^{-1} \left(\mathcal{P} \left({}^C_i T, \{c_j\}_i, {}^W \mathbf{x}_k \right) - {}^{S_i} \mathbf{x}_k \right) \right\|^2. \quad (1.11)$$

The transformation by ${}^{I_i} T^{-1}$ in Eq. (1.11) ensures that the backprojection error is measured in Cartesian image coordinates. It can be omitted if a film is used for image acquisition, on which Euclidean distances are measured in a Cartesian coordinate system, or as long as the pixel raster of the digital camera sensor is orthogonal ($\theta = 90^\circ$) and the pixels are quadratic ($\alpha_u = \alpha_v$). This special case corresponds to ${}^{S_i} T$ in Eq. (1.5) describing a similarity transform.

The bundle adjustment approach can be used for calibration of the intrinsic and extrinsic camera parameters, reconstruction of the three-dimensional scene structure, or estimation of object pose. Depending on the scenario, some or all of the parameters ${}^C_i T$, $\{c_j\}_i$, and ${}^W \mathbf{x}_k$ may be unknown and are obtained by a minimisation of the reprojection error E_B with respect to the unknown parameters. As long as the scene is static, utilising N simultaneously acquired images (stereo image analysis, cf. Section 1.3) is equivalent to evaluating a sequence of N images acquired by a single moving camera (structure from motion).

Minimisation of Eq. (1.11) involves nonlinear optimisation techniques such as the Gauss-Newton or the Levenberg-Marquardt approach (Press et al., 1992). The reprojection error of scene point ${}^W \mathbf{x}_k$ in image i influences the values of ${}^C_i T$ and $\{c_j\}_i$ only for images in which this scene point is also detected, leading to a sparse set of nonlinear equations. The sparsity of the optimisation problem is exploited in the algorithm by Lourakis and Argyros (2004). The error function defined by Eq. (1.11) may have a large number of local minima, such that reasonable initial guesses for the parameters to be estimated have to be provided. As long as no a priori knowledge about the camera positions is available, a general property of the bundle adjustment method is that it only recovers the scene structure up to an unknown constant scale factor, since an increase of the mutual distances between the scene points by a constant factor can be compensated by accordingly increasing the mutual distances between the cameras and their distances to the scene. However, this scale factor can be obtained if additional information about the scene, such as the distance between two scene points, is known.

Difficulties may occur in the presence of false correspondences or gross errors of the determined point positions in the images, corresponding to strong deviations of the distribution of reprojection errors from the assumed Gaussian distribution. Lourakis and Argyros (2004) point out that in realistic scenarios the assumption of a Gaussian distribution of the measurement errors systematically underestimates the fraction of large errors. Searching for outliers in the established correspondences can be performed e.g. using the random sample consensus (RANSAC) method (Fischler

and Bolles, 1981) in combination with a minimal case five point algorithm (Nister, 2004). Alternatively, it is often useful to reduce the weight of large reprojection errors, which corresponds to replacing the L_2 norm in Eq. (1.11) by a suitable different norm. This optimisation approach is termed M -estimator technique (Rey, 1983).

A further drawback of the correspondence-based geometric bundle adjustment approach is the fact that correspondences can only be reliably extracted in textured image parts, leading to a sparse three-dimensional reconstruction result in the presence of large weakly or repetitively textured regions.

1.3 Geometric Aspects of Stereo Image Analysis

The reconstruction of three-dimensional scene structure based on two images acquired from different positions and viewing directions is termed stereo image analysis. In this section we will regard the “classcial” Euclidean approach to this important field of image-based three-dimensional scene reconstruction (cf. Section 1.3.1) as well as its formulation in terms of projective geometry (cf. Section 1.3.2).

1.3.1 Euclidean Formulation of Stereo Image Analysis

In this section, we begin with an introduction in terms of Euclidean geometry, essentially following the derivation described by Horn (1986). We assume that the world coordinate system is identical with the coordinate system of camera 1, i.e. the transformation matrix ${}^C_W T$ corresponds to unity while the relative orientation of camera 2 with respect to camera 1 is given by ${}^C_W T$ and is assumed to be known. In Section 1.4 we will regard the problem of camera calibration, i.e. the determination of the extrinsic and intrinsic camera parameters. A point ${}^I \mathbf{x} = (\hat{u}_1, \hat{v}_1)^T$ in image 1 corresponds to a ray through the origin of the camera coordinate system according to

$${}^{C_1} \mathbf{x} = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} \hat{u}_1 s \\ \hat{v}_1 s \\ b s \end{pmatrix}, \quad (1.12)$$

where s is assumed to be a positive real number. In the coordinate system of camera 2, according to Eq. (1.2) the points on this ray have the coordinates

$${}^{C_2} \mathbf{x} = \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} = R {}^{C_1} \mathbf{x} + \mathbf{t} = \begin{pmatrix} (r_{11}\hat{u}_1 + r_{12}\hat{v}_1 + r_{13}b)s + t_1 \\ (r_{21}\hat{u}_1 + r_{22}\hat{v}_1 + r_{23}b)s + t_2 \\ (r_{31}\hat{u}_1 + r_{32}\hat{v}_1 + r_{33}b)s + t_3 \end{pmatrix} \quad (1.13)$$

with r_{ij} as the elements of the orthonormal rotation matrix R and t_i as the elements of the translation vector \mathbf{t} (cf. Eq. (1.2)). In the image coordinate system of camera 2, the coordinates of the vector ${}^I \mathbf{x} = (\hat{u}_2, \hat{v}_2)^T$ are given by

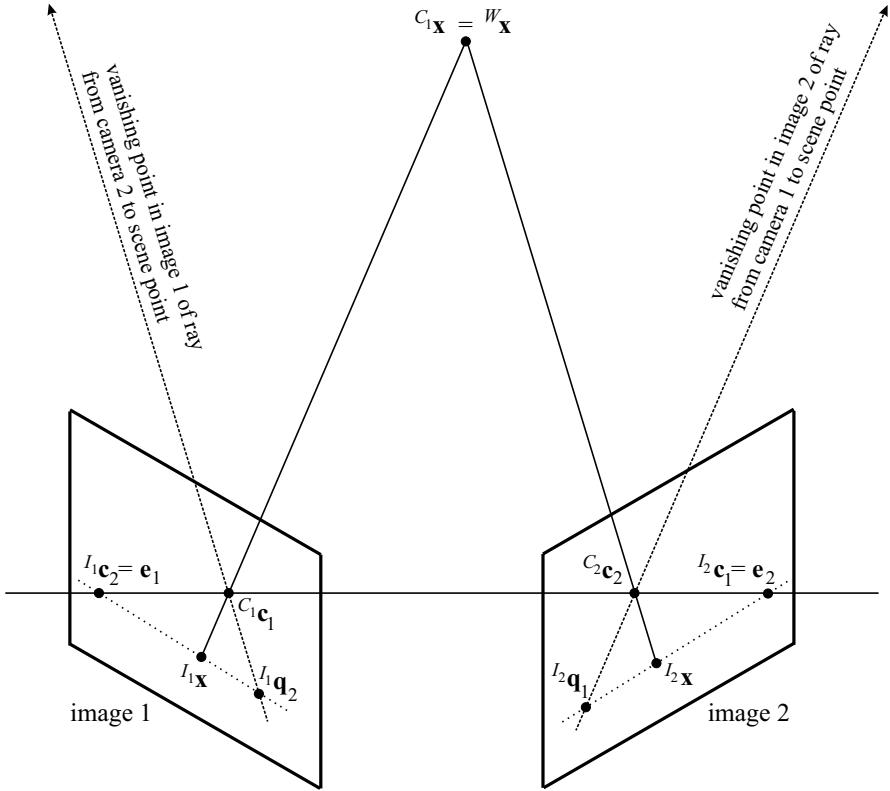


Fig. 1.2 Definition of epipolar geometry. The epipolar lines of the image points $I_1 \mathbf{x}$ and $I_2 \mathbf{x}$ are drawn as dotted lines, respectively.

$$\frac{\hat{u}_2}{b} = \frac{x_2}{z_2} \quad \text{and} \quad \frac{\hat{v}_2}{b} = \frac{y_2}{z_2}, \quad (1.14)$$

assuming identical principal distances for both cameras. With the abbreviations

$$x_2 = ds + p, \quad y_2 = es + q, \quad z_2 = fs + r$$

we now obtain the relations

$$\begin{aligned} \frac{\hat{u}_2}{b} &= \frac{d}{f} + \frac{fp - dr}{f} \frac{1}{fs + r} \\ \frac{\hat{v}_2}{b} &= \frac{e}{f} + \frac{fq - er}{c} \frac{1}{fs + r} \end{aligned}$$

describing a straight line connecting the point $(p/r, q/r)^T$ for $s = 0$ with the point $(d/f, e/f)^T$ for $s \rightarrow \infty$. The first of these points is the image of the principal point of camera 1, i.e. the origin of camera coordinate system 1, in the image of camera 2,

while the second point corresponds to the vanishing point of the ray in camera 2. The straight line describes a ray from the principal point of camera 2 which is parallel to the given ray through the principal point of camera 1. These geometric relations are illustrated in Fig. 1.2. The optical centre of camera 1 is at ${}^{C_1}\mathbf{c}_1$, and the scene point ${}^W\mathbf{x} = {}^{C_1}\mathbf{x}$ projects into the point ${}^{I_1}\mathbf{x}$ in image 1. The optical centre ${}^{C_2}\mathbf{c}_2$ of camera 2 is projected to ${}^{I_1}\mathbf{c}_2$ in image 1, and the vanishing point in image 1 of the ray from camera 2 to the scene point ${}^W\mathbf{x}$ is given by ${}^{I_1}\mathbf{q}_2$. The image points ${}^{I_1}\mathbf{c}_2$, ${}^{I_1}\mathbf{x}$, and ${}^{I_1}\mathbf{q}_2$ are located on a straight line, which corresponds to the intersection line between the image plane and a plane through the scene point ${}^W\mathbf{x}$ and the optical centres ${}^{C_1}\mathbf{c}_1$ and ${}^{C_2}\mathbf{c}_2$. A similar line is obtained for image 2. These lines are termed epipolar lines. A scene point projected to a point on the epipolar line in image 1 is always located on the corresponding epipolar line in image 2 constructed according to Fig. 1.2. This restriction on the image positions of corresponding image points is termed epipolar constraint. Each epipolar line is the intersection line of the image plane with an epipolar plane, i.e. a plane which contains the optical centres of both cameras. In image 1, all epipolar lines intersect in the image point ${}^{I_1}\mathbf{c}_2$ of the optical centre of camera 2, and vice versa. For real camera systems, the image plane may be of limited extent and will not always include the image of the optical centre of the other camera, respectively.

As long as the extrinsic relative camera orientation given by the rotation matrix R and the translation vector \mathbf{t} are known, it is straightforward to compute the three-dimensional position of a scene point ${}^W\mathbf{x}$ with image coordinates ${}^{I_1}\mathbf{x} = (\hat{u}_1, \hat{v}_1)^T$ and ${}^{I_2}\mathbf{x} = (\hat{u}_2, \hat{v}_2)^T$, expressed as ${}^{C_1}\mathbf{x}$ and ${}^{C_2}\mathbf{x}$ in the two camera coordinate systems. It follows from Eqs. (1.13) and (1.14) that

$$\begin{aligned} \left(r_{11} \frac{\hat{u}_1}{b} + r_{12} \frac{\hat{v}_1}{b} + r_{13} \right) z_1 + t_1 &= \frac{\hat{u}_2}{b} z_2 \\ \left(r_{21} \frac{\hat{u}_1}{b} + r_{22} \frac{\hat{v}_1}{b} + r_{23} \right) z_1 + t_2 &= \frac{\hat{v}_2}{b} z_2 \\ \left(r_{31} \frac{\hat{u}_1}{b} + r_{32} \frac{\hat{v}_1}{b} + r_{33} \right) z_1 + t_3 &= z_2. \end{aligned}$$

Any two of these equations can be used to solve for z_1 and z_2 . For the three-dimensional positions of the scene points we then obtain

$$\begin{aligned} {}^{C_1}\mathbf{x} &= \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} \hat{u}_1/b \\ \hat{v}_1/b \\ 1 \end{pmatrix} z_1 \\ {}^{C_2}\mathbf{x} &= \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} = \begin{pmatrix} \hat{u}_2/b \\ \hat{v}_2/b \\ 1 \end{pmatrix} z_2. \end{aligned} \quad (1.15)$$

Eq. (1.15) allows to compute the coordinates ${}^{C_i}\mathbf{x}$ of a scene point in any of the two camera coordinate systems based on the measured pixel positions of the corresponding image points, given the relative orientation of the cameras defined by

the rotation matrix R and the translation vector \mathbf{t} . Note that all computations in this section have been performed based on the metric image coordinates given by ${}^I_i \mathbf{x} = (\hat{u}_i, \hat{v}_i)^T$, which are related to the pixel coordinates given by ${}^S_i \mathbf{x} = (u_i, v_i)^T$ in the sensor coordinate system by Eq. (1.5).

1.3.2 Stereo Image Analysis in Terms of Projective Geometry

At this point it is illustrative to regard the derivation of the epipolar constraint in the framework of projective geometry. Two cameras regard a scene point ${}^W \tilde{\mathbf{x}}$ which is projected into the vectors ${}^I_1 \tilde{\mathbf{x}}'$ and ${}^I_2 \tilde{\mathbf{x}}'$ defined in the two image coordinate systems. Since these vectors are defined in homogeneous coordinates, ${}^W \tilde{\mathbf{x}}$ is of size 4×1 while ${}^I_1 \tilde{\mathbf{x}}'$ and ${}^I_2 \tilde{\mathbf{x}}'$ are of size 3×1 . The cameras are assumed to be pinhole cameras with the same principal distance b , and ${}^I_1 \tilde{\mathbf{x}}'$ and ${}^I_2 \tilde{\mathbf{x}}'$ are given in normalised coordinates (Birchfield, 1998), i.e. the vectors are scaled such that their last (third) coordinates are 1. Hence, their first two coordinates represent the position of the projected scene point in the image with respect to the principal point, measured in units of the principal distance b , respectively. As a result, the three-dimensional vectors ${}^I_1 \tilde{\mathbf{x}}'$ and ${}^I_2 \tilde{\mathbf{x}}'$ correspond to the Euclidean vectors from the optical centres to the projected points in the image planes.

The Essential Matrix

According to the epipolar constraint, the vector ${}^I_1 \tilde{\mathbf{x}}'$ from the first optical centre to the first projected point, the vector ${}^I_2 \tilde{\mathbf{x}}'$ from the second optical centre to the second projected point, and the vector \mathbf{t} connecting the two optical centres are coplanar. This condition can be expressed as

$${}^I_1 \tilde{\mathbf{x}}'^T (\mathbf{t} \times R {}^I_2 \tilde{\mathbf{x}}') = 0, \quad (1.16)$$

where R and \mathbf{t} denote the rotational and translational part of the coordinate transformation from the first into the second camera coordinate system. We now define $[\mathbf{t}]_\times$ as the 3×3 matrix for which we have $[\mathbf{t}]_\times \mathbf{y} = \mathbf{t} \times \mathbf{y}$ for any 3×1 vector \mathbf{y} . The matrix $[\mathbf{t}]_\times$ is termed cross product matrix of the vector \mathbf{t} . For $\mathbf{t} = (d, e, f)^T$, it is straightforward to show that

$$[\mathbf{t}]_\times = \begin{bmatrix} 0 & -f & e \\ f & 0 & -d \\ -e & d & 0 \end{bmatrix} \quad (1.17)$$

(Birchfield, 1998). Eq. (1.16) can then be rewritten as

$${}^I_1 \tilde{\mathbf{x}}'^T ([\mathbf{t}]_\times R {}^I_2 \tilde{\mathbf{x}}') = {}^I_1 \tilde{\mathbf{x}}'^T E {}^I_2 \tilde{\mathbf{x}}' = 0, \quad (1.18)$$

where $E = [\mathbf{t}]_x R$ is termed essential matrix and describes the transformation from the coordinate system of one pinhole camera into the coordinate system of the other pinhole camera. Eq. (1.18) shows that the epipolar constraint can be written as a linear equation in homogeneous coordinates, and it completely describes the geometric relationship between corresponding points in a pair of stereo images. The essential matrix contains five parameters, three for the relative rotation between the cameras, two for the direction of translation. It is not possible to recover the absolute magnitude of translation as increasing the distance between the cameras can be compensated by increasing the depth of the scene point by the same amount, thus leaving the coordinates of the image points unchanged. The determinant of the essential matrix is zero, and its two non-zero eigenvalues are equal (Birchfield, 1998).

The Fundamental Matrix

We now assume that the image points are not given in normalised coordinates but in sensor pixel coordinates by the projective 3×1 vectors ${}^{S_1}\tilde{\mathbf{x}}$ and ${}^{S_2}\tilde{\mathbf{x}}$. If the lenses are assumed to be distortion-free, the transformation from the normalised camera coordinate system into the sensor coordinate system is given by Eq. (1.9), leading to the linear relations

$$\begin{aligned} {}^{S_1}\tilde{\mathbf{x}} &= A_1 {}^{I_1}\tilde{\mathbf{x}}' \\ {}^{S_2}\tilde{\mathbf{x}} &= A_2 {}^{I_2}\tilde{\mathbf{x}}'. \end{aligned} \quad (1.19)$$

The matrices A_1 and A_2 contain the pixel size, pixel skew, and pixel coordinates of the principal point of the cameras, respectively. If lens distortion has to be taken into account e.g. according to Eqs. (1.3) and (1.4), the corresponding transformations may become nonlinear. Eqs. (1.18) and (1.19) yield the expressions

$$\begin{aligned} (A_2^{-1} {}^{S_2}\tilde{\mathbf{x}})^T (\mathbf{t} \times RA_1^{-1} {}^{S_1}\tilde{\mathbf{x}}) &= 0 \\ {}^{S_2}\tilde{\mathbf{x}}^T A_2^{-T} (\mathbf{t} \times RA_1^{-1} {}^{S_1}\tilde{\mathbf{x}}) &= 0 \\ {}^{S_2}\tilde{\mathbf{x}}^T F {}^{S_1}\tilde{\mathbf{x}} &= 0, \end{aligned} \quad (1.20)$$

where $F = A_2^{-T} EA_1^{-1}$ is termed fundamental matrix and provides a representation of both the intrinsic and the extrinsic parameters of the two cameras. The matrix F is always of rank 2 (Hartley and Zisserman, 2003), i.e. one of its eigenvalues is always zero. Eq. (1.20) is valid for all corresponding image points ${}^{S_1}\tilde{\mathbf{x}}$ and ${}^{S_2}\tilde{\mathbf{x}}$ in the images.

The fundamental matrix F relates a point in one stereo image to the line of all points in the other stereo image that may correspond to that point according to the epipolar constraint. In a projective plane, a line $\tilde{\mathbf{l}}$ is defined such that for all points $\tilde{\mathbf{x}}$ on the line the relation $\tilde{\mathbf{x}}^T \tilde{\mathbf{l}} = 0$ is fulfilled (Birchfield, 1998). At the same time, this relation indicates that in a projective plane, points and lines have the same representation and are thus dual with respect to each other. Especially, the epipolar line

$S_2\tilde{\mathbf{I}}$ in image 2 which corresponds to a point $S_1\tilde{\mathbf{x}}$ in image 1 is given by $S_2\tilde{\mathbf{I}} = F S_1\tilde{\mathbf{x}}$. Eq. (1.20) immediately shows that this relation must hold since all points $S_2\tilde{\mathbf{x}}$ in image 2 which may correspond to the point $S_1\tilde{\mathbf{x}}$ in image 1 are located on the line $S_2\tilde{\mathbf{I}}$. Accordingly, the line $S_1\tilde{\mathbf{I}} = F^T S_2\tilde{\mathbf{x}}$ in image 1 is the epipolar line corresponding to the point $S_1\tilde{\mathbf{x}}$ in image 2.

For an arbitrary point $S_1\tilde{\mathbf{x}}$ in image 1 except the epipole $\tilde{\mathbf{e}}_1$, the epipolar line $S_2\tilde{\mathbf{I}} = F S_1\tilde{\mathbf{x}}$ contains the epipole $\tilde{\mathbf{e}}_2$ in image 2 (Hartley and Zisserman, 2003). The epipoles $\tilde{\mathbf{e}}_1$ and $\tilde{\mathbf{e}}_2$ are defined in the sensor coordinate system of camera 1 and 2, respectively. We thus have $\tilde{\mathbf{e}}_2^T (F S_1\tilde{\mathbf{x}}) = (\tilde{\mathbf{e}}_2^T F) S_1\tilde{\mathbf{x}} = 0$ for all $S_1\tilde{\mathbf{x}}$, which implies $\tilde{\mathbf{e}}_2^T F = 0$. Accordingly, $\tilde{\mathbf{e}}_2$ is the left null-vector of F , corresponding to the eigenvector belonging to the zero eigenvalue of F^T . The epipole $\tilde{\mathbf{e}}_1$ in image 1 is given by the right null-vector of F according to $F\tilde{\mathbf{e}}_1 = 0$, i.e. it corresponds to the eigenvector belonging to the zero eigenvalue of F .

Projective Reconstruction of the Scene

In the framework of projective geometry, image formation by a pinhole camera is defined by the projection matrix P of size 3×4 as defined in Eq. (1.10). A projective scene reconstruction by two cameras is defined by $(P_1, P_2, \{{}^W\tilde{\mathbf{x}}_i\})$, where P_1 and P_2 denote the projection matrix of camera 1 and 2, respectively, and $\{{}^W\tilde{\mathbf{x}}_i\}$ are the scene points reconstructed from a set of point correspondences. Hartley and Zisserman (2003) show that a projective scene reconstruction is always ambiguous up to a projective transformation H , where H is an arbitrary 4×4 matrix. Hence, the projective reconstruction given by $(P_1, P_2, \{{}^W\tilde{\mathbf{x}}_i\})$ is equivalent to the one defined by $(P_1H, P_2H, \{H^{-1} {}^W\tilde{\mathbf{x}}_i\})$.

It is possible to obtain the camera projection matrices P_1 and P_2 from the fundamental matrix F in a rather straightforward manner. Without loss of generality, the projection matrix P_1 may be chosen such that $P_1 = [I \mid \mathbf{0}]$, i.e. the rotation matrix R is the identity matrix and the translation vector \mathbf{t} is zero, such that the world coordinate system W corresponds to the coordinate system C_1 of camera 1. The projection matrix of the second camera then corresponds to

$$P_2 = \left[[\tilde{\mathbf{e}}_2]_{\times} F \mid \tilde{\mathbf{e}}_2 \right]. \quad (1.21)$$

A more general form of P_2 is

$$P_2 = \left[[\tilde{\mathbf{e}}_2]_{\times} F + \tilde{\mathbf{e}}_2 \mathbf{v}^T \mid \lambda \tilde{\mathbf{e}}_2 \right], \quad (1.22)$$

where \mathbf{v} is an arbitrary 3×1 vector and λ a non-zero scalar (Hartley and Zisserman, 2003). Eqs. (1.21) and (1.22) show that the fundamental matrix F and the epipole $\tilde{\mathbf{e}}_2$, which is uniquely determined by F since it corresponds to its left null-vector, determine a projective reconstruction of the scene.

If two corresponding image points are situated exactly on their respective epipolar lines, Eq. (1.20) is exactly fulfilled, such that the rays described by the image

points ${}^S_1\tilde{\mathbf{x}}$ and ${}^S_2\tilde{\mathbf{x}}$ intersect in the point ${}^W\tilde{\mathbf{x}}$ which can be determined by triangulation in a straightforward manner. We will return to this scenario in Section 1.5 in the context of stereo image analysis in standard geometry, where the fundamental matrix F is assumed to be known. The search for point correspondences only takes place along corresponding epipolar lines, such that the world coordinates of the resulting scene points are obtained by direct triangulation. If, however, an unrestricted search for correspondences is performed, Eq. (1.20) is generally not exactly fulfilled due to noise in the measured coordinates of the corresponding points, and the rays defined by them do not intersect. The projective scene point ${}^W\tilde{\mathbf{x}}$ in the world coordinate system is obtained from ${}^S_1\tilde{\mathbf{x}}$ and ${}^S_2\tilde{\mathbf{x}}$ based on the relations ${}^S_1\tilde{\mathbf{x}} = P_1 {}^W\tilde{\mathbf{x}}$ and ${}^S_2\tilde{\mathbf{x}} = P_2 {}^W\tilde{\mathbf{x}}$, which can be combined into a linear equation of the form $G {}^W\tilde{\mathbf{x}} = 0$. The homogeneous scale factor is eliminated by computing the cross product ${}^S_1\tilde{\mathbf{x}} \times (P_1 {}^W\tilde{\mathbf{x}}) = 0$, which allows to express the matrix G as

$$G = \begin{bmatrix} u_1 \tilde{\mathbf{p}}_1^{(3)T} - \tilde{\mathbf{p}}_1^{(1)T} \\ v_1 \tilde{\mathbf{p}}_1^{(3)T} - \tilde{\mathbf{p}}_1^{(2)T} \\ u_2 \tilde{\mathbf{p}}_2^{(3)T} - \tilde{\mathbf{p}}_2^{(1)T} \\ v_2 \tilde{\mathbf{p}}_2^{(3)T} - \tilde{\mathbf{p}}_2^{(2)T} \end{bmatrix}, \quad (1.23)$$

where ${}^S_1\tilde{\mathbf{x}} = (u_1, v_1, 1)^T$, ${}^S_2\tilde{\mathbf{x}} = (u_2, v_2, 1)^T$, and $\tilde{\mathbf{p}}_i^{(j)T}$ corresponds to the j th row of the camera projection matrix P_i . The linear system of equations $G {}^W\tilde{\mathbf{x}} = 0$ is overdetermined since ${}^W\tilde{\mathbf{x}}$ only has three independent components due to its arbitrary projective scale, and generally only a least-squares solution exists due to noise in the measurements of ${}^S_1\tilde{\mathbf{x}}$ and ${}^S_2\tilde{\mathbf{x}}$. The solution for ${}^W\tilde{\mathbf{x}}$ corresponds to the unit singular vector that belongs to the smallest singular value of G .

However, as merely an algebraic error rather than a physically motivated geometric error is minimised by this linear approach to determine ${}^W\tilde{\mathbf{x}}$, Hartley and Zisserman (2003) suggest a projective reconstruction of the scene points by minimisation of the backprojection error in the sensor coordinate system. While ${}^S_1\tilde{\mathbf{x}}$ and ${}^S_2\tilde{\mathbf{x}}$ correspond to the measured image coordinates of a pair of corresponding points, the estimated point correspondences which exactly fulfill the epipolar constraint (1.20) are denoted by ${}^S_1\tilde{\mathbf{x}}^{(e)}$ and ${}^S_2\tilde{\mathbf{x}}^{(e)}$. We thus have ${}^S_2\tilde{\mathbf{x}}^{(e)T} F {}^S_1\tilde{\mathbf{x}}^{(e)} = 0$. The point ${}^S_1\tilde{\mathbf{x}}^{(e)}$ lies on an epipolar line ${}^S_1\tilde{\mathbf{l}}$ and ${}^S_2\tilde{\mathbf{x}}^{(e)}$ lies on the corresponding epipolar line ${}^S_2\tilde{\mathbf{l}}$. However, any other pair of points lying on the lines ${}^S_1\tilde{\mathbf{l}}$ and ${}^S_2\tilde{\mathbf{l}}$ also satisfies the epipolar constraint. Hence, the points ${}^S_1\tilde{\mathbf{x}}^{(e)}$ and ${}^S_2\tilde{\mathbf{x}}^{(e)}$ have to be determined such that the sum of the squared Euclidean distances $d^2({}^S_1\tilde{\mathbf{x}}, {}^S_1\tilde{\mathbf{l}})$ and $d^2({}^S_2\tilde{\mathbf{x}}, {}^S_2\tilde{\mathbf{l}})$ in the sensor coordinate system between ${}^S_1\tilde{\mathbf{x}}$ and ${}^S_1\tilde{\mathbf{l}}$ and between ${}^S_2\tilde{\mathbf{x}}$ and ${}^S_2\tilde{\mathbf{l}}$, respectively, i.e. the backprojection error, is minimised. Here, $d({}^S\tilde{\mathbf{x}}, {}^S\tilde{\mathbf{l}})$ denotes the perpendicular distance between the point ${}^S\tilde{\mathbf{x}}$ and the line ${}^S\tilde{\mathbf{l}}$. This minimisation approach is equivalent to bundle adjustment as long as the distance $d({}^S\tilde{\mathbf{x}}, {}^S\tilde{\mathbf{l}})$ is an Euclidean distance in the image plane rather than merely in the sensor coordinate system, which is the case for image sensors with zero skew and square pixels.

In each of the two images, the epipolar lines in the two images form a so-called pencil of lines, which is an infinite number of lines which all intersect in the same

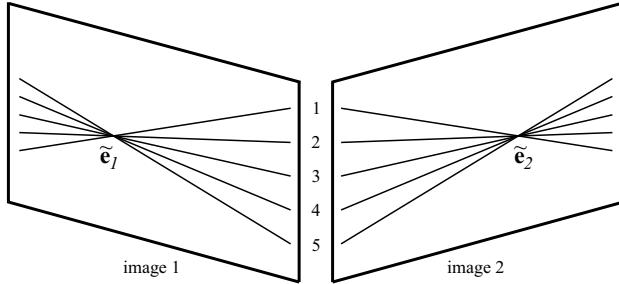


Fig. 1.3 In each of the two images, the epipolar lines form a pencil of lines. The intersection points correspond to the epipoles \tilde{e}_1 and \tilde{e}_2 . Corresponding pairs of epipolar lines are numbered consecutively.

point (cf. Fig. 1.3). For the pencils of epipolar lines in image 1 and 2, the intersection points correspond to the epipoles \tilde{e}_1 and \tilde{e}_2 . Hence, the pencil of epipolar lines can be parameterised by a single parameter t , such that an epipolar line in image 1 can be written as ${}^S_1\tilde{\mathbf{l}}(t)$. The corresponding epipolar line ${}^S_2\tilde{\mathbf{l}}(t)$ in image 2 is readily obtained based on the fundamental matrix F . Now the backprojection error term can be formulated as $d^2({}^S_1\tilde{\mathbf{x}}, {}^S_1\tilde{\mathbf{l}}(t)) + d^2({}^S_2\tilde{\mathbf{x}}, {}^S_2\tilde{\mathbf{l}}(t))$ and thus becomes a function of the single scalar variable t . Minimising the error term with respect to t effectively corresponds to finding the real roots of a polynomial of degree 6 (Hartley and Zisserman, 2003). The next step consists of selecting the points ${}^S_1\tilde{\mathbf{x}}^{(e)}$ and ${}^S_2\tilde{\mathbf{x}}^{(e)}$ which are closest to the lines ${}^S_1\tilde{\mathbf{l}}(t_{\min})$ and ${}^S_2\tilde{\mathbf{l}}(t_{\min})$, respectively, in terms of the Euclidean distance in the sensor coordinate system. The projective scene point ${}^W\tilde{\mathbf{x}}$ in the world coordinate system is obtained by replacing the measured normalised image point coordinates (u_1, v_1) and (u_2, v_2) in Eq. (1.23) by the normalised coordinates $(u_1^{(e)}, v_1^{(e)})$ and $(u_2^{(e)}, v_2^{(e)})$ of the estimated image points ${}^S_1\tilde{\mathbf{x}}^{(e)}$ and ${}^S_2\tilde{\mathbf{x}}^{(e)}$. Then an exact solution and not just a least-squares solution of the linear system of equations $G {}^W\tilde{\mathbf{x}} = 0$ with G given by Eq. (1.23) exists since the estimated image points ${}^S_1\tilde{\mathbf{x}}^{(e)}$ and ${}^S_2\tilde{\mathbf{x}}^{(e)}$ have been constructed such that they fulfill the epipolar constraint exactly, and the rays defined by ${}^S_1\tilde{\mathbf{x}}^{(e)}$ and ${}^S_2\tilde{\mathbf{x}}^{(e)}$ intersect in the point ${}^W\tilde{\mathbf{x}}$. Hence, in this case the solution for ${}^W\tilde{\mathbf{x}}$ is the unit singular vector of G that belongs to its zero singular value.

Estimating the fundamental matrix F and, accordingly, the projective camera matrices P_1 and P_2 and the projective scene points ${}^W\tilde{\mathbf{x}}_i$ from a set of point correspondences between the images can be regarded as the first (projective) stage of camera calibration. Subsequent calibration stages consist of determining a metric (Euclidean) scene reconstruction and camera calibration. These issues will be regarded further in Section 1.4.6 in the context of self-calibration of camera systems.

1.4 Geometric Calibration of Single and Multiple Cameras

Camera calibration aims for a determination of the transformation parameters between the camera lens and the image plane as well as between the camera and the scene based on the acquisition of images of a calibration rig with a known spatial structure. In photogrammetry, the transformation between the camera lens and the image plane is termed interior orientation. It is characterised by the matrix A of the intrinsic camera parameters, which contains the principal distance b , the pixel position (u_0, v_0) of the principal point in the image plane, the direction-dependent pixel scale, the pixel skew, and the lens distortion parameters (cf. Section 1.1). The exterior orientation of the camera, i.e. its orientation with respect to the scene, is defined by the rotation matrix R and the translation vector t which relate the camera coordinate system and the world coordinate system to each other as outlined in Section 1.1.

In this section, we first outline early camera calibration approaches exclusively devoted to the determination of the intrinsic camera parameters. We then describe classical techniques for simultaneous intrinsic and extrinsic camera calibration which are especially suited for fast and reliable calibration of standard video cameras and lenses which are commonly used in computer vision applications (Tsai, 1987; Zhang, 1999a; Bouguet, 2007). Furthermore, a short overview of self-calibration techniques is given. The section is concluded by a description of the semi-automatic calibration procedure for multi-camera systems introduced by Krüger et al. (2004), which is based on a fully automatic extraction of control points from the calibration images.

1.4.1 Methods for Intrinsic Camera Calibration

According to the detailed survey by Clarke and Fryer (1998), early approaches to camera calibration in the field of aerial photography in the first half of the 20th century mainly dealt with the determination of the intrinsic camera parameters, which was carried out in a laboratory. This was feasible in practice due to the fact that aerial (metric) camera lenses are focused to infinity in a fixed manner and do not contain iris elements. The principal distance, in this case being equal to the focal length, was computed by observing the angles through the lens to a grid plate displaying finely etched crosses. By analysing the values for the principal distance obtained along several radial lines in the image plane, an average “calibrated” value was selected that best compensated the effects of radial distortion, which was only taken into account in an implicit manner. The principal point was determined based on an autocollimation method. In stereoplottting devices, radial distortion was compensated by optical correction elements. Due to the low resolution of the film used for image acquisition, there was no need to take into account tangential distortion.

In these scenarios, important sources of calibration errors are the considerable difference in temperature between the laboratory and during flight, leading e.g. to

insufficient flatness of the glass plates still used for photography at that time and irregular film shrinkage (Hothmer, 1958). Hence, so-called field calibration techniques were introduced in order to determine the camera parameters under the conditions encountered during image acquisition. Radial distortion curves were produced based on stereo images of the flat surfaces of frozen lakes which were just about to melt, thus showing a sufficient amount of texture on their icy surfaces to facilitate stereo analysis. Other field calibration techniques rely on terrestrial control points (Merrit, 1948). A still different method is based on the well-known angular positions of stars visible in the image (Schmid, 1974). Although this method turned out to yield very accurate calibration results, an essential drawback is the necessity to identify each star and to take into account corrections for atmospheric refraction and diurnal aberration.

An analytic model of radial and tangential lens distortion based on a power series expansion has been introduced by Brown (1958) and by Brown (1966), which is still utilised in modern calibration approaches (cf. also Eqs. (1.3) and (1.4)). These approaches involve the simultaneous determination of lens parameters, extrinsic camera orientation, and coordinates of control points in the scene in the camera coordinate system, based on the bundle adjustment method. A different method for the determination of radial and tangential distortion parameters is plumb line calibration (Brown, 1971), exploiting the fact that straight lines in the real world remain straight in the image. Radial and tangential distortions can directly be inferred from deviations from straightness in the image. These first calibration methods based on bundle adjustment with additional parameters for lens distortion, focal length, position of the principal point, flatness of the photographic plate, and film shrinkage (Brown, 1958, 1966, 1971) are usually termed on-the-job calibration (Clarke and Fryer, 1998).

1.4.2 The Direct Linear Transform (DLT) Method

In its simplest form, the direct linear transform (DLT) calibration method (Abdel-Aziz and Karara, 1971) aims for a determination of the intrinsic and extrinsic camera parameters according to Eq. (1.1). This goal is achieved by establishing an appropriate transformation which translates the world coordinates of known control points in the scene into image coordinates. An illustrative description of the DLT method is given by Kwon (1998). The DLT method assumes a pinhole camera, for which, according to the introduction given in Section 1.1, it is straightforward to derive the relation

$$\begin{pmatrix} \hat{u} \\ \hat{v} \\ -b \end{pmatrix} = c R \begin{pmatrix} x - x_0 \\ y - y_0 \\ z - z_0 \end{pmatrix}. \quad (1.24)$$

In Eq. (1.24), R denotes the rotation matrix that relates the world coordinate system to the camera coordinate system as described in Section 1.1, \hat{u} and \hat{v} the metric pixel coordinates in the image plane relative to the principal point, and x, y, z are the

components of a scene point ${}^W\mathbf{x}$ in the world coordinate system. The values x_0, y_0 , and z_0 can be inferred from the translation vector \mathbf{t} introduced in Section 1.1, while c is a scalar scale factor. This scale factor amounts to

$$c = -\frac{b}{r_{31}(x-x_0) + r_{32}(y-y_0) + r_{33}(z-z_0)}, \quad (1.25)$$

where the coefficients r_{ij} denote the elements of the rotation matrix R . Assuming rectangular sensor pixels without skew, the coordinates of the image point in the sensor coordinate system, i.e. the pixel coordinates, are given by $u - u_0 = k_u \hat{u}$ and $v - v_0 = k_v \hat{v}$, where u_0 and v_0 denote the position of the principal point in the sensor coordinate system. Inserting Eq. (1.25) into Eq. (1.24) then yields the relations

$$\begin{aligned} u - u_0 &= -\frac{b}{k_u} \frac{r_{11}(x-x_0) + r_{12}(y-y_0) + r_{13}(z-z_0)}{r_{31}(x-x_0) + r_{32}(y-y_0) + r_{33}(z-z_0)} \\ v - v_0 &= -\frac{b}{k_v} \frac{r_{21}(x-x_0) + r_{22}(y-y_0) + r_{23}(z-z_0)}{r_{31}(x-x_0) + r_{32}(y-y_0) + r_{33}(z-z_0)} \end{aligned} \quad (1.26)$$

Rearranging Eq. (1.26) results in expressions for the pixel coordinates u and v which only depend on the coordinates x, y , and z of the scene point and eleven constant parameters that comprise intrinsic and extrinsic camera parameters:

$$\begin{aligned} u &= \frac{L_1 x + L_2 y + L_3 z + L_4}{L_9 x + L_{10} y + L_{11} z + 1} \\ v &= \frac{L_5 x + L_6 y + L_7 z + L_8}{L_9 x + L_{10} y + L_{11} z + 1}. \end{aligned} \quad (1.27)$$

If we use the abbreviations $b_u = b/k_u$, $b_v = b/k_v$, and $D = -(x_0 r_{31} + y_0 r_{32} + z_0 r_{33})$, the parameters $L_1 \dots L_{11}$ can be expressed as

$$\begin{aligned} L_1 &= \frac{u_0 r_{31} - b_u r_{11}}{D} \\ L_2 &= \frac{u_0 r_{32} - b_u r_{12}}{D} \\ L_3 &= \frac{u_0 r_{33} - b_u r_{13}}{D} \\ L_4 &= \frac{(b_u r_{11} - u_0 r_{31}) x_0 + (b_u r_{12} - u_0 r_{32}) y_0 + (b_u r_{13} - u_0 r_{33}) z_0}{D} \\ L_5 &= \frac{v_0 r_{31} - b_v r_{21}}{D} \\ L_6 &= \frac{v_0 r_{32} - b_v r_{22}}{D} \\ L_7 &= \frac{v_0 r_{33} - b_v r_{23}}{D} \\ L_8 &= \frac{(b_v r_{21} - v_0 r_{31}) x_0 + (b_v r_{22} - v_0 r_{32}) y_0 + (b_v r_{23} - v_0 r_{33}) z_0}{D} \end{aligned}$$

$$\begin{aligned} L_9 &= \frac{r_{31}}{D} \\ L_{10} &= \frac{r_{32}}{D} \\ L_{11} &= \frac{r_{33}}{D} \end{aligned} \quad (1.28)$$

It is straightforward but somewhat tedious to compute the intrinsic and extrinsic camera parameters from these expressions for $L_1 \dots L_{11}$.

Radial and tangential distortions introduce offsets Δu and Δv with respect to the position of the image point expected according to the pinhole model. Using the polynomial laws defined in Eqs. (1.3) and (1.4) and setting $\xi = u - u_0$ and $\eta = v - v_0$, these offsets can be formulated as

$$\begin{aligned} \Delta u &= \xi(L_{12}r^2 + L_{13}r^4 + L_{14}r^6) + L_{15}(r^2 + 2\xi^2) + L_{16}\eta\xi \\ \Delta v &= \eta(L_{12}r^2 + L_{13}r^4 + L_{14}r^6) + L_{15}\eta\xi + L_{16}(r^2 + 2\eta^2) \end{aligned} \quad (1.29)$$

(Kwon, 1998). The additional parameters $L_{12} \dots L_{14}$ describe the radial and L_{15} and L_{16} the tangential lens distortion, respectively.

By replacing in Eq. (1.27) the values of u by $u + \Delta u$ and v by $v + \Delta v$ and defining the abbreviation $Q_i = L_9x_i + L_{10}y_i + L_{11}z_i + 1$, where x_i , y_i and z_i denote the world coordinates of scene point i ($i = 1, \dots, N$), an equation for determining the parameters $L_1 \dots L_{16}$ is obtained according to

$$\left[\begin{array}{ccccccccccccc} \frac{x_1}{Q_1} & \frac{y_1}{Q_1} & \frac{z_1}{Q_1} & \frac{1}{Q_1} & 0 & 0 & 0 & 0 & \frac{-u_1x_1}{Q_1} & \frac{-u_1y_1}{Q_1} & \frac{-u_1z_1}{Q_1} \\ 0 & 0 & 0 & 0 & \frac{x_1}{Q_1} & \frac{y_1}{Q_1} & \frac{z_1}{Q_1} & \frac{1}{Q_1} & \frac{-v_1x_1}{Q_1} & \frac{-v_1y_1}{Q_1} & \frac{-v_1z_1}{Q_1} \\ \vdots & \vdots \\ \frac{x_N}{Q_N} & \frac{y_N}{Q_N} & \frac{z_N}{Q_N} & \frac{1}{Q_N} & 0 & 0 & 0 & 0 & \frac{-u_Nx_N}{Q_N} & \frac{-u_Ny_N}{Q_N} & \frac{-u_Nz_N}{Q_N} \\ 0 & 0 & 0 & 0 & \frac{x_N}{Q_N} & \frac{y_N}{Q_N} & \frac{z_N}{Q_N} & \frac{1}{Q_N} & \frac{-v_Nx_N}{Q_N} & \frac{-v_Ny_N}{Q_N} & \frac{-v_Nz_N}{Q_N} \end{array} \right] \begin{pmatrix} L_1 \\ L_2 \\ \vdots \\ L_{16} \end{pmatrix} = \begin{pmatrix} \frac{u_1}{Q_1} \\ \frac{v_1}{Q_1} \\ \vdots \\ \frac{u_N}{Q_N} \\ \frac{v_N}{Q_N} \end{pmatrix} \quad (1.30)$$

Eq. (1.30) is of the form

$$M \mathbf{L} = \mathbf{B}, \quad (1.31)$$

where M is a rectangular matrix of size $2N \times 16$, \mathbf{B} a column vector of length $2N$, and \mathbf{L} a column vector of length 16 containing the parameters $L_1 \dots L_{16}$. The number of control points in the scene required to solve Eq. (1.31) amounts to eight if all 16 parameters are desired to be recovered. In the absence of lens distortions, only eleven parameters need to be recovered based on at least six control points. It is of course favourable to utilise more than the minimum necessary number of control points since the measured pixel coordinates u_i and v_i are not error-free. In this case,

Eq. (1.31) is overdetermined, and the vector \mathbf{L} is obtained according to

$$\mathbf{L} = (M^T M)^{-1} M^T \mathbf{B}, \quad (1.32)$$

where the matrix $(M^T M)^{-1} M^T$ is the pseudoinverse of M . Eq. (1.32) yields a least-squares solution for the parameter vector \mathbf{L} . It is important to note that the coefficient matrix A in Eq. (1.31) contains the values Q_i , which in turn depend on the parameters L_9 , L_{10} , and L_{11} . Initial values for these parameters have to be chosen, and the solution (1.32) has to be computed iteratively.

It is worth noting that the control points must not be coplanar but have to obtain a volume in three-dimensional space if the projection of arbitrary scene points onto the image plane is required. Otherwise, the pseudoinverse of M does not exist. A reduced, two-dimensional DLT can be formulated by setting $z = 0$ in Eq. (1.27) for scene points situated on a plane in three-dimensional space. In this special case it is always possible to choose the world coordinate system such that $z = 0$ for all regarded scene points.

The DLT method is a simple and easy-to-use camera calibration method, but it has two essential drawbacks. The first one is that the computed elements of the matrix R do not form an orthonormal matrix, as it would be expected for a rotation matrix. Incorporating orthonormality constraints into the DLT scheme would require nonlinear optimisation methods instead of the simple iterative linear solution scheme defined by Eq. (1.32). Another drawback is the fact that the optimisation scheme is not equivalent to bundle adjustment. While bundle adjustment minimises the backprojection error in the image plane, Eq. (1.30) illustrates that the DLT method minimises the error of the backprojected scaled pixel coordinates $(u_i/Q_i, v_i/Q_i)$. It is not guaranteed that this somewhat arbitrary error measure is always a reasonable choice.

1.4.3 The Camera Calibration Method by Tsai (1987)

In contrast to the DLT method, the camera calibration method by Tsai (1987) minimises the backprojection error in the image plane. It recovers the intrinsic camera parameters, including the distortion parameters, and the extrinsic camera parameters based on a set of accurately known control points in the scene. The optimisation starts with closed-form linear estimates of some parameters which are used as initial estimates for a nonlinear optimisation scheme that determines all camera parameters simultaneously. The formulation of the method is slightly different for planar calibration rigs or calibration rigs covering a certain volume in space (Horn, 2000).

In the absence of lens distortion, one readily obtains from the pinhole model (cf. Section 1.1) the relation

$$\frac{\hat{u}}{b} = \frac{r_{11}x + r_{12}y + r_{13}z + t_x}{r_{31}x + r_{32}y + r_{33}z + t_z} \quad (1.33)$$

$$\frac{\hat{v}}{b} = \frac{r_{21}x + r_{22}y + r_{23}z + t_y}{r_{31}x + r_{32}y + r_{33}z + t_z}, \quad (1.34)$$

where the coefficients r_{ij} again denote the elements of the rotation matrix R , and the translation vector \mathbf{t} corresponds to $\mathbf{t} = (t_x, t_y, t_z)^T$. Lens distortions are modelled as described by Eqs. (1.3) and (1.4). A control point ${}^W\mathbf{x}$ is defined in the world coordinate system by ${}^W\mathbf{x} = (x, y, z)^T$.

Data acquisition for camera calibration corresponds to imaging a calibration rig of known geometry. Correspondences between the control points and their images are assumed to be given, i.e. the control points have to be identified in the images either manually or automatically. The method by Tsai (1987) attempts to obtain estimates of several camera parameters by linear least-squares optimisation methods based on the pseudo-inverse matrix approach. Constraints between parameters such as the orthonormality of the rotation matrix are not enforced in this initial stage. Similar to the DLT method, it is not the backprojection error in the image plane which is minimised initially but different error measures which are chosen such that the optimisation can be formulated as a linear optimisation problem. This approach is justified since the parameters obtained from linear optimisation are merely used as initial values for the subsequent nonlinear optimisation stage, during which all camera parameters are determined simultaneously.

Combining Eqs. (1.33) and (1.34) yields the expression

$$\frac{\hat{u}}{\hat{v}} = \frac{r_{11}x + r_{12}y + r_{13}z + t_x}{r_{21}x + r_{22}y + r_{23}z + t_y}, \quad (1.35)$$

which is independent of the principal distance b and the radial lens distortion, since it only depends on the direction from the principal point to the image point. Rearranging Eq. (1.35) yields a linear homogeneous equation in the eight unknowns r_{11} , r_{12} , r_{13} , r_{21} , r_{22} , r_{23} , t_x , and t_y according to

$$(x\hat{u})r_{11} + (y\hat{v})r_{12} + (z\hat{v})r_{13} + \hat{v}t_x - (x\hat{u})r_{21} - (y\hat{u})r_{22} - (z\hat{u})r_{23} - \hat{u}t_y = 0. \quad (1.36)$$

The coefficients in Eq. (1.36) consist of the coordinates of control points and their corresponding image points. Each such correspondence yields one equation of the form (1.36). Due to the fact that Eq. (1.36) is homogeneous, each solution for the eight unknowns remains a solution when it is multiplied by a uniform scale factor. Hence, it is useful to normalise one of the unknowns, e.g. t_y , by setting $t_y = 1$. At least seven correspondences between control points and their respective image points are then required to obtain a solution for the remaining seven unknowns. If more correspondences are available, the system of equations can be solved in the least-squares sense based on the pseudoinverse method.

A first estimate of the unknown scale factor of the solution can be obtained by exploiting the fact that the rows of the rotation matrix R are supposed to be normal. Accordingly, the scale factor is chosen such that $r_{11}^2 + r_{12}^2 + r_{13}^2 = 1$ or $r_{21}^2 + r_{22}^2 + r_{23}^2 = 1$. In addition to normalisation, orthogonality of the first two rows of R is enforced (Horn, 2000). For two given vectors \mathbf{a} and \mathbf{b} , two orthogonal vectors \mathbf{a}'

and \mathbf{b}' which are as close as possible to \mathbf{a} and \mathbf{b} are obtained by $\mathbf{a}' = \mathbf{a} + k\mathbf{b}$ and $\mathbf{b}' = \mathbf{b} + k\mathbf{a}$, resulting in a quadratic equation for the factor k according to

$$\mathbf{a}' \cdot \mathbf{b}' = \mathbf{a} \cdot \mathbf{b} + k(|\mathbf{a}|^2 + |\mathbf{b}|^2) + k^2 \mathbf{a} \mathbf{b} = 0. \quad (1.37)$$

After enforcing orthogonality of the first two rows based on the value of k inferred from Eq. (1.37), the rows have to be renormalised. The third row of R is then obtained by setting it equal to the cross product of the first two rows, finally resulting in an orthonormal rotation matrix R .

Similar to the DLT method, this first optimisation stage of the calibration method by Tsai (1987) does not minimise the backprojection error in the image plane but a different quantity defined by Eq. (1.36) which is chosen not because it is physically meaningful but because it leads to a system of homogeneous linear equations that can be solved easily. It turns out in practice, however, that the resulting camera parameters are sufficiently accurate to serve as initial values for the second, nonlinear bundle adjustment based optimisation stage.

At this point it is interesting to regard the special case of a planar calibration rig. The world coordinate system can then always be chosen such that $z = 0$ for all control points. Eq. (1.36) reduces to a linear homogeneous equation in the six unknowns $r_{11}, r_{12}, r_{21}, r_{22}, t_x$, and t_y . Again, one such equation is obtained for each correspondence between a control point and its image point. After setting e.g. $t_y = 1$, at least five correspondences are required to obtain a solution for the remaining five unknowns.

We assume that the set of linear equations yields the values $r'_{11}, r'_{12}, r'_{21}, r'_{22}, t'_x$, and $t'_y = 1$. Now the full 3×3 rotation matrix R needs to be estimated based on its upper left 2×2 submatrix. The correct scale factor is estimated by taking into account that the rotation matrix R is supposed to be orthonormal, i.e. its first rows need to be normalised with an appropriate factor k . This factor is obtained by setting

$$\begin{aligned} r'^2_{11} + r'^2_{12} + r'^2_{13} &= k^2 \\ r'^2_{21} + r'^2_{22} + r'^2_{23} &= k^2 \\ r'_{11}r'_{21} + r'_{12}r'_{22} + r'_{13}r'_{23} &= 0. \end{aligned} \quad (1.38)$$

From these equations it is straightforward to derive a biquadratic equation for k (Horn, 2000) according to

$$k^4 - k^2 (r'^2_{11} + r'^2_{12} + r'^2_{21} + r'^2_{22}) + (r'_{11}r'_{22} - r'_{12}r'_{21})^2 = 0, \quad (1.39)$$

such that the missing elements r'_{13} and r'_{23} of the first two rows of R amount to

$$\begin{aligned} r'^2_{13} &= k^2 - (r'^2_{11} + r'^2_{12}) \\ r'^2_{23} &= k^2 - (r'^2_{21} + r'^2_{22}). \end{aligned} \quad (1.40)$$

It is shown by Horn (2000) that only the more positive of the two solutions for k^2 yields positive right hand sides of Eq. (1.40). The resulting solution for k^2 is

$$k^2 = \frac{1}{2} \left((r_{11}^2 + r_{12}^2 + r_{21}^2 + r_{22}^2) + \sqrt{[(r_{11} - r_{22})^2 + (r_{12} + r_{21})^2][(r_{11} + r_{22})^2 + (r_{12} - r_{21})^2]} \right). \quad (1.41)$$

The first two rows of the rotation matrix are then normalised by dividing their elements by k . The third row is set equal to the cross product of the first two rows. The signs of r'_{13} and r'_{23} , however, are ambiguous. A straightforward way to resolve this ambiguity is to use the resulting transformation for projecting the control points into the image and to adopt the solution which yields the smallest backprojection error.

We have now estimated the rotation matrix R and the first two components t_x and t_y of the translation vector \mathbf{t} relative to its third component t_z , which has been normalised to 1. The translation component t_z and the principal distance b are obtained from Eqs. (1.33) and (1.34) which can be transformed into two linear equations for t_z and b , given the estimates for the elements of R

$$\begin{aligned} (r_{11}x + r_{12}y + r_{13}z + t_x)b - \hat{u}t_z &= (r_{31}x + r_{32}y + r_{33}z)\hat{u} \\ (r_{21}x + r_{22}y + r_{23}z + t_x)b - \hat{v}t_z &= (r_{31}x + r_{32}y + r_{33}z)\hat{v}. \end{aligned} \quad (1.42)$$

Eq. (1.42) yields a solution for t_z and b based on one or more correspondences between control points and their image points. To recover t_z and b separately, the calibration rig must span a range of depth values. If all control points have the same distance from the camera it will only be possible to determine the ratio b/t_z . Furthermore, calibration accuracy increases with the depth range covered by the control points.

Based on linear optimisation, we have now estimated the elements of the rotation matrix R , the translation vector \mathbf{t} , and the principal distance b , having enforced orthonormality of R . The principal point (u_0, v_0) and the radial and tangential distortion coefficients still need to be determined, and the parameters already obtained based on linear methods are refined by minimising the backprojection error in the image plane according to the bundle adjustment method (cf. Eq. (1.11)), utilising nonlinear optimisation techniques such as the Levenberg-Marquardt method (Press et al., 1992). The linear estimates for R , \mathbf{t} , and b are used as initial values for the nonlinear optimisation procedure. Concerning the principal point (u_0, v_0) , it is often a good initial approximation to assume that it is located in the image centre. In the case of weak radial and tangential lens distortion, it may be sufficient to set the corresponding initial parameter values to zero, while otherwise initial estimates can be obtained separately e.g. based on plumb line calibration (cf. Section 1.4.1).

1.4.4 The Camera Calibration Method by Zhang (1999a)

The camera calibration method by Zhang (1999a) is specially designed for utilising a planar calibration rig which is viewed by the camera at different viewing angles and distances. This calibration approach is derived in terms of the projective geometry framework.

For a planar calibration rig, the world coordinate system can always be chosen such that we have $Z = 0$ for all points on it. The image formation can then be described in homogeneous normalised coordinates by

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = A [R \mid \mathbf{t}] \begin{pmatrix} X \\ Y \\ 0 \\ 1 \end{pmatrix} = A [\mathbf{r}_1 \mid \mathbf{r}_2 \mid \mathbf{t}] \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix}, \quad (1.43)$$

where the vectors \mathbf{r}_i denote the columns of the rotation matrix R . Let the vector $\mathbf{M} = (X, Y)^T$ denote a point on the calibration rig with $Z = 0$. The corresponding vector in normalised homogeneous coordinates is given by $\tilde{\mathbf{M}} = (X, Y, 1)^T$. According to Eq. (1.43), in the absence of lens distortion the image point $\tilde{\mathbf{m}}$ and its corresponding scene point $\tilde{\mathbf{M}}$ are related by a homography H . A homography denotes a linear transform of a vector (of length 3) in the projective plane. It is given by a 3×3 matrix and has eight degrees of freedom as a projective transform is unique only up to a scale factor (cf. Section 1.1). This leads to

$$\tilde{\mathbf{m}} = H\tilde{\mathbf{M}} \quad \text{with} \quad H = A [\mathbf{r}_1 \ \mathbf{r}_2 \ \mathbf{t}]. \quad (1.44)$$

To compute the homography H , Zhang (1999a) proposes a nonlinear optimisation procedure which minimises the Euclidean backprojection error of the scene points projected into the image plane. The column vectors of H are denoted by \mathbf{h}_1 , \mathbf{h}_2 , and \mathbf{h}_3 . We therefore have

$$[\mathbf{h}_1 \ \mathbf{h}_2 \ \mathbf{h}_3] = \lambda A [\mathbf{r}_1 \ \mathbf{r}_2 \ \mathbf{t}], \quad (1.45)$$

where λ is an arbitrary scale factor. It follows from Eq. (1.45) that $\mathbf{r}_1 = (1/\lambda)A^{-1}\mathbf{h}_1$ and $\mathbf{r}_2 = (1/\lambda)A^{-1}\mathbf{h}_2$ with $\lambda = 1/\|A^{-1}\mathbf{h}_1\| = 1/\|A^{-1}\mathbf{h}_2\|$. The orthonormality of \mathbf{r}_1 and \mathbf{r}_2 yields $\mathbf{r}_1^T \cdot \mathbf{r}_2 = 0$ and $\mathbf{r}_1^T \cdot \mathbf{r}_1 = \mathbf{r}_2^T \cdot \mathbf{r}_2$, implying

$$\begin{aligned} \mathbf{h}_1^T A^{-T} A^{-1} \mathbf{h}_2 &= 0 \\ \mathbf{h}_1^T A^{-T} A^{-1} \mathbf{h}_1 &= \mathbf{h}_2^T A^{-T} A^{-1} \mathbf{h}_2 \end{aligned} \quad (1.46)$$

as constraints on the intrinsic camera parameters. In Eq. (1.46), the expression A^{-T} is an abbreviation for $(A^T)^{-1}$.

A closed-form solution for the extrinsic and intrinsic camera parameters is obtained by defining the symmetric matrix

$$B = A^{-T} A^{-1}, \quad (1.47)$$

which can alternatively be defined by a six-dimensional vector $\mathbf{b} = (B_{11}, B_{12}, B_{22}, B_{13}, B_{23}, B_{33})$. With the notation $\mathbf{h}_i = (h_{i1}, h_{i2}, h_{i3})^T$ for the column vectors \mathbf{h}_i of the homography H we obtain

$$\mathbf{h}_i^T B \mathbf{h}_j = \mathbf{v}_{ij} \mathbf{b}, \quad (1.48)$$

where the six-dimensional vector \mathbf{v}_{ij} corresponds to

$$\mathbf{v}_{ij} = (h_{i1}h_{j1}, h_{i1}h_{i2} + h_{j1}, h_{i2}h_{j2}, h_{i3}h_{j1} + h_{i1}h_{j3}, h_{i3}h_{j2} + h_{i2}h_{j3}, h_{i3}h_{j3})^T. \quad (1.49)$$

The two fundamental constraints given by Eq. (1.46) can now be rewritten as two homogeneous equations:

$$\begin{pmatrix} \mathbf{v}_{12}^T \\ (\mathbf{v}_{11} - \mathbf{v}_{22})^T \end{pmatrix} \mathbf{b} = 0. \quad (1.50)$$

Acquiring n images of the planar calibration rig yields n equations of the form (1.50), leading to the homogeneous linear equation

$$V\mathbf{b} = 0 \quad (1.51)$$

for \mathbf{b} , where V is a matrix of size $2n \times 6$. For $n \leq 3$ Eq. (1.51) yields a solution for \mathbf{b} which is unique up to a scale factor. Zhang (1999a) shows that for $n = 2$ images and an image sensor without skew, corresponding to the matrix element A_{12} being zero, adding the appropriate constraint $(0, 1, 0, 0, 0, 0)\mathbf{b} = 0$ yields a solution for \mathbf{b} also in this special case. If only a single calibration image is available, Zhang (1999a) proposes to assume a pixel sensor without skew ($A_{12} = 0$), set the principal point given by u_0 and v_0 equal to the image centre, and estimate only the two matrix elements A_{11} and A_{22} from the calibration image. It is well known from linear algebra that the solution to a homogeneous linear equation of the form (1.51) corresponds to the eigenvector of the 6×6 matrix $V^T V$ associated with its smallest eigenvalue (or, equivalently, the right singular value of V associated with its smallest singular value).

Once \mathbf{b} is determined, the intrinsic camera parameters can be determined based on the relation $B = vA^{-T}A$, where v is a scale factor, as follows:

$$\begin{aligned} v_0 &= A_{23} = (B_{12}B_{13} - B_{11}B_{23})/(B_{11}B_{22} - B_{12}^2) \\ v &= B_{33} - [B_{13}^2 + v_0(B_{12}B_{13} - B_{11}B_{23})]/B_{11} \\ \alpha_u &= A_{11} = \sqrt{v/B_{11}} \\ \alpha_v &= A_{22} = \sqrt{vB_{11}/(B_{11}B_{22} - B_{12}^2)} \\ \alpha_u \cot \theta &= A_{12} = -B_{12}\alpha_u^2\alpha_v/v \\ u_0 &= A_{13} = A_{12}v_0/\alpha_v - B_{13}\alpha_u^2/v \end{aligned} \quad (1.52)$$

(Zhang, 1998). The extrinsic parameters for each image are then obtained according to

$$\begin{aligned}
\mathbf{r}_1 &= \lambda A^{-1} \mathbf{h}_1 \\
\mathbf{r}_2 &= \lambda A^{-1} \mathbf{h}_2 \\
\mathbf{r}_3 &= \mathbf{r}_1 \times \mathbf{r}_2 \\
\mathbf{t} &= \lambda A^{-1} \mathbf{h}_3 .
\end{aligned} \tag{1.53}$$

The matrix R computed according to Eq. (1.53), however, does not generally satisfy the orthonormality constraints imposed on a rotation matrix. For initialisation of the subsequent nonlinear bundle adjustment procedure, Zhang (1998) therefore describes a method to estimate the rotation matrix which is closest to a given 3×3 matrix in terms of the Frobenius norm.

Similar to the DLT method, the intrinsic and extrinsic camera parameters computed so far have been obtained by minimisation of an algebraic error measure which is not physically meaningful. Zhang (1999a) uses these parameters as initial values for a bundle adjustment step which is based on the minimisation of the error term

$$\sum_{i=1}^n \sum_{j=1}^m \| \mathbf{m}_{ij} - A(R_i \mathbf{M}_j + \mathbf{t}) \|^2 . \tag{1.54}$$

In the optimisation, a rotation R is described by the Rodrigues vector \mathbf{r} . The direction of this vector indicates the rotation axis and its magnitude denotes the rotation angle in radians. The matrix R and the vector \mathbf{r} are related by the Rodrigues formula (Faugeras, 1993). Zhang (1999a) utilises the Levenberg-Marquardt algorithm (Press et al., 1992) to minimise the bundle adjustment error term (1.54).

To take into account radial lens distortion, Zhang (1999a) utilises the model defined by Eq. (1.3). Tangential lens distortion is neglected. Assuming small radial distortions, such that only the coefficients k_1 and k_3 in Eq. (1.3) are significantly different from zero, a procedure is suggested for estimating k_1 and k_3 by alternation. An initial solution for the camera parameters is obtained by setting $k_1 = k_3 = 0$, which yields projected control points according to the pinhole model. The parameters k_1 and k_3 are computed in a second step by minimising the average Euclidean distance in the image plane between the projected and the observed image points, based on an overdetermined system of linear equations. The final values for k_1 and k_3 are obtained by iteratively applying this procedure.

Due to the observed slow convergence of the iterative technique, Zhang (1999a) proposes an alternative approach to determine lens distortion by incorporating the distortion parameters appropriately into the error term (1.54) and estimating them simultaneously with the other camera parameters.

1.4.5 The Camera Calibration Method by Bouguet (2007)

Bouguet (2007) provides a toolbox for calibration of multiple cameras implemented in MATLAB. This method performs an initial estimation of the planar homography for each image which is identical to the procedure proposed by Zhang (1999a).

The initial closed-form solution for the intrinsic camera parameters is obtained in a slightly different manner by explicitly exploiting the orthogonality of vanishing points. The distortion coefficients are not estimated during the initialisation phase. In the camera model by Bouguet (2007), the intrinsic camera parameters are given by the horizontal and vertical principal distances $b_u = bk_u$ and $b_v = bk_v$, measured in pixels, where b is the principal distance measured in metric units (cf. Section 1.1), the principal point (u_0, v_0) , a skew coefficient (which is set to zero by default), and three radial and two tangential distortion parameters according to Eqs. (1.3) and (1.4) (Heikkilä and Silvén, 1997).

In this model, the plane of the image sensor is assumed to be perpendicular to the optical axis. Deviations of the true sensor orientation from this assumption are translated into non-zero tangential distortion coefficients. The final estimation of all intrinsic and extrinsic camera parameters is performed based on bundle adjustment (minimisation of the Euclidean backprojection error in the image plane) similar to Zhang (1999a), but an alternating optimisation scheme for the intrinsic and the extrinsic parameters is employed. The lens distortion parameters are initially set to zero.

1.4.6 Self-calibration of Camera Systems from Multiple Views of a Static Scene

The camera calibration approaches regarded so far (cf. Sections 1.4.2–1.4.5) all rely on a set of images of a calibration rig of known geometry with well-defined control points that can be extracted at high accuracy from the calibration images. Camera calibration without a dedicated calibration rig, thus exclusively relying on feature points extracted from a set of images of a scene of unknown geometry and the established correspondences between them, is termed self-calibration.

In principle, a bundle adjustment procedure based on multiple views of a static scene relying on a minimisation of the error term (1.11) allows to determine the unknown three-dimensional structure of the scene points simultaneously with all intrinsic and extrinsic camera parameters (up to a uniform scale factor), effectively resulting in a calibrated camera. However, good initial values should be available for the intrinsic and extrinsic camera parameters to ensure convergence of the nonlinear optimisation algorithm. In photogrammetric applications, such prior knowledge is usually available since the optical properties of the lens and the approximate three-dimensional scene structure are often fairly well known, while in the domain of computer vision such information may often be completely absent. This situation may especially occur when the source of the image data to be analysed is unknown, such as in the case of television image sequence analysis. Hence, self-calibration methods in computer vision generally consist of a projective calibration and reconstruction step based on linear equations for the intrinsic and extrinsic camera parameters (except for the lens distortion parameters, which are usually set to zero or for

which an initial guess has to be obtained independently), followed by a refinement of the inferred parameters based on bundle adjustment.

All self-calibration methods described in this section are not only suitable for determining the intrinsic and extrinsic camera parameters without involving a calibration rig with known control points but are also fundamental to understand the general scenario of three-dimensional scene reconstruction based on multiple images of the scene acquired with unknown cameras.

1.4.6.1 Projective Reconstruction: Determination of the Fundamental Matrix

According to Hartley and Zisserman (2003), the first step of self-calibration from multiple views of an unknown static scene is the determination of the fundamental matrix F between image pairs as defined in Section 1.3.2. This procedure immediately allows to compute a projective reconstruction of the scene based on the camera projection matrices P_1 and P_2 which can be computed with Eqs. (1.21) and (1.22). Given a sufficient number of point correspondences $({}^S_1 \tilde{\mathbf{x}}, {}^S_2 \tilde{\mathbf{x}})$ (seven or more), the fundamental matrix F can be computed based on Eq. (1.20). We express the image points ${}^S_1 \tilde{\mathbf{x}}$ and ${}^S_2 \tilde{\mathbf{x}}$ in normalised coordinates by the vectors $(u_1, v_1, 1)^T$ and $(u_2, v_2, 1)^T$. Each point correspondence generates one linear equation in the unknown matrix elements of F according to

$$u_1 u_2 F_{11} + u_2 v_1 F_{12} + u_2 F_{13} + u_1 v_2 F_{21} + v_1 v_2 F_{22} + v_2 F_{23} + u_1 F_{31} + v_1 F_{32} + F_{33} = 0. \quad (1.55)$$

The coefficients of this equation can be written in terms of the known (measured) coordinates of ${}^S_1 \tilde{\mathbf{x}}'$ and ${}^S_2 \tilde{\mathbf{x}}'$. We define the vector \mathbf{f} of length 9 as being composed of the matrix elements of F in row-major order. Eq. (1.55) then becomes

$$(u_1 u_2, u_2 v_1, u_2, u_1 v_2, v_1 v_2, v_2, u_1, v_1, 1) \mathbf{f} = 0. \quad (1.56)$$

Accordingly, a set of n point correspondences yields a set of linear equations for the matrix elements of F of the form

$$G\mathbf{f} = \begin{bmatrix} u_1^{(1)} u_2^{(1)} & u_2^{(1)} v_1^{(1)} & u_2^{(1)} & u_1^{(1)} v_2^{(1)} & v_1^{(1)} v_2^{(1)} & v_2^{(1)} & u_1^{(1)} & v_1^{(1)} & 1 \\ \vdots & \vdots \\ u_1^{(n)} u_2^{(n)} & u_2^{(n)} v_1^{(n)} & u_2^{(n)} & u_1^{(n)} v_2^{(n)} & v_1^{(n)} v_2^{(n)} & v_2^{(n)} & u_1^{(n)} & v_1^{(n)} & 1 \end{bmatrix} \mathbf{f} = \mathbf{0}. \quad (1.57)$$

The matrix F can only be obtained up to a scale factor because Eq. (1.57) is a homogeneous equation. A solution which is unique up to scale is directly obtained if the coefficient matrix G is of rank 8. However, if we assume that the established point correspondences are not exact due to measurement noise, the rank of the coefficient matrix G is 9 even if only eight point correspondences are taken into account, and the accuracy of the solution for F generally increases if still more point correspondences are regarded. In this case, the least-squares solution for \mathbf{f} is given by the singular vector of G corresponding to its smallest singular value. This solution

minimises the term $\|G\mathbf{f}\|$ subject to the constraint $\|\mathbf{f}\| = 1$. This method essentially corresponds to the eight-point algorithm for determination of the fundamental matrix F as introduced by Longuet-Higgins (1981).

A problem with this approach is the fact that the fundamental matrix obtained from Eq. (1.57) is generally not of rank 2 due to measurement noise, while the epipoles of the image pair are given by the left and right null-vectors of F , i.e. the eigenvectors belonging to the zero eigenvalues of F^T and F , respectively. These do not exist if the rank of F is higher than 2. A convenient way to enforce the constraint that F is of rank 2 is to replace the solution found by the singular value decomposition of the coefficient matrix G as defined in Eq. (1.57) by the matrix \bar{F} which minimises the Frobenius norm $\|F - \bar{F}\|_F$ subject to the constraint $\det \bar{F} = 0$. The Frobenius norm $\|A\|_F$ of a matrix A with elements a_{ij} is given by

$$\|A\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2 = \text{trace}(A^* A) = \sum_{i=1}^{\min(m,n)} \sigma_i^2 \quad (1.58)$$

with A^* as the conjugate transpose of A and σ_i as its singular values. If we assume that $F = UDV^T$ is the singular value decomposition of F with D as a diagonal matrix $D = \text{diag}(r, s, t)$, where $r \geq s \geq t$, the matrix \bar{F} which minimises the Frobenius norm $\|F - \bar{F}\|_F$ is given by $\bar{F} = U \text{diag}(r, s, 0) V^T$.

If only seven point correspondences are available, the coefficient matrix G in Eq. (1.57) is generally of rank 7. In this situation a solution for F can still be found by taking into account the constraint that the matrix is singular. The corresponding solution to the linear set of equations defined by $G\mathbf{f} = \mathbf{0}$ is a two-dimensional space of the form $\alpha F_1 + (1 - \alpha) F_2$ with α as a scalar variable. The matrices F_1 and F_2 as the matrices corresponding to the two right null-vectors \mathbf{f}_1 and \mathbf{f}_2 of the coefficient matrix G . We furthermore make use of the constraint $\det F = 0$, which is equivalent to $\det(\alpha F_1 + (1 - \alpha) F_2) = 0$. With the known matrices F_1 and F_2 this relation implies a cubic equation in α , which has either one or three real solutions. If complex solutions occur, they are discarded. Substituting these solutions back into the relation $\alpha F_1 + (1 - \alpha) F_2$ in turn yields one or three solutions for the fundamental matrix F (Hartley and Zisserman, 2003).

The eight-point algorithm described above is a simple method to determine the fundamental matrix F , which is implemented in a rather straightforward manner. Since the elements of the fundamental matrix may be of strongly different orders of magnitude, it is favourable to normalise by a translation and scaling transformation in each image the sensor (pixel) coordinates of the image points, given by $(u_i^{(j)}, v_i^{(j)}, 1)^T$ with $i \in \{1, 2\}$ indicating the image index and j denoting the index of the pair of corresponding points. This transformation is chosen such that the centroid of the image points, which are given in normalised homogeneous coordinates, is equal to the origin and that their root-mean-square distance to the origin amounts to $\sqrt{2}$. This transformation is performed according to

$$\begin{pmatrix} \check{u}_i^{(j)} \\ \check{v}_i^{(j)} \\ 1 \end{pmatrix} = T_i \begin{pmatrix} u_i^{(j)} \\ v_i^{(j)} \\ 1 \end{pmatrix}, \quad (1.59)$$

where the transformation matrices T_i are given by

$$T_i = \begin{bmatrix} s_i & 0 & -s_i \left\langle u_i^{(j)} \right\rangle_j \\ 0 & s_i & -s_i \left\langle v_i^{(j)} \right\rangle_j \\ 0 & 0 & 1 \end{bmatrix}$$

with $s_i = \frac{\sqrt{2}}{\sqrt{\left\langle \left(u_i^{(j)} - \left\langle u_i^{(j)} \right\rangle_j \right)^2 + \left(v_i^{(j)} - \left\langle v_i^{(j)} \right\rangle_j \right)^2 \right\rangle_j}}. \quad (1.60)$

A normalised fundamental matrix \check{F} is then obtained based on Eq. (1.57), where the image points $(u_i^{(j)}, v_i^{(j)}, 1)^T$ are replaced by the normalised image points $(\check{u}_i^{(j)}, \check{v}_i^{(j)}, 1)^T$, followed by enforcing the singularity constraint on \check{F} using the SVD-based procedure described above. Denormalisation of \check{F} according to $F = T_2^T \check{F} T_1$ then yields the fundamental matrix F for the original image points.

As a further possibility to determine the fundamental matrix, Hartley and Zisserman (2003) describe an approach which directly yields a singular matrix \tilde{F} based on an iterative linear optimisation procedure during which the epipole $\tilde{\mathbf{e}}_1$ of camera 1 is estimated.

The linear methods to determine the fundamental matrix F all rely on error measures which are purely algebraic rather than physically motivated. To obtain a fundamental matrix which is optimal in terms of the geometric distance in the image plane between the measured point correspondences ${}^{S_1}\tilde{\mathbf{x}}_i$ and ${}^{S_2}\tilde{\mathbf{x}}_i$ and the estimated point correspondences ${}^{S_1}\tilde{\mathbf{x}}_i^{(e)}$ and ${}^{S_2}\tilde{\mathbf{x}}_i^{(e)}$, which exactly satisfy the relation ${}^{S_2}\tilde{\mathbf{x}}_i^{(e)T} F {}^{S_1}\tilde{\mathbf{x}}_i^{(e)} = 0$, it is necessary to minimise the error term

$$E_G = \sum_i \left[d^2 \left({}^{S_1}\tilde{\mathbf{x}}_i, {}^{S_1}\tilde{\mathbf{x}}_i^{(e)} \right) + d^2 \left({}^{S_2}\tilde{\mathbf{x}}_i, {}^{S_2}\tilde{\mathbf{x}}_i^{(e)} \right) \right]. \quad (1.61)$$

In Eq. (1.61), the distance measure $d \left({}^{S_1}\tilde{\mathbf{x}}, {}^{S_1}\tilde{\mathbf{x}}^{(e)} \right)$ describes the Euclidean distance in the image plane between the image points ${}^{S_1}\tilde{\mathbf{x}}$ and ${}^{S_1}\tilde{\mathbf{x}}^{(e)}$, i.e. the backprojection error. To minimise the error term (1.61), Hartley and Zisserman (2003) suggest to define the camera projection matrices as $P_1 = [I \mid \mathbf{0}]$ (the so-called canonical form) and $P_2 = [M \mid \mathbf{t}]$ and the set of three-dimensional scene points that belong to the measured point correspondences ${}^{S_1}\tilde{\mathbf{x}}_i$ and ${}^{S_2}\tilde{\mathbf{x}}_i$ as ${}^W\mathbf{x}_i$. It then follows that ${}^{S_1}\tilde{\mathbf{x}}_i^{(e)} = P_1 {}^W\tilde{\mathbf{x}}_i$ and ${}^{S_2}\tilde{\mathbf{x}}_i^{(e)} = P_2 {}^W\tilde{\mathbf{x}}_i$, and the projection matrix P_2 , defined by the matrix M and the

vector \mathbf{t} , and the scene points ${}^W\tilde{\mathbf{x}}_i$ are varied such that the geometric error term E_G according to Eq. (1.61) is minimised. Due to the special form of camera matrix P_1 , the matrix F follows as $F = [\mathbf{t}] \times M$. The correspondingly estimated image points $S_1 \tilde{\mathbf{x}}_i^{(e)}$ and $S_2 \tilde{\mathbf{x}}_i^{(e)}$ exactly satisfy the relation $S_1 \tilde{\mathbf{x}}_i^{(e)T} F S_2 \tilde{\mathbf{x}}_i^{(e)} = 0$. The geometric error term (1.61) is minimised with a nonlinear optimisation algorithm such as the Levenberg-Marquardt method (Press et al., 1992). An initial estimate of the matrix elements of F may be obtained by the previously described normalised eight-point algorithm, followed by projective reconstruction of the scene points ${}^W\tilde{\mathbf{x}}_i$ as outlined at the end of Section 1.3.2. The nonlinear optimisation approach to determine the fundamental matrix F is computationally somewhat expensive. Lourakis and Argyros (2004) describe a computationally more efficient method which exploits the sparse structure of the matrix of the derivatives of the error term E_G .

1.4.6.2 The Step Towards Metric Self-calibration

In a metric coordinate system, the cameras are calibrated and the scene structure is represented in an Euclidean world coordinate system. Each of the m cameras is defined by its projection matrix $P_i^{(M)}$ which projects a point ${}^W\tilde{\mathbf{x}}_i^{(M)}$ into an image point $S_i \tilde{\mathbf{x}}_i^{(M)} = P_i^{(M)} {}^W\tilde{\mathbf{x}}_i^{(M)}$. The index M denotes that the projection matrices as well as the scene and image points, although given in homogeneous coordinates, are represented in Euclidean coordinate systems. The projection matrices can be written as $P_i^{(M)} = A_i[R_i \mid \mathbf{t}_i]$ for $i = 1, \dots, m$. A projective reconstruction (cf. Section 1.3.2) yields projection matrices P_i which are related to the corresponding Euclidean matrices $P_i^{(M)}$ by

$$P_i^{(M)} = P_i H \quad (1.62)$$

for $i = 1, \dots, m$, where H is an unknown 4×4 projective transformation. According to Hartley and Zisserman (2003), the aim of metric self-calibration is the determination of H in Eq. (1.62).

In the following we will assume that the world coordinate system is identical to the coordinate system of camera 1, i.e. $R_1 = I$ and $\mathbf{t}_1 = \mathbf{0}$. The matrices R_i and the translation vectors \mathbf{t}_i denote the transformation between camera i and camera 1, and $P_1^{(M)} = A_1[I \mid \mathbf{0}]$. In the projective reconstruction the first projection matrix is set to the canonical form $P_1 = [I \mid \mathbf{0}]$. If H is written as

$$H = \begin{bmatrix} B & \mathbf{t} \\ \mathbf{v}^T & k \end{bmatrix}, \quad (1.63)$$

Eq. (1.62) becomes $[A_1 \mid \mathbf{0}] = [I \mid \mathbf{0}]H$, which implies $B = A_1$ and $\mathbf{t} = \mathbf{0}$. Furthermore, due to the non-singularity of H the matrix element $H_{44} = k$ must be non-zero and is therefore set to 1, which fixes the scale of the reconstruction. Hence, H is of the form

$$H = \begin{bmatrix} A_1 & \mathbf{0} \\ \mathbf{v}^T & 1 \end{bmatrix}. \quad (1.64)$$

Under these conditions, the plane at infinity corresponds to

$$\tilde{\pi}_\infty = H^{-T} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{bmatrix} A_1^{-T} & -A_1^{-T}\mathbf{v} \\ \mathbf{0}^T & 1 \end{bmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -A_1^{-T}\mathbf{v} \\ 1 \end{pmatrix} \quad (1.65)$$

and hence $\mathbf{v} = -\mathbf{p}^T A$ with the upper triangular matrix $A \equiv A_1$ denoting the intrinsic calibration of the first camera (Hartley and Zisserman, 2003). Eq. (1.64) then becomes

$$H = \begin{bmatrix} A & \mathbf{0} \\ -\mathbf{p}^T A & 1 \end{bmatrix}. \quad (1.66)$$

The step from projective to metric reconstruction consists of a determination of the three components of \mathbf{p} and the five independent matrix elements of A .

The Basic Equations for Self-calibration

To determine the basic equations for self-calibration, the camera projection matrices of the projective reconstruction are denoted as $P_i = [B_i \mid \mathbf{b}_i]$. Combining Eqs. (1.62) and (1.63) yields

$$A_i R_i = (B_i - \mathbf{b}_i \mathbf{p}^T) A_1 \quad (1.67)$$

for $i = 2, \dots, m$, which corresponds to $R_i = A_i^{-1} (B_i - \mathbf{b}_i \mathbf{p}^T) A_1$. It follows from the orthonormality of the rotation matrices that $R_i R_i^T = I$ and thus

$$A_i A_i^T = (B_i - \mathbf{b}_i \mathbf{p}^T) A_1 A_1^T (B_i - \mathbf{b}_i \mathbf{p}^T)^T. \quad (1.68)$$

A geometric entity which is important in this context is the absolute conic Ω_∞ . In projective geometry, curves generated by intersecting a cone with a plane (circles, ellipses, parabolas, and hyperbolas) can all be represented by appropriate matrices C , and for all points $\tilde{\mathbf{x}}$ that are part of the conic the relation

$$\tilde{\mathbf{x}}^T C \tilde{\mathbf{x}} = 0 \quad (1.69)$$

holds. Points $\tilde{\mathbf{x}} = (X, Y, Z, W)^T$ on the absolute conic Ω_∞ are situated on the plane at infinity $\tilde{\pi}_\infty$. In a metric coordinate system we have $\tilde{\pi}_\infty = (0, 0, 0, 1)^T$, and points on Ω_∞ satisfy the two relations

$$\begin{aligned} X^2 + Y^2 + Z^2 &= 0 \\ W &= 0. \end{aligned} \quad (1.70)$$

For points on $\tilde{\pi}_\infty$ with $W = 0$, Eq. (1.70) can be written as $(X, Y, Z)I(X, Y, Z)^T = 0$. Accordingly, the matrix representation of Ω_∞ is the identity matrix I . All points on Ω_∞ are located on $\tilde{\pi}_\infty$ and are purely imaginary. According to Hartley and Zisserman (2003), the absolute conic Ω_∞ is the geometric representation of the five degrees of

freedom required to define metric properties in an affine coordinate system. The absolute conic is invariant with respect to any similarity transform.

The dual of the absolute conic Ω_∞ is the so-called absolute dual quadric denoted by Q_∞^* . Geometrically, Q_∞^* is represented by the planes which are tangent to Ω_∞ (the “envelope” of Ω_∞). Algebraically, Q_∞^* is represented by a homogeneous 4×4 matrix of rank 3, which in its canonical form in a metric coordinate system corresponds to

$$Q_\infty^* = \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix}. \quad (1.71)$$

The plane at infinity $\tilde{\pi}_\infty$ is the null-vector of Q_∞^* (Hartley and Zisserman, 2003).

An important geometric entity in the context of self-calibration is the image of the absolute conic (IAC). Its determination requires knowledge about the projection from the plane at infinity $\tilde{\pi}_\infty$ to the image plane. Points on $\tilde{\pi}_\infty$ are of the form ${}^W\tilde{\mathbf{x}}_\infty = (\mathbf{d}^T, 0)^T$. They are imaged by a general camera with a projection matrix $P = AR[I | \mathbf{a}]$ with $R\mathbf{a} = \mathbf{t}$ according to the relation

$${}^S\tilde{\mathbf{x}} = P {}^W\tilde{\mathbf{x}}_\infty = AR[I | \mathbf{a}] \begin{pmatrix} \mathbf{d} \\ 0 \end{pmatrix} = AR\mathbf{d}. \quad (1.72)$$

Hence, the transformation between $\tilde{\pi}_\infty$ and an image corresponds to a planar homography according to ${}^S\tilde{\mathbf{x}} = H\mathbf{d}$ with $H = AR$. Under a point homography H , a point ${}^W\tilde{\mathbf{x}}$ is mapped to $H {}^W\tilde{\mathbf{x}}$ while a conic C becomes $H^{-T}CH^{-1}$ (Birchfield, 1998). This implies that the conic $C = \Omega_\infty = I$ on $\tilde{\pi}_\infty$ is transformed to $\omega = (AR)^{-T}I(AR)^{-1} = A^{-T}RR^{-1}A^{-1} = (AA^T)^{-1}$. The IAC is thus given by the conic

$$\omega = (AA^T)^{-1} = A^{-T}A^{-1}. \quad (1.73)$$

Accordingly, the dual image of the absolute conic (DIAC) is given by the symmetric matrix $\omega^* = \omega^{-1} = AA^T$, corresponding to the image of the absolute dual quadric Q_∞^* . As soon as ω^* (or equivalently ω) is known, the intrinsic camera parameters given by the upper-triangular matrix A can be obtained based on Cholesky factorisation (Press et al., 1992). The DIAC $\omega^* = AA^T$ can be written in terms of the intrinsic camera parameters as

$$\omega^* = \begin{bmatrix} \alpha_u^2 + (\alpha_u \cot \theta)^2 + u_0^2 & \alpha_u \alpha_v \cot \theta + u_0 v_0 & u_0 \\ \alpha_u \alpha_v \cot \theta + u_0 v_0 & \alpha_v^2 + v_0^2 & v_0 \\ u_0 & v_0 & 1 \end{bmatrix} \quad (1.74)$$

with α_u and α_v as the scale parameters of the image sensor, u_0 and v_0 as the coordinates of the principal point in the sensor coordinate system, and θ as the skew angle of the pixel columns (cf. Section 1.1). Based on the inferred relations for the IAC ω and the DIAC ω^* , Eq. (1.68) now yields the basic equations for self-calibration:

$$\begin{aligned} \omega_i^* &= (B_i - \mathbf{b}_i \mathbf{p}^T) \omega_1^* (B_i - \mathbf{b}_i \mathbf{p}^T)^T \\ \omega_i &= (B_i - \mathbf{b}_i \mathbf{p}^T)^{-T} \omega_1 (B_i - \mathbf{b}_i \mathbf{p}^T)^{-1}. \end{aligned} \quad (1.75)$$

These equations provide relations between the unknown elements of ω_i or ω_i^* , the unknown elements of \mathbf{p} , and the known elements of the camera projection matrices given by B_i and \mathbf{b}_i . Constraints on the intrinsic camera parameters given by the matrices A_i , e.g. knowledge about their zero elements, can be used to generate equations for the eight unknown parameters of \mathbf{p} and A_1 from Eq. (1.75). Most self-calibration approaches first determine ω_i or ω_i^* and then the matrices A_i .

A special case of high practical relevance is the situation where all cameras have the same intrinsic parameters. Eq. (1.75) then becomes

$$AA^T = (B_i - \mathbf{b}\mathbf{p}^T)AA^T(B_i - \mathbf{b}\mathbf{p}^T)^T. \quad (1.76)$$

Since each side of Eq. (1.76) is a symmetric 3×3 matrix and the equation is homogeneous, each view apart from the first provides five additional constraints, such that a solution for the eight unknown parameters and thus for A can be obtained for $m \geq 3$ views.

Self-calibration Based on the Absolute Dual Quadric

For the general scenario of self-calibration, when the m cameras are not identical, Triggs (1997) introduces a method based on the absolute dual quadric Q_∞^* , which is a degenerate dual quadric represented by a 4×4 homogeneous matrix of rank 3. In this context, “degenerate” means that the quadric is “flat” in the direction defined by its null-vector, which is in turn given by the normal vector $\tilde{\pi}_\infty$ of the plane at infinity. Hence, all points on Q_∞^* are situated on the plane at infinity. The image ω^* of the absolute dual quadric Q_∞^* corresponds to a conic defined by

$$\omega^* = PQ_\infty^*P^T, \quad (1.77)$$

where P is the projection matrix of the camera. Accordingly, the projection of Q_∞^* in the image plane corresponds to the dual image of the absolute conic $\omega^* = AA^T$. Hartley and Zisserman (2003) show that Eq. (1.77) is equivalent to the basic equations of self-calibration (1.75). The basic idea of self-calibration based on Q_∞^* is to employ Eq. (1.77) to transform a constraint on ω^* to a constraint on Q_∞^* , relying on the known projection matrix P . In an Euclidean coordinate system, Q_∞^* has the canonical form $Q_\infty^* = \tilde{I} = \text{diag}(1, 1, 1, 0)$. In a general projective coordinate system we have $Q_\infty^* = H\tilde{I}H^T$, which follows from the projective transformation rule for quadrics (Hartley and Zisserman, 2003). If Q_∞^* has been estimated in a projective coordinate system, the homogeneous transformation H from the projective to the metric coordinate system can be extracted by decomposing Q_∞^* according to $Q_\infty^* = H\tilde{I}H^T$, which is computed from the eigenvalue decomposition of Q_∞^* e.g. with the Jacobi algorithm (Press et al., 1992). The metric camera projection matrix $P^{(M)}$ is then obtained by $P^{(M)} = PH$ and the metric scene point coordinates by ${}^W\tilde{\mathbf{x}}_i^{(M)} = H^{-1} {}^W\tilde{\mathbf{x}}_i$.

Linear constraints on Q_∞^* are obtained once the principal point (u_0, v_0) is known, as the image coordinate system can then be shifted such that $u_0 = v_0 = 0$. The linear equations are obtained from the resulting zero entries $\omega_{13}^* = \omega_{23}^* = \omega_{31}^* = \omega_{32}^* = 0$ in Eq. (1.74), using the projection equation (1.77) for ω^* . Additional constraints on the camera matrices A_i may yield further linear constraints on Q_∞^* , e.g. for zero skew ($\theta = 90^\circ$) it is $\omega_{12}^* = \omega_{21}^* = 0$. A further constraint is provided by a known aspect ratio of the pixels, e.g. for square pixels with $\alpha_u = \alpha_v$. The 4×4 matrix Q_∞^* is parameterised by its 10 independent matrix elements on the diagonal and above the diagonal, which may be represented by a vector of length 10. A least-squares solution for the parameters of Q_∞^* can be obtained by singular value decomposition.

In the special case of constant intrinsic camera parameters, we have $\omega_i^* = \omega_j^*$ or $P_i Q_\infty^* P_i^T = P_j Q_\infty^* P_j^T$ for all i and j . However, the equality only holds up to an unknown arbitrary projective scale. Hence, the ratios between corresponding matrix elements of ω_i^* and ω_j^* are all identical, which yields a set of quadratic equations for the matrix elements of Q_∞^* . Under the weaker assumption of zero pixel skew for all images, Eq. (1.74) yields the constraint $\omega_{12}^* \omega_{33}^* = \omega_{13}^* \omega_{23}^*$, corresponding to one quadratic equation in the matrix elements of Q_∞^* provided by each image. A further constraint is given by $\det Q_\infty^* = 0$, which follows from the fact that the absolute dual quadric is degenerate. As the projective scale factor of Q_∞^* is arbitrary, at least eight images are necessary to compute Q_∞^* . A different nonlinear approach to determining Q_∞^* is to minimise the error term $\sum_{i=1}^m \|A_i A_i^T - P_i Q_\infty^* P_i^T\|_F^2$, where $\|\dots\|_F$ denotes the Frobenius norm, and the matrices $A_i A_i^T$ and $P_i Q_\infty^* P_i^T$ are normalised to unit Frobenius norm, respectively. This error term is minimised e.g. with the Levenberg-Marquardt algorithm. Since this is merely an algebraic rather than a physically meaningful error term, it is advantageous to refine the solution found for the matrix elements of Q_∞^* and ω_i^* by a full bundle adjustment step.

Self-calibration Based on the Kruppa Equations

Based on the method originally suggested by Kruppa (1913) to determine the intrinsic parameters of a camera, for a pair of images the relation

$$[\tilde{\mathbf{e}}_2]_\times \omega^* [\tilde{\mathbf{e}}_2]_\times = F \omega^* F^T \quad (1.78)$$

between the epipole $\tilde{\mathbf{e}}_2$, the DIAC ω^* , and the fundamental matrix F is established by Viéville and Lingrand (1995). In Eq. (1.78), $[\tilde{\mathbf{e}}_2]_\times$ denotes the cross product matrix defined according to Eq. (1.17) of the epipole $\tilde{\mathbf{e}}_2$. Eq. (1.78) is termed Kruppa equation; due to its algebraic form, however, this formulation cannot be easily applied in practice. Hartley (1997) provides a different derivation of the Kruppa equation and arrives at the formulation

$$\frac{\mathbf{u}_2^T \omega_2^* \mathbf{u}_2}{\sigma_1^2 \mathbf{v}_1^T \omega_1^* \mathbf{v}_1} = - \frac{\mathbf{u}_2^T \omega_2^* \mathbf{u}_2}{\sigma_1 \sigma_2 \mathbf{v}_1^T \omega_1^* \mathbf{v}_2} = \frac{\mathbf{u}_1^T \omega_2^* \mathbf{u}_1}{\sigma_2^2 \mathbf{v}_2^T \omega_1^* \mathbf{v}_2}, \quad (1.79)$$

where \mathbf{u}_i and \mathbf{v}_i are the column vectors of the matrices U und V obtained by singular value decomposition of the fundamental matrix F according to $F = UDV^T$, with D as a diagonal matrix, and σ_i are the singular values. In Eq. (1.79) the DIACs ω_1^* and ω_2^* are not assumed to be identical in the two images. Cross-multiplication of both sides of the equation yields two quadratic equations in the parameters of ω_1^* and ω_2^* . If the DIACs are assumed to be identical, six quadratic equations in the five unknown parameters of ω^* are given by three images.

The main advantage of using the Kruppa equations to determine the DIAC ω^* is that this approach does not require a projective reconstruction as it is the case for methods that rely on a determination of the absolute dual quadric Q_∞^* . Furthermore, for two views the Kruppa equation provides the only constraint available for ω^* (Lu et al., 2004). However, due to the fact that the Kruppa equation does not enforce the degeneracy of the absolute dual quadric and the constraint that the absolute conic lies in the same plane (the plane at infinity) for all images, the accuracy and performance of self-calibration based on the Kruppa equations is often observed to be inferior to methods based on the determination of the absolute dual quadric Q_∞^* .

Stratified Self-calibration

The calibration approach based on the absolute dual quadric simultaneously estimates both the intrinsic camera parameters given by the DIAC ω^* and the plane at infinity $\tilde{\pi}_\infty$ given as the null-space of Q_∞^* . The resulting necessity to determine many parameters in a single optimisation run may lead to numerical instabilities. Hence, an alternative approach is to proceed step by step, from projective to affine and finally to metric reconstruction.

Affine reconstruction is possible as soon as the plane at infinity $\tilde{\pi}_\infty$ is known, since $\tilde{\pi}_\infty$ is invariant with respect to affine transformations. The plane at infinity can be computed based on prior knowledge about the scene structure, such as parallel lines and the resulting vanishing points. Without such prior knowledge, determining $\tilde{\pi}_\infty$ is possible by exploiting the modulus constraint (Pollefeys and van Gool, 1999), which is a constraint on the position of $\tilde{\pi}_\infty$ in the form of a polynomial equation. If we assume constant intrinsic camera parameters, Eq. (1.67) with $A_i = A$ yields

$$B_i - \mathbf{b}_i \mathbf{p}^T = \mu A R_i A^{-1}, \quad (1.80)$$

where μ is the arbitrary projective scale factor and the vector \mathbf{p} represents the coordinates of the plane at infinity given by the first three elements of $\tilde{\pi}_\infty$, i.e. $\tilde{\pi}_\infty = (\mathbf{p}^T, 1)$. The matrix $A R_i A^{-1}$ is conjugate to a rotation and therefore has the eigenvalues $\{1, e^{i\phi}, e^{-i\phi}\}$. As a consequence, the eigenvalues of $(B_i - \mathbf{b}\mathbf{p}^T)$ are $\{\mu, \mu e^{i\phi}, \mu e^{-i\phi}\}$ and thus have equal absolute values (moduli). This result corresponds to the modulus constraint on the coordinates \mathbf{p} of the plane at infinity $\tilde{\pi}_\infty$. To compute the eigenvalues of the matrix $B_i - \mathbf{b}\mathbf{p}^T$ one needs to compute the roots of its characteristic polynomial $\det(\lambda I - B_i + \mathbf{b}\mathbf{p}^T)$, which leads to a fourth-degree polynomial in the three elements of \mathbf{p} (Hartley and Zisserman, 2003). In principle, from three

images a number of 4^3 possible solutions can be inferred for \mathbf{p} , but an additional cubic equation for the elements of \mathbf{p} can be established which often allows to eliminate most spurious solutions. While the Kruppa equation represents constraints on ω^* but does not involve $\tilde{\pi}_\infty$, the modulus constraint determines $\tilde{\pi}_\infty$ but does not explicitly take into account ω^* .

Once the plane at infinity $\tilde{\pi}_\infty$ is determined, affine reconstruction has been performed. The step from affine to metric reconstruction corresponds to determining the intrinsic camera parameters given by the matrix A based on $\tilde{\pi}_\infty$. Transformation of the IAC or the DIAC yields a linear algorithm for A . At this point the concept of the homography H induced by a plane $\tilde{\pi}$ turns out to be helpful. The ray corresponding to a point ${}^{S_1}\tilde{\mathbf{x}}$ in image 1 meets the plane $\tilde{\pi}$ in a point ${}^W\tilde{\mathbf{x}}_\pi$, which is then projected into the point ${}^{S_2}\tilde{\mathbf{x}}$ in image 2. The transformation between the two image points is a homography H , i.e. ${}^{S_2}\tilde{\mathbf{x}} = H {}^{S_1}\tilde{\mathbf{x}}$. Given the projection matrices $P_1 = [I \mid \mathbf{0}]$ and $P_2 = [B \mid \mathbf{b}]$ for the two cameras and the plane $\tilde{\pi} = (\mathbf{v}^T, 1)^T$, where $\tilde{\pi}^T {}^W\mathbf{x} = 0$ for all scene points ${}^W\mathbf{x}$ on $\tilde{\pi}$, the homography H induced by the plane $\tilde{\pi}$ is ${}^{S_2}\tilde{\mathbf{x}} = H {}^{S_1}\tilde{\mathbf{x}}$ with $H = B - \mathbf{b}\mathbf{v}^T$ (Hartley and Zisserman, 2003). Note that the vector \mathbf{v} is not a projective vector.

The infinite homography H^∞ is the plane projective transformation between two images induced by the plane at infinity $\tilde{\pi}_\infty$. Once the plane at infinity $\tilde{\pi}_\infty = (\mathbf{p}^T, 1)^T$ and the camera projection matrices $P_i = [B_i \mid \mathbf{b}_i]$ are known for the projective coordinate system of each view, the infinite homography corresponds to

$$H_i^\infty = B_i - \mathbf{b}_i \mathbf{p}^T, \quad (1.81)$$

where H_i^∞ is the plane homography from the image of a camera with projection matrix $[I \mid \mathbf{0}]$ to the image of camera i with projection matrix $[B_i \mid \mathbf{b}_i]$. In Eq. (1.81), the first camera is assumed to be in its canonical form $[I \mid \mathbf{0}]$. Otherwise, Eq. (1.81) becomes

$$H_i^\infty = (B_i - \mathbf{b}_i \mathbf{p}^T) (B_1 - \mathbf{b}_1 \mathbf{p}^T)^{-1}. \quad (1.82)$$

The image of the absolute conic, which lies on $\tilde{\pi}_\infty$, is mapped between camera i and camera 1 by H_i^∞ . The IAC ω_i and the DIAC ω_i^* for camera i now become

$$\begin{aligned} \omega_i &= (H_i^\infty)^{-T} \omega_1 (H_i^\infty)^{-1} \\ \omega_i^* &= H_i^\infty \omega_1^* (H_i^\infty)^T. \end{aligned} \quad (1.83)$$

(Lu et al., 2004). Eq. (1.83) is equivalent to the basic equations of self-calibration according to Eq. (1.75). This is an especially useful formulation since it yields linear constraints on ω_i . Constraints imposed on ω_i in one camera can be directly transferred to another camera, which may lead to a number of constraints that is sufficient to determine ω_1 by means of linear optimisation methods alone. The matrix A_1 of the intrinsic camera parameters can be obtained from ω_1 based on Cholesky decomposition (Press et al., 1992).

An important special case is the configuration with constant intrinsic parameters, i.e. $A_i = A$ and $\omega_i^* = \omega^*$ for all cameras. Eq. (1.83) then becomes $\omega^* = H_i^\infty \omega^* (H_i^\infty)^T$. Although this equation is homogeneous, the projective scale factor

Essential matrix: $I_1 \tilde{\mathbf{x}}'^T E^T I_2 \tilde{\mathbf{x}}' = 0$	Fundamental matrix: $S_2 \tilde{\mathbf{x}}^T F^T S_1 \tilde{\mathbf{x}} = 0$ $F = A_2^{-T} E A_1^{-1}$
Epipoles: $\tilde{\mathbf{e}}_2^T F = 0 \quad F \tilde{\mathbf{e}}_1 = 0$	Projective reconstruction (linear method): $P_1 = [I \mid \vec{0}] \quad P_2 = [[\tilde{\mathbf{e}}_2] \times F \mid \tilde{\mathbf{e}}_2]$
Projection matrices in Euclidean coordinate system: $P_i^{(M)} = P_i H$ $H = \begin{bmatrix} A_1 & \mathbf{0} \\ \mathbf{v}^T & 1 \end{bmatrix}$	$G^T W \tilde{\mathbf{x}} = 0$ $G = \begin{bmatrix} u_1 \tilde{\mathbf{p}}_1^{(3)T} - \tilde{\mathbf{p}}_1^{(1)T} \\ v_1 \tilde{\mathbf{p}}_1^{(3)T} - \tilde{\mathbf{p}}_1^{(2)T} \\ u_2 \tilde{\mathbf{p}}_2^{(3)T} - \tilde{\mathbf{p}}_2^{(1)T} \\ v_2 \tilde{\mathbf{p}}_2^{(3)T} - \tilde{\mathbf{p}}_2^{(2)T} \end{bmatrix}$ Nonlinear projective reconstruction is based on minimising Euclidean distances in the image plane.
Image of the absolute conic (IAC): $\omega = (AA^T)^{-1} = A^{-T}A^{-1}$	Dual image of the absolute conic (DIAC): $\omega^* = \omega^{-1} = AA^T$
Basic equations of self-calibration: $\omega_i^* = (B_i - \mathbf{b}_i \mathbf{p}^T) \omega_1^* (B_i - \mathbf{b}_i \mathbf{p}^T)^T$ $\omega_i = (B_i - \mathbf{b}_i \mathbf{p}^T)^{-T} \omega_1 (B_i - \mathbf{b}_i \mathbf{p}^T)^{-1}$	where $P_i = [B_i \mid \vec{b}_i]$
Self-calibration based on the absolute dual quadric: $\omega^* = PQ_\infty^*P^T \quad Q_\infty^* = H \tilde{H}^T$ $P^{(M)} = PH \quad W \tilde{\mathbf{x}}_i^{(M)} = H^{-1} W \tilde{\mathbf{x}}_i$	determination of H based on eigenvalue decomposition of the absolute dual quadric transformation into Euclidean coordinates
Self-calibration based on the Kruppa equations: $[\tilde{\mathbf{e}}_2] \times \omega^* [\tilde{\mathbf{e}}_2] \times = F \omega^* F^T$ $\frac{\mathbf{u}_2^T \omega_2^* \mathbf{u}_2}{\sigma_1^2 \mathbf{v}_1^T \omega_1^* \mathbf{v}_1} = - \frac{\mathbf{u}_2^T \omega_2^* \mathbf{u}_2}{\sigma_1 \sigma_2 \mathbf{v}_1^T \omega_1^* \mathbf{v}_2}$ $= \frac{\mathbf{u}_1^T \omega_2^* \mathbf{u}_1}{\sigma_2^2 \mathbf{v}_2^T \omega_1^* \mathbf{v}_2}$	Stratified self-calibration: Step 1: Affine reconstruction by determination of the plane at infinity, relying on the basic equations of self-calibration and exploiting the modulus constraint Step 2: Infinite homography, determination of the IAC and DIAC $H_i^\infty = (B_i - \mathbf{b}_i \mathbf{p}^T) (B_i - \mathbf{b}_i \mathbf{p}^T)^{-1}$ $\omega_i = (H_i^\infty)^{-T} \omega_1 (H_i^\infty)^{-1}$ $\omega_i^* = H_i^\infty \omega_1^* (H_i^\infty)^T$.

Fig. 1.4 Summary of projective techniques for scene reconstruction and self-calibration.

amounts to unity as long as all matrices H_i^∞ are normalised such that $\det H_i^\infty = 1$. This relation yields six linear equations in the six independent elements of ω^* . Hartley and Zisserman (2003) state that the resulting system of linear equations is conditioned such that a unique solution for ω^* can only be obtained if more than two images are available.

A summary of the scene reconstruction and self-calibration methods based on projective geometry described so far in this section is given in Fig. 1.4.

Self-calibration Based on Scene Constraints (Vanishing Points)

The self-calibration methods described so far require a large number of point correspondences between the acquired images. In certain scenes, especially those containing human-made objects such as buildings, it may be more convenient to extract lines instead of well-defined points from the acquired images. Hence, to conclude this section, we will give a short outline of how scene constraints such as pairs of lines in the images that correspond to parallel lines in the scene can be exploited for the purpose of self-calibration.

Two parallel lines in the scene intersect at infinity. The projected images of the lines are generally not parallel and intersect in an image point ${}^S\tilde{\mathbf{v}}$ with well-defined image coordinates, which is termed vanishing point. Cipolla et al. (1999) present a framework to directly estimate the elements of the matrix A of the intrinsic camera parameters based on three mutually orthogonal vanishing points. They suggest that parallel lines are extracted manually from the image. Similarly, vanishing points fulfilling the orthogonality condition can be used for computing the IAC ω . A pair of vanishing points ${}^S\tilde{\mathbf{v}}_1$ and ${}^S\tilde{\mathbf{v}}_2$ corresponding to orthogonal directions in the scene are shown by Hartley and Zisserman (2003) to represent conjugate points with respect to ω , thus fulfilling the relation

$${}^S\tilde{\mathbf{v}}_1^T \omega {}^S\tilde{\mathbf{v}}_2 = 0. \quad (1.84)$$

The five independent parameters defining the matrix ω can be obtained based on the linear constraints imposed by three mutually orthogonal vanishing points. If correspondences are established between at least two vanishing points detectable in two images, respectively, it is possible to infer the rotation between the views. The translation cannot be recovered due to the fact that vanishing points are situated at infinite distance.

As it is fairly unlikely that three mutually orthogonal vanishing points can be reliably detected in a single image, Grammatikopoulos et al. (2004) extend the approach based on vanishing points towards a framework involving several images acquired independently with the same camera (thus not necessarily all showing the same object), each displaying two orthogonal vanishing points. The lines extracted from the images to determine the vanishing points are also used to infer the radial lens distortion coefficients. An automated version of this technique is described by Grammatikopoulos et al. (2006). Lines are extracted and vanishing points are determined automatically based on the method by Rother (2000). Multiple images acquired independently with the same camera, each displaying either two or three mutually orthogonal vanishing points, are utilised for this method.

1.4.7 Semi-automatic Calibration of Multiocular Camera Systems

In principle, all intrinsic and extrinsic camera parameters can be determined by self-calibration as a “by-product” of the three-dimensional reconstruction of the scene based on a large number of point correspondences between the images. An alternative approach to the determination of the intrinsic camera parameters exploits scene constraints such as pairs of parallel lines in the scene. However, not every distribution of scene points is equally well suited for an accurate determination of camera parameters based on self-calibration—as an example, many linear methods fail if the scene points do not occupy a volume in space. As a consequence, during the camera calibration phase a “cooperative” scene from which scene points and their mutual correspondences can be extracted at high accuracy and reliability is highly desirable. Later on, once the intrinsic camera parameters (and the relative extrinsic camera transformations in the case of multiple cameras) are known, the scenes to be reconstructed may become less cooperative.

Especially in self-calibration scenarios, false correspondences are a serious problem as when they remain undetected, the final reconstruction result may display gross errors. Outlier rejection methods such as RANSAC (Fischler and Bolles, 1983) or robust optimisation techniques like M-estimators (Rey, 1983) are often unable to fully solve this problem. In some cases, the scene may even contain no well-defined point features at all, e.g. if it is desired to reconstruct a weakly-textured or textureless surface. Furthermore, the majority of automatic methods for establishing point correspondences assume that the surfaces of the objects in the scene are Lambertian, while most realistic surface materials also have a specular component. This specular component may lead to inaccurate image point measurements—if e.g. a specular reflection is detected as a point feature in one image, it will be associated with a different physical surface point in an image acquired from another viewpoint, leading to inaccuracies or even gross errors in the measured image coordinates of the corresponding points. Under unfavourable circumstances, the determination of point correspondences may become impossible at all since the specular object surface may look completely different from different viewpoints (the problem of stereo image analysis in the presence of specularly reflecting surfaces and the physically correct determination of point correspondences is regarded in detail in Section 4.4).

Hence, it is observed in many real-world scenarios that more accurate intrinsic and extrinsic camera parameters (and in turn three-dimensional scene reconstruction results) can be obtained by a camera calibration procedure which is based on a calibration rig displaying control points which are easy to detect in the image and which have accurately known world coordinates. An aspect not regarded in the previous sections about camera calibration methods is the automatic extraction of the control points of the calibration rig and their mutual correspondences from the acquired set of calibration images. This problem is examined in some detail in this section.

In the domain of classical photogrammetry, camera calibration methods based on accurately manufactured non-planar calibration rigs are used which determine the intrinsic and extrinsic camera parameters based on bundle adjustment (Luhmann, 2003). In the domain of three-dimensional computer vision, the methods by Tsai

(1987) and by Zhang (1999a) (cf. Sections 1.4.3 and 1.4.4) are of high practical importance. The main reasons are that they rely on planar calibration rigs which are generally more easily manufactured than three-dimensional rigs and that they do not require any prior knowledge about the intrinsic or extrinsic camera parameters due to the initialisation steps, which are based on linear optimisation and therefore do not tend to get stuck in local optima.

Planar calibration rigs can be produced fairly easily, e.g. by printing the calibration pattern on paper or by producing a composite structure with a photolithographically printed pattern. For non-planar calibration rigs, usually rigidly connected planar components are used, e.g. oriented by a right angle with respect to each other or composed of a layered structure. While planar rigs are easier to manufacture, non-planar rigs tend to result in a higher accuracy of the determined parameters (Krüger, 2007).

An important type of calibration pattern consists of a white background covered with black circular dots placed on a quadratic grid. The control points correspond to the dot centres relative to one specific dot, e.g. the top left one. White retro-reflective dots on black background may also be used. The major advantage of a circular dot pattern is the fact that the centres of the dots can be extracted at high subpixel accuracy if the dots appear sufficiently large in the images (Luhmann, 2003). A correction term for the error of the dot centre coordinates induced by lens distortion, which is typically of the order of a few 10^{-2} pixels, is introduced by Heikkilä and Silvén (1997). A drawback of the circular dot pattern is the fact that the mutual orientations of multiple cameras cannot be recovered without additional markers on the rig. A solution to this problem is provided by the commercial system of the company Aicon, where binary, rotationally symmetric ring patterns are added to each circular dot. The binary code assigns a unique number to each point, such that a robust identification of all control points is possible even when the rig is not completely visible in the image. A drawback of this solution, however, is the fact that the calibration system only works in combination with a custom-made, highly accurate but also fairly expensive carbon fibre calibration rig and that it is required to use a commercial software which cannot always be adapted to the application-specific needs.

A problem which is common for the described systems is that an automatic extraction of the control points on the calibration rig from the calibration images is desirable, but a robust method to perform this task is not always accessible. Commercial systems, e.g. those developed by the companies Aicon or PointGrey Research, and open source systems such as the Open Computer Vision library (OpenCV) or the Matlab calibration toolbox by Bouguet (2007) are available for computing the intrinsic and extrinsic camera parameters, having different features and being of a different degree of maturity. The system by Bouguet (2007) has become very popular. This software is comparable in accuracy to other systems, but superior to all other systems in practicability as the calibration rig can be printed out with a standard laser printer. Even though it is able to calibrate binocular cameras, the system is not useable for more than two cameras or if the cameras are arbitrarily positioned in space rather than arranged in standard epipolar geometry. Additionally, the rig

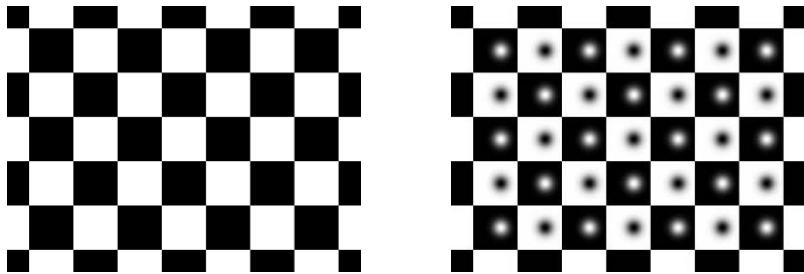


Fig. 1.5 Left: The calibration rig proposed by Bouguet (2007). It provides no information about the rig orientation if the black and white fields are of square shape and ambiguous information otherwise. Right: The calibration rig proposed by Krüger et al. (2004). It provides unambiguous orientation information regardless of the number and size of the squares.

detector in the toolbox is a manual solution, i.e. the user has to mark the rig corners by clicking into the images.

For these reasons we propose a semi-automatic camera calibration approach (Krüger et al., 2004), which is employed in the application scenarios described in Chapters 5 and 6. While in contrast to self-calibration techniques relying on point features extracted from a scene of unknown geometry, our method requires a calibration rig with accurately known control points, no user interaction is necessary since the control points are automatically extracted from the calibration images. After automatic extraction of the control points, the calibration technique by Bouguet (2007) is used, which in turn relies on the lens distortion model by Brown (1958) and on the methods introduced by Zhang (1999a) and by Heikkilä and Silvén (1997) for determination of the intrinsic and extrinsic camera parameters.

1.4.7.1 The Rig Extraction Algorithm

The calibration rig as proposed by Bouguet (2007) consists of a chequerboard pattern, e.g. generated by a laser printer, attached to a planar surface. The parts of this rig will be named as follows: The term “square” denotes the black or white fields of the rig, which do not need to be square, though. A “corner” is a point where four squares touch, while “rig corner” denotes the outermost four corners where one large and three small squares touch. This calibration rig is very simple to generate but does not provide an orientation. With a slight modification consisting of acentrical marks without sharp edges, an orientation can be obtained even if only a part of the rig visible. Fig. 1.5 illustrates the original calibration rig proposed by Bouguet (2007) and the calibration rig proposed in this work.

Existing Algorithms for Extracting the Calibration Rig

To extract the image coordinates of the control points from the images, the software package by Bouguet (2007) provides a manual solution which prompts the user to click near the outer corners of the rig. While being completely out of scope for an automatic calibration process, it makes even the occasional calibration process very cumbersome and prone to errors. A good calibration needs about 20 images per camera. At four clicks per image (if no mistake is made), this amounts to 240 clicks with a precision of less than four pixels for a trinocular camera system.

The OpenCV library provides a corner detection algorithm as well. It is based on an initial scanning process with a subsequent subpixel accurate search of the corners. The initial scanning process operates on binary images, extracts the potential corner candidates and tries to sort them with a polygonal approximation in a loop. If all expected corners are determined, the result is assigned to the subpixel accurate location. It is a gradient minimum search with a relocation of a neighborhood window until the center stays within a given threshold. In practice, this contour analysis turns out to be rather instable. The OpenCV algorithm is not able to properly detect and sort the corners under lighting conditions occurring under outdoor conditions or in the industrial production environment. In order to avoid singularities due to parallel image and calibration rig planes, the calibration rig must be imaged at a certain obliquity. The corner analysis of the OpenCV implementation often fails if the angle or the camera-to-rig distance is too large, mainly due to an imprecise polygon approximation.

Two proprietary implementations based on cross-correlation template matching, line fitting and subsequent precise position estimation at subpixel accuracy, have been conceived as the predecessors of the algorithm proposed in this work. The first one operates on the assumption of a fully visible calibration rig. Hence, it extracts a number of most prominent features, equal to the number of corners in the calibration rig, by means of cross-correlation matching (Aschwanden, 1993). Subsequently, an outlier detection is performed based on a Hough transform (Jähne, 2005) and geometric constraint evaluation. Upon this, the lines are approximated by least-squares methods. Independent of the previously detected features the line intersections are used as coarse corner guesses, and a maximum search followed by fitting a paraboloid to the correlation coefficients is performed. The second method discards the assumption of a fully visible calibration rig. Instead of extracting a fixed number it selects features according to their reliability. The coarse feature sorting and outlier detection is performed as described above. After that, it is determined for each potential rig corner if it displays a rig-corner template. The final sorting before the determination of feature position at subpixel accuracy is performed by accumulating features along line segments into approximated lines.

These two state-of-the-art algorithms are very robust with respect to occlusion, illumination, and noise. Due to the efficient implementation of the correlation process they are fast enough for real-time processing. The limit of usability is the assumption that the straight lines through the corners appear approximately straight in the image. Once the lens distortion effects are comparable to the corner spacing

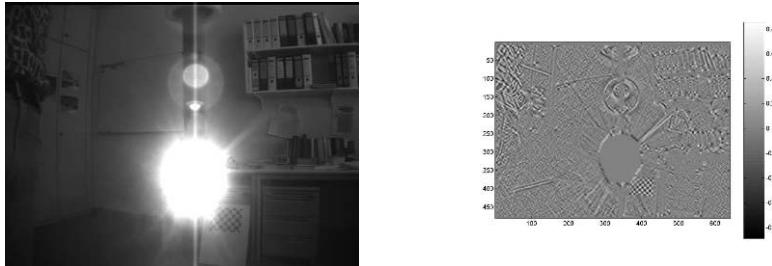


Fig. 1.6 Left: Image of the calibration rig under unfavourable illumination conditions which may, however, occur in real-world environments such as industrial production facilities. Right: Resulting cross-correlation coefficient.

in the image, this assumption is not valid any more since it is only true for lenses with weak distortion. If it is required to cope with wide angle lenses, lens distortion increases. In the image, lens distortion makes straight lines appear bent. This effect directly leads to blurring in the Hough accumulator of the first sorting stage when compared to a lens with weak distortion. In the blurred Hough accumulator, maxima cannot be detected precisely, such that false positive lines appear or existing lines remain undetected. Obviously, this behaviour is directly related to the fact that the Hough transform is a global algorithm. One way to cope with the distortion is to change from a global approach to a local approach.

A Graph-based Rig Extraction Algorithm

The rig extraction algorithm by Krüger et al. (2004) described in this section aims for establishing correspondences between the control points on the calibration rig and the respective positions of the corner points in a large number of calibration images in a reliable manner under possibly difficult image acquisition conditions. The proposed algorithm is based on the previously mentioned features, i.e. the cross-correlation coefficient between the image and a corner mask. The local extrema are identified, and their position is determined at subpixel accuracy by means of weighted mean or bivariate quadratic interpolation. The major difference to the existing approaches is the integration of these local extrema to a complete rig. Both the Hough transform and the proposed algorithm are bottom-up algorithms: Starting with atomic features, more complex entities are constructed, cumulating in the complete rig. The integration in this algorithm is done by topological methods. This approach is guided by a general principle: Prefer discarding the whole image to accepting a false positive identification of the corners. This strategy results from the fact that images are easy to acquire while errors are hard to cope with during subsequent processing stages. One observes that positive and negative correlation coefficient peaks interchange. Each positive peak has four negative neighbours directly connected along the black/white edges of the squares and vice versa. This effect is shown in Fig. 1.6. The first step is to identify this neighbourhood relation,

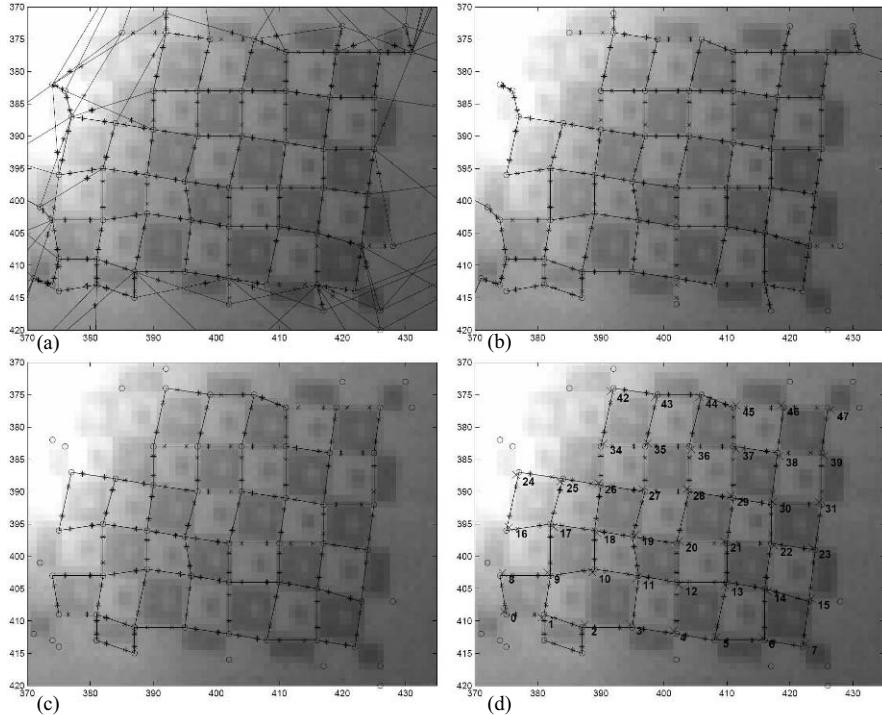


Fig. 1.7 Results of the first two topological filters and the corner enumeration. The images display an enlargement of the relevant part of Fig. 1.6, demonstrating the performance of the processing steps under difficult conditions. The asterisks denote the direction of the edges. (a) Initial graph. (b) Graph without non-bidirectional edges. (c) Graph containing circles only. (d) Final corner enumeration.

which is illustrated in Fig. 1.7a for a close-up of the image shown in Fig. 1.6. The quite extreme situation shown in Fig. 1.6 was chosen to illustrate the effect that under realistic conditions e.g. in the industrial production environment (cf. Chapter 6) highlights may appear in the calibration images. It is thus required that the rig extraction algorithm is insensitive with respect to such perturbations.

The next steps verify the resulting directed graph according to a few simple rules that eliminate the edges to false positive corners and cut the graph such that the rig becomes a single graph component. Each corner candidate contains the edges to the four neighbours, labelled with the respective directions (left, right, up, down). The processing steps operate on a directed graph G defined as

$$G = \{V, E\} \quad (1.85)$$

$$V = \left\{ s_{\mathbf{x}_i} = \begin{pmatrix} u_i \\ v_i \end{pmatrix} \mid i = 1, \dots, N_v \right\} \quad (1.86)$$

$$E = \left\{ \mathbf{e}_i = (s_i, t_i, d_i) \mid i = 1, \dots, N_e; \right.$$

Table 1.1 Definitions of the graph functions used by the edge circle filter.

d	$cw(d)$	$ccw(d)$	$opposite(d)$
left	up	down	right
up	right	left	down
right	down	up	left
down	left	right	up

$$s_i, t_i \in 1, \dots, N_v; \\ d_i \in D = \{\text{left,right,up,down}\} \quad (1.87)$$

The corner candidates in the correlation coefficient image are obtained by a straightforward non-maximum suppression followed by false-positive removal. The non-maximum suppression is performed by counting the number of pixels with a lower absolute value than the centre pixel in the eight-neighbourhood. If the count exceeds a threshold (default: 6) and the centre value exceeds another threshold (default: 0.75), the pixel is assumed to be a corner candidate. This simple non-maximum suppression provides a robust detection with a reasonable amount of false positives. The false positives are the neighbouring pixels of the true positive. Deciding which of the pixels is the true positive is done during the position estimation at subpixel accuracy. As soon as this position is available, the same data are used to determine the interpolated cross-correlation value. The candidate with the larger cross-correlation value is assumed to be the true positive. Two algorithms for computing the subpixel position are investigated: Weighted mean (WM) and bivariate quadratic interpolation (BVI). The subpixel position of WM is the average of the eight-neighbourhood positions weighted by the corresponding correlation coefficients. The interpolated value is the mean of the correlation coefficients involved. The subpixel position of BVI is the location of the extremum of the bivariate quadratic function, assuming it is appropriately shaped and does not form a saddle point. The interpolated cross-correlation value is the function value at the extremum. This procedure yields the set V in Eq. (1.85). The corresponding set E is constructed by finding the nearest neighbour t of each vertex s in the respective direction d . The correlation coefficients of v_s and v_t must be of opposite sign.

The first filter eliminates non-bidirectional graph edges. Fig. 1.7b illustrates a situation where a false positive is eliminated this way. This procedure corresponds to deleting all graph edges $e = (s, t, d)$ subject to the condition $(t, s, opposite(d)) \notin E$. The functions $cw(d)$ (clockwise), $ccw(d)$ (counter-clockwise), and $opposite(d)$ operating on the graph are defined according to Table 1.1.

The second filter checks for circles (closed, non-reversing paths in E) of length 4. These circles in the graph map directly to the edges of the squares. Fig. 1.7 illustrates incomplete circles (e.g. Fig. 1.7b, top row, leftmost corner) and complete circles only (Fig. 1.7c). The filter is implemented as a so-called mark-and-sweep algorithm. In the first run, all corner candidates which are part of at least one circle of length 4 are marked. In the second run, all candidates which have not been marked are deleted.

The circle check is performed by marking a vertex that fulfills the following three conditions:

$$e_1 = (v_0, v_1, d_1) \in E \quad \text{for all } d_1 \in D \quad (1.88)$$

$$e_i = (v_{i-1}, v_i, d_i = \text{cw}(d_{i-1})) \in E \quad \text{for } i \in \{2, 3, 4\} \quad (1.89)$$

$$v_0 = v_4. \quad (1.90)$$

An analogous check operates in counterclockwise direction, i.e. the function $\text{cw}(\dots)$ in Eq. (1.89) is replaced by the function $\text{ccw}(\dots)$.

The third filter eliminates graph edges that have an exceptional difference in length. The lengths are compared along one axis only, i.e. left/right and up/down, which avoids problems with strongly tilted rigs and false positives next to the rig (not shown here). For each vertex with edges oriented in opposite directions, i.e. $(s, t_1, d) \in E$ and $(s, t_2, \text{opposite}(d)) \in E$, the lengths $l_1 = \|v_s - v_{t_1}\|$ and $l_2 = \|v_s - v_{t_2}\|$ are determined. If the condition

$$1 - \frac{\min(l_1, l_2)}{\max(l_1, l_2)} > \delta \quad (1.91)$$

is fulfilled, the lengths of the edges are assumed to be so strongly different that this difference cannot be explained as a result of lens distortion or a slanted calibration rig. The threshold δ depends on the expected lens distortion. We will use a default value of $\delta = 0.4$; the larger the distortion, the smaller should be the threshold value. A further run of the circle filter is performed to eliminate corners which have become unconnected due to differences in length.

At this point the rig is assumed to be one component of the graph. The next step is to identify the component which describes the rig. This is done by traversing the graph component starting from all rig corner candidates. During the traversal each square corner is assigned its presumed number (cf. Fig. 1.7d). A single inconsistency in this enumeration starting from different rig corners discards the complete image. During the traversal the number of vertices per component is counted. The largest component is assumed to be the rig. If the number of corners is not the expected one, the complete image is discarded. If the rig is rectangular, the algorithm checks for the correct size in both axes, i.e. a rig turned by 90° is discarded. This algorithm performs multiple runs, starting at different assumed rig corners and neglecting the image if the obtained enumerations are not identical.

The last step is to detect the direction marks on the rig (cf. Fig. 1.5) and to change the corner enumeration accordingly. For arbitrary camera orientations, the direction marks resolve the 180° ambiguity (or 90° ambiguity in the case of a quadratic rig) of the chequerboard orientation. In order to detect the direction of each notch the grey values along two lines are extracted. These lines start in the middle of one square border and end in the middle of the opposite square border. The grey levels on the horizontal line lead to a higher standard deviation than those on the vertical line if the rig is horizontal, as in Fig. 1.6. Additionally, the weighted mean of these grey values yields a position that clearly detects them to be off-center. It is possible

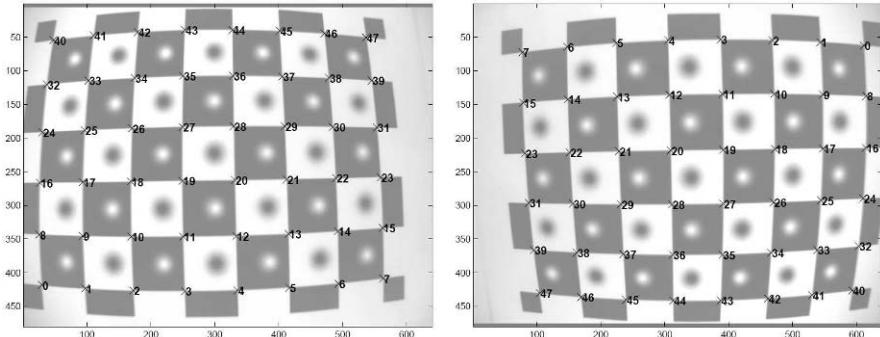


Fig. 1.8 Typical rig extraction results for images acquired with the PointGrey Digiclops wide-angle camera system. Strong lens distortion effects are apparent.

to threshold the standard deviation to detect missing notch marks if the processing options require them. In this case the image is discarded because expected notch marks cannot be found. The direction of the rig is determined by computing the mean direction vector from the intersection points of the two lines to the center of gravity. The direction of the mean vector is quantised into four directions (left, right, up, down) and the corner identifiers obtained in the previous step are adjusted accordingly based on the square counts.

A typical rig extraction result is shown in Fig. 1.8 for the trinocular PointGrey Digiclops wide-angle camera system with a field of view of 70° . The size of each square amounts to $30 \times 30 \text{ mm}^2$. The rig was printed out at 600 dpi and attached to a wooden board.

1.4.7.2 Comparison Between Automatic and Manual Calibration Rig Extraction

The computation times of the automatic methods are divided into two parts: The first and faster part is the rig finder implemented in C++. It takes about 0.3 seconds per image tuple on a standard Pentium IV processor. The second part is the intrinsic and extrinsic calibration, which is implemented in Matlab and takes additional 3.3 and 3.6 seconds per image tuple for the WM and the BVI approach, respectively. A manual calibration performed with the original software package by Bouguet (2007) based on the same images requires about 1 minute per image tuple.

The calibration procedure itself is performed after manual or automatic extraction of the control points. It is similar to the optimisation approach proposed by Bouguet (2007). In a first step, the intrinsic parameters of the cameras are determined independently, based on a bundle adjustment stage involving the control points “seen” by each of the cameras, respectively. This nonlinear optimisation procedure is initialised with the linear, homography-based method according to Zhang (1999a). In a second bundle adjustment step, the extrinsic camera parameters are de-

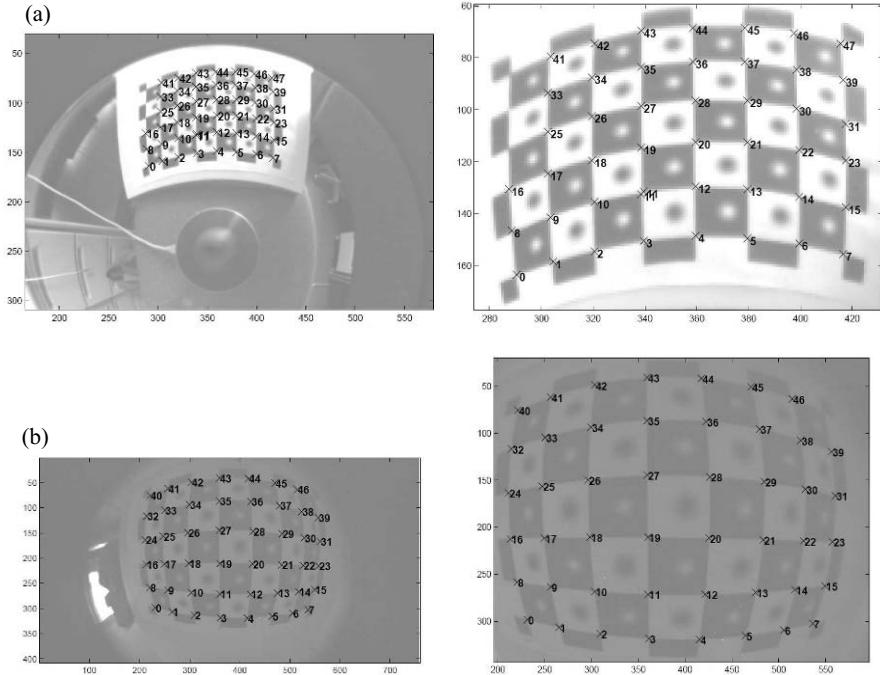


Fig. 1.9 (a) Extraction of the calibration rig in an image acquired with a catadioptric omnidirectional camera. (b) Extraction of the calibration rig in an image acquired with a fisheye lens. For both examples, the full image and an enlarged section are shown, respectively. The contrast has been reduced for visualisation purposes.

terminated by nonlinear optimisation, where the previously obtained intrinsic camera parameters remain unchanged.

The residual backprojection error strongly depends on the number of images available for calibration and on the positions in space where the calibration rig is located during image acquisition. Krüger (2007) states four “rules of thumb” to yield a suitable set of calibration images:

1. In order to be able to determine the lens distortion coefficients, images in which the calibration rig completely fills the image must be acquired for each camera.
2. The rig should be positioned at three different distances to the cameras, at the beginning, in the middle, and at the end of the working volume.
3. For all three distances, the calibration rig should also be placed such that it appears close to the corners of the images, respectively.
4. At each position, the calibration rig should be placed orthogonal to the optical axis as well as in four orientations rotated by an angle of about 30° around the horizontal and the vertical axis, respectively.

As long as these rules are adhered to, the residual root-mean-square backprojection error obtains typical values of approximately 0.1–0.2 pixels when 30–100 calibra-

tion images are used, independent of whether the control points are extracted manually or with the proposed automatic method. This order of magnitude is obtained both for the convergent stereo camera setup employed in the close-range inspection scenarios described in Chapter 5, which consists of two Baumer CCD cameras of 1032×776 pixels image size equipped with Cosmicar video lenses of $f = 25$ mm, and for the trinocular PointGrey DigiClops system, which is used in the application scenario of human–robot interaction described in Chapter 6. In the first scenario, the distance to the scene amounts to about 0.5 m, in the second scenario to 2–7 m. The uncertainty of the measured image coordinates of the control points due to pixel noise is examined in Section 1.4.8.

The proposed method for calibration rig extraction successfully performs the difficult task of automatically assigning the control points on the rig to their counterparts in the acquired calibration images. The subsequent camera calibration step depends on the camera model, which can be exchanged readily. We therefore conclude this section with two examples of automatic calibration rig extraction from images acquired with somewhat “exotic” lenses, an omnidirectional catadioptric lens and a fisheye lens (cf. Fig. 1.9). The proposed graph-based rig extraction algorithm performs well despite the strongly distorted shape of the calibration rig in the images. Even when not all control points are found (cf. Fig. 1.9a), the assignment of the extracted control points is correct. Hence, the proposed method may serve as a general-purpose approach to automatic calibration rig extraction for a variety of lens architectures.

1.4.8 Accurate Localisation of Chequerboard Corners

1.4.8.1 Different Types of Calibration Targets and Their Localisation in Images

Usually the calibration of a camera system—which may consist of multiple rigidly connected cameras—is performed by observing a calibration rig. At least the positions of special points on this planar or non-planar rig are known up to a reasonably high accuracy. These points are termed targets in photogrammetry. Given the positions of the targets in a set of images and their known spatial coordinates, a least mean squares error between the projected spatial points and the measured points is minimised. This requires suitable initial parameters such as the per-image transformation or the parameters of the camera model for each camera to be calibrated. For planar calibration rigs and pinhole cameras these parameters can be estimated using homographies (cf. Sections 1.4.3 and 1.4.4). Similar techniques can be used to reconstruct the spatial positions of the targets along with the camera calibration.

Four aspects of this process influence the accuracy of the calibration: The suitability of the camera model for the given cameras and lenses, the accuracy of the calibration rig, the accuracy of the target detection algorithm, and the placement of the rig with respect to the camera. The accuracy of the calibration rig is a manu-

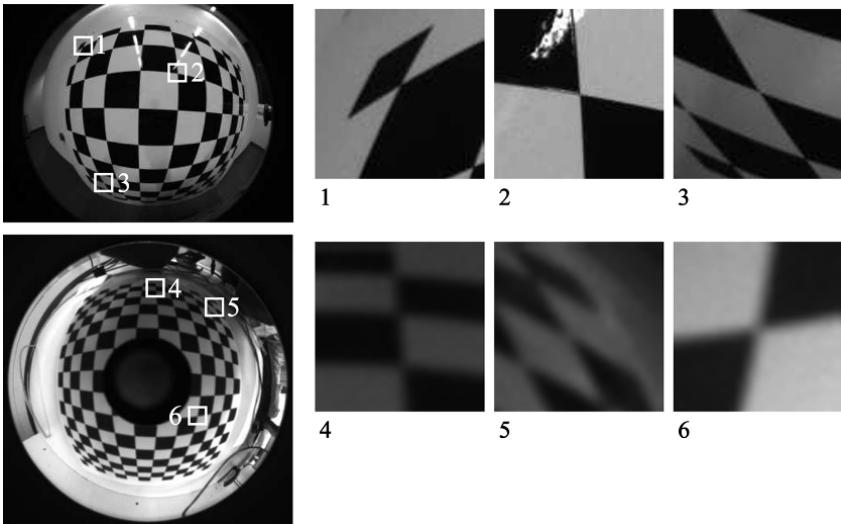


Fig. 1.10 Top: Image of a chequerboard pattern acquired with a camera of 1.9 megapixels image size, equipped with a fisheye lens of 185° field of view. Bottom: Image taken with a catadioptric omnidirectional camera of 640×480 pixels resolution. Enlarged example corners are shown on the right, respectively.

factoring problem. Different camera models are e.g. described by Brown (Brown, 1966) (cf. also Section 1.1). The placement strategy of the calibration rig is outlined by Mason (1994) and Krüger (2007). In this section we examine the accuracy of the target detection algorithm for a novel technique and three reference approaches.

Photogrammetric applications usually deal with nearly ideal pinhole cameras, resulting in higher hardware costs but better image quality. Computer vision applications are often concerned with optics that display strong distortion effects. Especially, many mobile robotic systems employ non-pinhole optical systems such as fisheye lenses or catadioptric optics. Generally, corners, chequerboard corners, or circular markers are used. Examples of images displaying a chequerboard pattern acquired with such non-pinhole optical systems are shown in Fig. 1.10 along with typical chequerboard corners extracted from these images.

Chequerboard corners and are commonly used for camera calibration in computer vision applications, whereas circular dots are more common in the domain of photogrammetry. Mallon and Whelan (2006) compare these targets and find the chequerboard corners to be bias-free with respect to projective transformations and nonlinear distortions, while circular markers are influenced by these effects. A simple explanation for this behaviour relies on the fact that chequerboard corners are invariant with respect to scale as long as the localisation window only contains the corner. The coordinate of interest is the intersection of the edges of the four adjacent chequerboard fields, i.e. a point. The projection of a point is invariant with respect to these influences as it remains a point. In contrast, the centre of a projected circle is not necessarily found at the same location as the projection of the centre of

the observed ellipse, such that circular markers suffer from a bias in their centre coordinates. This effect becomes increasingly pronounced with stronger distortion effects.

Mallon and Whelan (2006) present an edge-based nonlinear chequerboard corner localiser. Their method performs a least-squares fit of a parametrically modelled edged image to the acquired edge image. The presented approach relies on the method by Li and Lavest (1995), where it is applied to calibration rigs consisting of white lines on black background (KTH calibration cube). The drawback of this method is its dependence on edge images. Furthermore, the original method by Li and Lavest (1995) performs a rather ad-hoc approach to the modelling of the corner image, since it uses 11 parameters compared to the 7 parameters utilised here. The four additional parameters account for an illumination gradient, different reflectivities of the white lines, and different line widths. These additional parameters impose an unnecessary computational burden on the optimiser—in general, more model parameters increase the probability to get stuck in local minima.

One of the popular algorithms for finding the centre of a circular white target is the computation of the centre of gravity of a window around the detected position (Luhmann, 2003), where the square of the grey value serves as the weight of the pixel. This algorithm is simple and fast but sensitive to inhomogeneous grey values caused e.g. by defects of the target or incident light. This method is used as a reference in this section.

Locating a chequerboard corner can be achieved by various methods. Chen and Zhang (2005) use the intermediate values computed in the Harris-corner-like detection stage to obtain the second-order Taylor expansion of the image and find its saddle point. The underlying corner model is restricted to orthogonal corners. The result is invariant with respect to rotation, translation, contrast, and brightness, but not to projective transformations or nonlinear distortions. The algorithm by Lucchese and Mitra (2002) performs a least-squares fit of a second-order polynomial to a low-pass version of the input image. The position of the saddle point is obtained from the polynomial coefficients. The underlying model is invariant to affine transformations, contrast, and brightness. If the neighbourhood of the corner is chosen small enough, the affine invariance yields a suitable approximation in the presence of projective transforms and nonlinear distortions. It copes with the distortions introduced by fisheye lenses but requires a careful tuning of the size of the low-pass filter and at the same time the window size for fitting. The low-pass parameters depend mostly on the size of the point spread function (PSF) of the lens. We use this method as a further reference approach. Krüger et al. (2004) utilise intermediate information of the corner detector (cf. also Section 1.4.7). A correlation coefficient between the image and a corner template is computed at every pixel position, where peaks in this correlation coefficient image indicate the presence of chequerboard corners. The subpixel accurate position is obtained by computing the position of the local extremum by fitting a second-order polynomial to the correlation coefficient. Incidentally, this operation is identical to the method by Lucchese and Mitra (2002) but operates on different input data. For this reason we include the method as another reference approach.

Section 1.4.8.2 describes the method for chequerboard corner localisation suggested by Krüger and Wöhler (2009). A chequerboard corner model is derived along with a model adaptation procedure that copes with strong distortions and different PSF sizes while providing a high accuracy. As a byproduct, the radius of the PSF is obtained along with the target position.

Olague and Hernández (2005) present a technique that can be seen as a precursor to the method proposed in this study. They model an L-corner by two overlapping smoothed step functions using a Gaussian kernel as the smoothing operator and therefore model their step function by the Gaussian error function. We are using a similar step function model in our approach but perform an additional approximation. Since the Gaussian error function is provided by most numerical libraries as a piecewise polynomial approximation, it is quite slow to compute. We thus propose using a sigmoidal function based on a single exponential, which exists as a floating-point operation in many processor architectures and is therefore fast to compute. Furthermore, we use chequerboard corners instead of L-corners and assume a circular PSF, thus reducing the number of parameters to 7, compared to 12 by Olague and Hernández (2005).

The comparison between the different methods is performed using image sets labelled with a displacement in metric units, which in combination with the camera parameters yields the displacement in pixels as a ground truth. This experimentation setup differs from the approach of comparing the reprojection error of a subsequent bundle adjustment stage usually encountered in the literature (Luhmann, 2003).

1.4.8.2 A Model-based Method for Chequerboard Corner Localisation

We regard a square image window I^* with a width and height of $(2r + 1)$ pixels. The central pixel is assumed to have the coordinates $(0, 0)$ and to be the location of the most likely position of the corner as computed by the corner detector. An ideally sharp image \hat{I} of a corner with a brightness of -1 in the dark and $+1$ in the bright areas can be modelled by the product of two oriented step functions according to

$$\begin{aligned} \hat{I}(x, y) &= \delta(x \cos \alpha_1 + y \sin \alpha_1) \delta(x \cos \alpha_2 + y \sin \alpha_2) \\ \text{with } x, y \in \mathbb{R} \text{ and } \delta(t) &= \begin{cases} -1 & \text{if } t < 0 \\ +1 & \text{otherwise} \end{cases} \end{aligned} \quad (1.92)$$

The angles α_1 and α_2 denote the directions of the normals to the black-white edges. This notation is identical to the affine transformation of an orthogonal, infinitely large corner. Since we assume that r is sufficiently small (e.g. $r = 9$ pixels), the affine transformation is a suitable approximation of the projective transform and the lens distortions. Otherwise the straight line edges may be replaced by a different model such as cubic curves.

The ideal image \hat{I} is subject to blurring by the lens. An exact description of the PSF due to diffraction of monochromatic light at a circular aperture is given by the radially symmetric Airy pattern $A(r) \propto [J_1(r)/r]^2$, where $J_1(r)$ is a Bessel function

of the first kind of first order (Hecht, 2001). Consequently, the image of a point light source is radially symmetric and displays an intensity maximum at its centre and concentric rings surrounding the maximum with brightnesses which decrease with increasing ring radius. It is explained in more detail in Chapter 3 that for practical purposes a radially symmetric Gaussian function is a reasonable approximation to the PSF. It is thus assumed that the PSF is an ideal circular Gaussian filter G of radius σ . Hence, the continuous image $\tilde{I}(x,y)$ of the ideal chequerboard corner $\hat{I}(x,y)$ corresponds to

$$\tilde{I}(x,y) = G\left(\sqrt{x^2+y^2}, \sigma\right) * \hat{I}(x,y) \quad \text{with} \quad G(t, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2\sigma^2}}. \quad (1.93)$$

It is formed by convolving the step image with the circular Gaussian filter. Since the Gaussian filter is separable, we may exchange the step function $\delta(t)$ in $\hat{I}(x,y)$ by the step response of the Gaussian filter in $\tilde{I}(x,y)$. Hence, the observed intensity pattern corresponds to the step response $H(t, \sigma)$ of the PSF $G(t, \sigma)$ with

$$H(t, \sigma) = \operatorname{erf}\left(\frac{t}{\sqrt{2}\sigma}\right). \quad (1.94)$$

The error function $\operatorname{erf}(t)$ is twice the integral of the Gaussian distribution with zero mean and variance $1/2$, i.e. $\operatorname{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-s^2} ds$. It is scaled such that its infimum and supremum are $+1$ and -1 , respectively. A model of the observed intensity distribution of the chequerboard corner is then given by

$$\tilde{I}(x,y) = H(x \cos \alpha_1 + y \sin \alpha_1, \sigma) H(x \cos \alpha_2 + y \sin \alpha_2, \sigma). \quad (1.95)$$

The error function $\operatorname{erf}(t)$ is of sigmoidal shape but cannot be expressed in closed form. In practice it is approximated numerically and quite expensive to compute. In order to achieve an acceptable computational performance of the implementation we approximate the Gaussian error integral by the function $S(t, \tilde{\sigma})$ according to

$$\begin{aligned} \tilde{I}(x,y) &\approx S(x \cos \alpha_1 + y \sin \alpha_1, \tilde{\sigma}) S(x \cos \alpha_2 + y \sin \alpha_2, \tilde{\sigma}) \\ \text{with } S(t, \tilde{\sigma}) &= \frac{2}{1 + e^{-t/\tilde{\sigma}}} - 1. \end{aligned} \quad (1.96)$$

The function $S(t, \tilde{\sigma})$ is also of sigmoidal shape and similar to the logistic function $L(t) = 1/(1 + e^{-t})$. In Eq. (1.96), $\tilde{\sigma}$ is a scaling factor which is proportional to the width parameter σ of the Gaussian PSF $G(t, \sigma)$. Setting $\tilde{\sigma} = \sigma\sqrt{\pi/8}$ yields an identical slope of $H(t, \sigma)$ and $S(t, \tilde{\sigma})$ at $t = 0$ (Fig. 1.11). Note that the function $S(t, \tilde{\sigma})$ defined in Eq. (1.96) can also be expressed as

$$S(t, \tilde{\sigma}) = \tanh\left(\frac{t}{2\tilde{\sigma}}\right). \quad (1.97)$$

To determine the discrete model I_{uv} of the image, we assume a linear camera response described by gain β and offset γ and sample $\tilde{I}(x,y)$ at the integer-valued

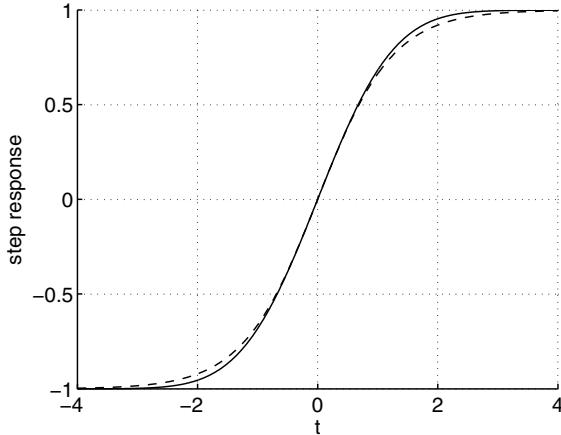


Fig. 1.11 True step response $H(t, \sigma)$ with $\sigma = 1$ (solid curve) and approximation by the function $S(t, \tilde{\sigma})$ with $\tilde{\sigma} = \sigma \sqrt{\pi/8}$ (dashed curve).

pixel positions (u, v) according to

$$I_{uv} = \beta \tilde{I}(u + u_0, v + v_0) + \gamma \quad \text{with } u, v \in \mathbb{N} \quad \text{and} \quad u_0, v_0 \in \mathbb{R}. \quad (1.98)$$

Again this is an approximation as each pixel of the sensor actually performs an integral over the area covered by it. Since the corner model has been fixed to the previously detected corner point, we have to move the corner with respect to the centre pixel of I_{uv} . In order to obtain the subpixel accurate corner position (u_0, v_0) of the corner in the input image I_{uv}^* we find the simulated corner image I_{uv} that is the best approximation of I_{uv}^* in the least-mean-squares sense by determining

$$\arg \min_{u_0, v_0, \beta, \gamma, \tilde{\sigma}, \alpha_1, \alpha_2} \left(\sum_{u,v} [I_{uv} - I_{uv}^*]^2 \right) \quad (1.99)$$

using the Levenberg-Marquardt algorithm. For clarity, the dependence of I_{uv} on u_0 , v_0 , β , γ , $\tilde{\sigma}$, α_1 , and α_2 has been omitted in Eq. (1.99). Gain and offset are initialised from the minimum and maximum grey value in I_{uv}^* . The angles α_1 and α_2 are initialised based on the polynomial coefficients obtained with the approach by Lucchese and Mitra (2002), but can also be set to fixed values. The parameter $\tilde{\sigma}$ is initialised by 1 and u_0 and v_0 by zero values.

1.4.8.3 Experimental Evaluation

The performance of the proposed nonlinear model adaptation procedure is evaluated based on real-world images. Traditionally, calibration target localisation ap-

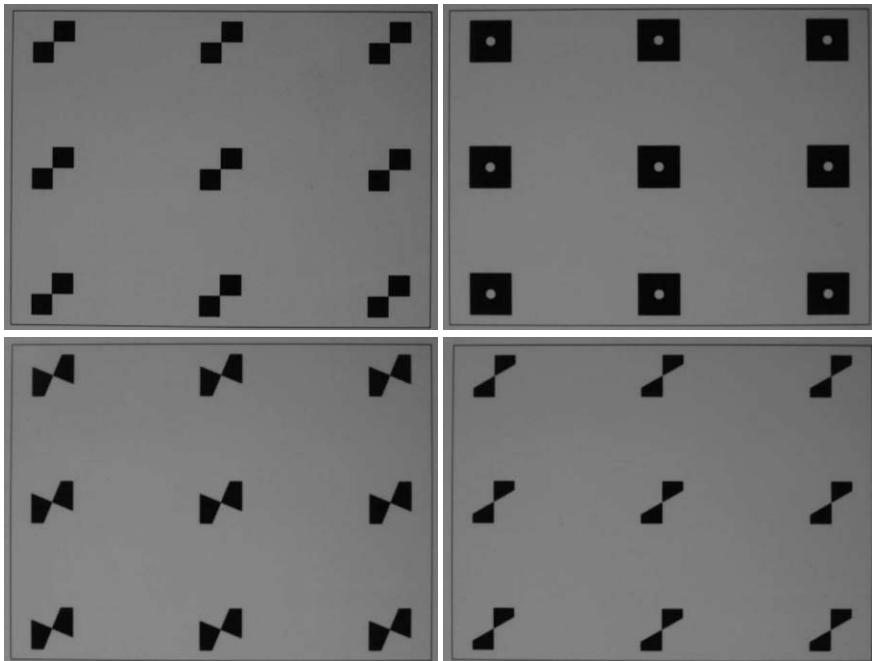


Fig. 1.12 Example input images for the evaluation of the examined chequerboard corner localisation methods. The x axis is in horizontal and the y axis in vertical image direction.

proaches are evaluated either using synthetic data (Lucchese and Mitra, 2002) or by performing a bundle adjustment based scene reconstruction or camera calibration (Luhmann, 2003; Olague and Hernández, 2005).

The proposed algorithm is not compared with the reference methods on synthetic data since synthetic data only show how well the synthesis and the analysis algorithms match. To a limited extent, they may provide quantitative information about noise sensitivity. Furthermore, the number of influencing parameters that have to be modelled by synthetic data are quite large. Apart from general and statistical influences (e.g. brightness, contrast, digitalisation noise) there are a lot of local influences with even more parameters (e.g. brightness gradient, inhomogeneity of illumination and pattern). Modelling all these influences realistically in a way that it does not accidentally favour one method over another is difficult, if not unfeasible at all. In a bundle adjustment based calibration (Olague and Hernández, 2005), the various sources of error such as the camera model or the calibration rig error are mixed such that clear distinctions between them are hard or even impossible to obtain when it is desired to determine the accuracy of the employed technique for chequerboard corner localisation.

For these reasons, we recorded a set of 100 images of a board with chequerboard corners printed on it (cf. Fig. 1.12 for examples), which was attached to a high-precision positioning unit, moved the board horizontally in front of the camera, and

Table 1.2 Plot labels and examined methods for chequerboard corner localisation.

Plot label	Method
linear	Fit a second order polynomial to low-pass filtered input image (Lucchese and Mitra, 2002)
nonlinear	Fit a corner model using Levenberg-Marquardt (this section)
parabolic	Fit a second order polynomial to correlation coefficient image (Krüger et al., 2004)
circle	Weighted mean (for circular markers only) (Luhmann, 2003)

recorded the next set of images. In order to attribute the images with the relative displacements in pixels, we obtained the overall scaling factor by measuring the subpixel accurate width of the box of size $64 \times 48 \text{ mm}^2$ around the targets. The grey levels of the box borders along a row were each approximated by a parabola and the positions of the local extrema were used to compute the width of the box in pixels at an accuracy of about 0.2 pixels. The resulting horizontal pixel scale amounts to $s_x = 8.910 \pm 0.003 \text{ pixels mm}^{-1}$ on the surface of the positioning unit, implying a maximum systematic error of the chequerboard corner displacements due to inaccurate knowledge of the pixel scale s_x of about 0.03 percent.

The images were acquired with a Sony DFW-V500 grey level camera with an image size of 640×480 pixels and 8 bits per pixel, using a lens with a focal length of 12.5 mm, giving a horizontal field of view of 21.7° . Regarding the random fluctuations of the pixel grey values, the signal-to-noise ratio amounts to about 50 for the bright and 10 for the dark image regions. These values are fairly typical of standard industrial CCD cameras. The camera was mounted at a distance of 190 mm from the positioning unit with the optical axis parallel to the normal of its surface. Specifically, the mechanical setup ensures that the deviations β_x and β_y from orthogonality of the optical axis with respect to the x and y axis of the positioning unit are smaller than 0.5° . If we assume that the optical axis of the camera is inclined horizontally by the angle β_x , perspectival foreshortening of the surface of the positioning unit implies an unforeshortened horizontal pixel scale $s_x^{(0)} = s_x / \cos \beta_x$ as long as β_x is smaller than a few degrees (for larger angles the finite distance of the camera from the positioning unit would become relevant). At the same time, a horizontal displacement Δx in millimetres on the positioning unit translates into a horizontal displacement Δu in pixel coordinates of $\Delta u = \Delta x s_x^{(0)} \cos \beta_x = \Delta x s_x$. Hence, systematic errors of the pixel scale s_x and the horizontal displacement Δu in pixel coordinates due to a small nonzero horizontal deviation β_x from orthogonality compensate each other, and no systematic errors are introduced.

A non-zero deviation β_y of the optical axis from orthogonality in y direction has the effect that the upper and lower row of targets (cf. Fig. 1.12) have effective horizontal pixel scales which are different from the value s_x determined for the middle row. For $\beta_y = 0.5^\circ$, the relative differences amount to 0.1 percent in our setup but compensate each other on the average—when the pixel scale is higher for the top row it is lower by the same relative amount for the bottom row and vice versa.

Based on a calibration of the intrinsic camera parameters according to Section 1.4.7, we found that the lens displays only insignificant distortions. The lens distortion was measured to be less than 0.2 pixels difference between distorted and undistorted image near the image corners, leading to a maximum relative shortening of displacements in the image corners of 0.1 percent.

As the random scatter (standard deviation) of all chequerboard corner localisation techniques examined in this study is always higher than 1 percent of the displacement and corresponds to about 10 percent for the most accurate method when averaged over all displacements, it can be safely assumed that the ground truth is sufficiently accurate for the evaluation.

The positioning unit was moved to the positions 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 300, and 400 μm along the x (horizontal) axis. The images were processed beginning with manually entered starting points. The window size was set to 19×19 pixels, corresponding to $r = 9$ pixels. For each of the nine corner targets in each image the position was determined using the nonlinear technique described in this section and the methods by Lucchese and Mitra (2002) and Krüger et al. (2004). The location of each of the nine circular markers was obtained based on the weighted mean method with the square of the grey levels as weights (Luhmann, 2003). In order to investigate the influence of the bias due to perspective transformation, the targets are arranged in a 3×3 grid. The target in the image centre is subject to the smallest amount of bias as the camera observes it nearly perpendicularly. The targets near the image edges and those situated in the image corners are observed under a more oblique angle, hence they are subject to a stronger bias.

Since no absolute ground truth positions are available but accurate displacements, the ground truth displacements are compared to the displacements estimated by the various operators. This is accomplished by selecting two random images of the same target and computing the pixel distance between the estimated absolute corner positions. The random selection chooses both image pairs that have a zero ground truth displacement and those that have a non-zero ground truth displacement. This ensures an appropriate consideration of the image noise which influences the localisation process. Identical image pairs are selected for the different operators. Additionally, the stability of the three methods is evaluated with respect to the starting position of the nonlinear model adaptation procedure. The subpixel accurate corner positions are computed starting from the eight pixels around the manually selected starting position and the starting position itself. The mean and standard deviation of the deviation between ground truth and estimated position is shown for all four methods in Figs. 1.13–1.15. Furthermore, the deviations depending on the target position in the image (centre, edge, corner), on the target rotation, and on the target shear are indicated. All targets were acquired with two different contrasts, white on black and light grey on dark grey. A further image set shows the targets with a different PSF, acquired after slightly changing the focus setting of the lens.

Table 1.2 relates the methods to the plot labels. Averaged over all examined targets, the proposed nonlinear fitting algorithm is about two times more accurate than both the linear algorithm and the centre of gravity based circle localiser. It is about three times more accurate than the parabolic approximation of the filter response

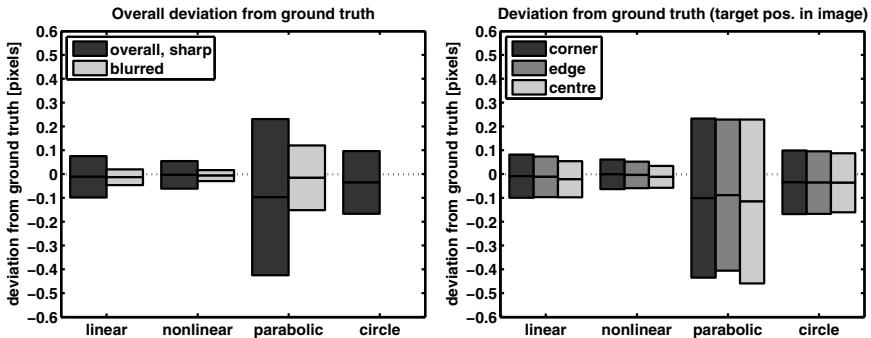


Fig. 1.13 Left: Overall deviation from ground truth for the different methods. Right: Dependence of the deviation on the target position in the image. The bars indicate the root mean square error, the central line of each bar the mean deviation between the measured and the true target displacement.

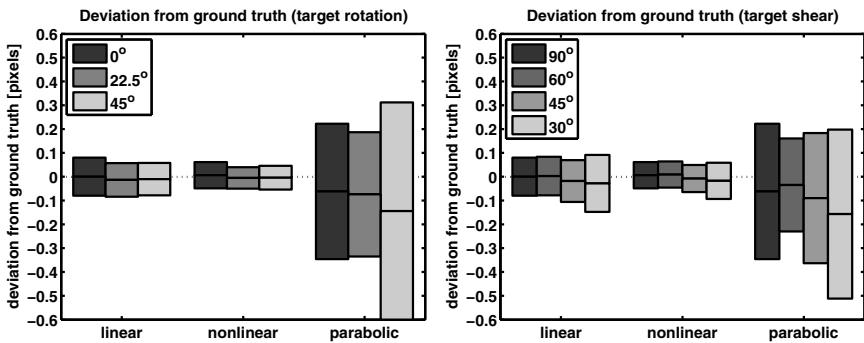


Fig. 1.14 Left: Dependence of the deviation from ground truth on the amount of target rotation. Right: Dependence of the deviation on the amount of target shear.

peak (Fig. 1.13). Furthermore, a much better performance of the linear and nonlinear algorithms is observed when the image is slightly blurred. Fig. 1.13 also shows that the performance of the algorithms is virtually independent of the target position in the image. The observed slight differences can be attributed to random variations.

Fig. 1.14 shows that the linear and nonlinear chequerboard corner localisers are basically rotation invariant as the standard deviations are nearly identical for all rotations. A slight bias increasing with decreasing angle can be observed. This trend is of minor importance, thus it can be attributed to random variations in the data set. For a large amount of shear, the proposed nonlinear algorithm is still of superior performance, compared to the linear fit. Even for an angle of only 30° between the sides of the chequerboard corner it performs better than the linear fit under any condition. The parabolic peak approximation has a standard deviation about three times larger than that of the other methods.

Fig. 1.15 shows that the poor performance of the circle localiser over all images is caused by its reduced robustness regarding inhomogeneous illumination and that

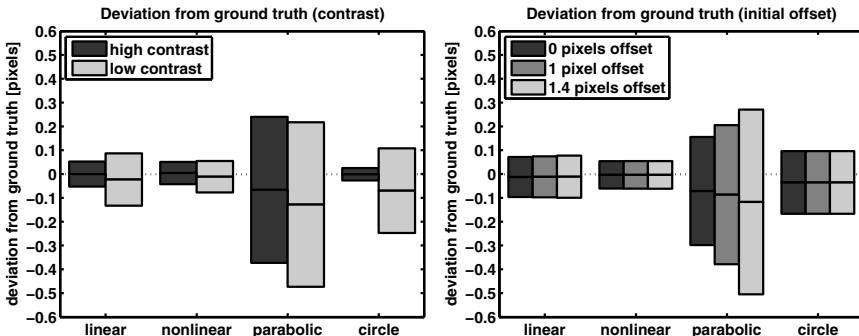


Fig. 1.15 Left: Dependence of the deviation from ground truth on the image contrast. Right: Dependence of the deviation on the starting position.

the poor performance of the parabolic peak approximation is caused in part by its reduced robustness with respect to the offset from the starting position. The performance of the linear and the nonlinear corner modelling rivals that of the circle localiser for non-blurred high contrast images. This behaviour is caused by the better signal-to-noise ratio since the location relevant signal (i.e. the step response) is of lower spatial frequency than in the non-blurred case and therefore provides more data for the model fit. Furthermore, Fig. 1.15 shows that the nonlinear method performs significantly better than both reference methods under low-contrast conditions. The fact that the centre of gravity circle localiser requires uniform illumination and reflectivity for best results is well-known from previous studies (Luhmann, 2003). In practice, this method requires constructive measures such as retro-reflective targets and confocal illumination. Our experiment did not provide these in order to investigate the influences of a purposefully simple imaging situation. Although the nonlinear method results in a higher standard deviation and a slight bias due to the inhomogeneous illumination, the performance drop is not as pronounced as for the reference methods. The nearly quadrupled standard deviation of the circle localiser along with a bias larger than the standard deviation of the corresponding high contrast result illustrate the superior robustness of the nonlinear method, which shows only an insignificant bias. The peak approximation is not significantly influenced by the contrast as the correlation coefficient is brightness and contrast invariant. Even in a low contrast setting the peak approximation is outperformed by both the linear and the nonlinear corner model. Three of the examined algorithms are robust to variations of the initial position in a 3×3 neighbourhood. The peak approximation shows a clear trend regarding its accuracy, since if the initial corner detector is off by only one pixel diagonally, the deviation from ground truth increases by about one third.

At this point it should be noted that all stated accuracies refer to displacements, i.e. the differences of target positions. According to the law of error propagation (Bronstein and Semendjajew, 1989) and assuming identical error distributions for the position measurements used to determine a set of differences, the actual errors

of the corner positions are smaller by a factor of $1/\sqrt{2}$ than the displacement errors stated above. Hence, the proposed nonlinear chequerboard corner localisation algorithm yields a corner localisation accuracy of better than 0.041 pixels regardless of target rotation, shear, or image position for high contrast images. The accuracy is better than 0.047 pixels for low contrast images and better than 0.016 pixels for slightly blurred images. These values refer to the accuracy of the determined corner point position and are inferred from the accuracies obtained for differences between pairs of corner positions.

Compared to the traditional centre of gravity based localisation algorithm for circular markers, the proposed algorithm allows to obtain accurate calibration and scene reconstruction results largely independent of the camera model under less favourable illumination conditions. Furthermore, the proposed algorithm yields the PSF width parameter σ , which is necessary for the computation of depth from defocus (Kuhl et al., 2006; Wöhler et al., 2009), and is therefore also suitable for the calibration of the depth–defocus function of the optical system as described in Section 3.2.

1.5 Stereo Image Analysis in Standard Geometry

When the intrinsic and extrinsic camera parameters are known or can be inferred from a reference set of point correspondences, it is advantageous to rectify the stereo image pair to standard geometry, corresponding to parallel optical axes, identical principal distances, collinear horizontal image axes, and image planes which are orthogonal to the optical axes and parallel to the stereo baseline. Accordingly, pairs of corresponding epipolar lines become parallel to each other and to the horizontal image axes. The important advantage of performing stereo analysis in image pairs rectified to standard geometry is the fact that the search for point correspondences needs to be performed only along corresponding pixel rows of the rectified images.

1.5.1 *Image Rectification According to Standard Geometry*

In a first step, the original images are warped such that the radial and tangential lens distortions described by Eqs. (1.3) and (1.4) are compensated. Image rectification then essentially corresponds to determining a projective transformation for each image such that the conditions stated above are fulfilled. To obtain parallel epipolar lines, Hartley and Zisserman (2003) propose a method to map the epipoles to infinity and determine an appropriate projective transformation for each image based on a set of point correspondences. Their approach does not require that the extrinsic camera parameters are known. A more compact algorithm for the rectification of stereo image pairs, requiring calibrated cameras, i.e. knowledge about their intrinsic parameters A_i and their extrinsic parameters R_i and t_i , is introduced by Fusiello

et al. (2000). Basically, geometric camera calibration is achieved by applying the principle of bundle adjustment to a reference object of known size and shape. An overview of geometric camera calibration methods is given in Section 1.4. In the algorithm by Fusiello et al. (2000), the image formation process of camera i is defined by the corresponding projective transformation

$$P_i = A_i [R_i \mid \mathbf{t}_i] = [B_i \mid \mathbf{b}_i] \quad \text{with} \quad B_i = A_i R_i \quad \text{and} \quad \mathbf{b}_i = A_i \mathbf{t}_i. \quad (1.100)$$

The optical centre ${}^W\mathbf{c}_i$ of camera i in the world coordinate system is related to the translation vector \mathbf{t}_i by $\mathbf{t}_i = -R_i {}^W\mathbf{c}_i$. According to Eq. (1.100), we then obtain

$$\mathbf{b}_i = -B_i {}^W\mathbf{c}_i. \quad (1.101)$$

The image point ${}^{S_i}\tilde{\mathbf{x}}$, defined in pixel coordinates in image i , is obtained from the corresponding scene point ${}^W\tilde{\mathbf{x}}$ by ${}^{S_i}\tilde{\mathbf{x}} = P_i {}^W\tilde{\mathbf{x}}$. The optical ray \mathbf{s} associated with this image point is given in parametric form by

$$\mathbf{s} = {}^W\mathbf{c}_i + \lambda B_i^{-1} {}^{S_i}\tilde{\mathbf{x}} \quad (1.102)$$

with λ as a real number.

Since we assume that the stereo cameras are calibrated, the projective transformations P_1 and P_2 are known. The rectification procedure now consists of defining two new projective transformations $P_1^{(s)}$ and $P_2^{(s)}$ obtained by rotating the original cameras around their optical centres such that the image planes become coplanar and are parallel to the baseline, i.e. the line connecting the optical centres of the cameras. This operation ensures that the epipolar lines become parallel. The optical centres themselves remain unchanged. To obtain horizontal epipolar lines, the baseline must be parallel to the new horizontal image axis of both cameras. Additionally, corresponding points must have the same vertical image coordinate v . This configuration is obtained by requiring that the new cameras have the same intrinsic parameters, i.e. $A_1 = A_2 \equiv A$. The new projective transformations $P_i^{(s)}$ are then given by

$$P_i^{(s)} = A [R \mid -R {}^W\mathbf{c}_i]. \quad (1.103)$$

The optical centres ${}^W\mathbf{c}_i$ are obtained by ${}^W\mathbf{c}_i = -B_i^{-1}\mathbf{b}_i$ according to Eq. (1.101). Since the optical axes are parallel, the new rotation matrix R is the same for both cameras. It is defined by its row vectors according to

$$R = \begin{bmatrix} \mathbf{r}_1^T \\ \mathbf{r}_2^T \\ \mathbf{r}_3^T \end{bmatrix}, \quad (1.104)$$

denoting the x , y , and z axis, respectively, of the camera coordinate system, expressed in world coordinates. To take into account the previously described conditions for standard geometry, we define the new x axis to be parallel to the baseline, i.e. $\mathbf{r}_1 = ({}^W\mathbf{c}_1 - {}^W\mathbf{c}_2) / |{}^W\mathbf{c}_1 - {}^W\mathbf{c}_2|$. We then define a unit vector \mathbf{k} which corre-

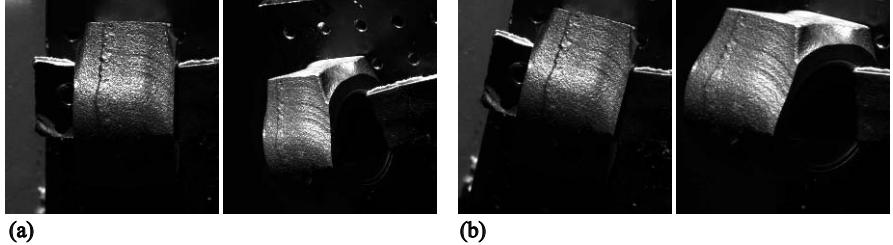


Fig. 1.16 (a) Pair of stereo images, acquired by two cameras with non-parallel optical axes. (b) The same image pair after rectification to standard geometry.

sponds to the z unit vector of the original coordinate system of camera 1. The new y axis is given by $\mathbf{r}_2 = \mathbf{k} \times \mathbf{r}_1$, thus being orthogonal to the new x axis and orthogonal to the z axis of the original coordinate system of camera 1. The new z axis corresponds to $\mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2$. This algorithm fails when the optical axes are parallel to the baseline.

In order to rectify the images, it is necessary to compute the transformation that maps the original image plane defined by $P_i = [B_i \mid \mathbf{b}_i]$ to the new image plane defined by $P_i^{(s)} = [B_i^{(s)} \mid \mathbf{b}_i^{(s)}]$. For an arbitrary scene point ${}^W\mathbf{x}$ the expressions $S_i \tilde{\mathbf{x}} = P_i {}^W\tilde{\mathbf{x}}$ and $S_i^{(s)} \tilde{\mathbf{x}} = P_i^{(s)} {}^W\tilde{\mathbf{x}}$ are valid. According to Eq. (1.102), the corresponding optical rays are given by

$$\begin{aligned} {}^W\mathbf{x} &= {}^W\mathbf{c}_i + \lambda B_i^{-1} S_i \tilde{\mathbf{x}} \\ {}^W\mathbf{x} &= {}^W\mathbf{c}_i + \lambda^{(s)} (B_i^{(s)})^{-1} S_i^{(s)} \tilde{\mathbf{x}}, \end{aligned} \quad (1.105)$$

which directly implies

$$S_i^{(s)} \tilde{\mathbf{x}} = \frac{\lambda}{\lambda^{(s)}} B_i^{(s)} B_i^{-1} S_i \tilde{\mathbf{x}}. \quad (1.106)$$

Note that the integer-valued pixel positions in the rectified image generally correspond to non-integer positions in the original image. Hence, the pixel grey values of the rectified image are computed based on bilinear interpolation. An example of a stereo image pair originally acquired with a convergent camera setup and rectified to standard geometry is shown in Fig. 1.16.

The first two coordinates of the projective vector $S_i^{(s)} \tilde{\mathbf{x}}$ correspond to the rectified pixel values $u_i^{(s)}$ and $v_i^{(s)}$. The row $v_1^{(s)}$ of image 1 and the row $v_2^{(s)} = v_1^{(s)}$ of image 2 now form a corresponding pair of epipolar lines. As a consequence, the problem of stereo image analysis becomes a problem of establishing corresponding points along image rows. We will now regard the computation of three-dimensional scene structure from the coordinates of image points measured in stereo image pairs rectified to standard geometry. Without loss of generality, at this point the coordinate system of camera 1 is used as the world coordinate system, such that the corresponding rotation matrix R_1 corresponds to the identity matrix and the translation vector

\mathbf{t}_1 is zero, and we define $\mathbf{t} \equiv \mathbf{t}_2$. The distance $\|\mathbf{t}\|$ between the optical centres of the two cameras corresponds to the baseline distance of the stereo camera system. As the images are rectified to standard geometry, both optical axes are perpendicular to the baseline. Due to the rectification to standard geometry, both cameras have the same effective principal distance b_0 .

The image coordinates in standard geometry are denoted by $u_1^{(s)}$ and $v_1^{(s)}$ in image 1 and by $u_2^{(s)}$ and $v_2^{(s)}$ in image 2, respectively. As we assume that the world coordinate system corresponds to the coordinate system of camera 1, a scene point can be described by ${}^W\mathbf{x} = {}^{C_1}\mathbf{x} = (x, y, z)^T$. We furthermore assume square pixels with $k_u = k_v$, corresponding to a pixel edge length $d_p = 1/k_u$, and without skew. Eq. (1.1) then implies

$$\begin{aligned} \frac{d_p u_1^{(s)}}{b_0} &= \frac{x}{z} \\ \frac{d_p u_2^{(s)}}{b_0} &= \frac{x - \|\mathbf{t}\|}{z} \\ \frac{d_p v_1^{(s)}}{b_0} &= \frac{d_p v_2^{(s)}}{b_0} = \frac{y}{z} \end{aligned} \quad (1.107)$$

The three unknowns x , y , and z can be obtained by solving these three equations, leading to the expressions

$$\begin{aligned} x &= \|\mathbf{t}\| \frac{u_1^{(s)}}{u_1^{(s)} - u_2^{(s)}} \\ y &= \|\mathbf{t}\| \frac{v_1^{(s)}}{u_1^{(s)} - u_2^{(s)}} \\ z &= \|\mathbf{t}\| \frac{b_0}{d_p} \frac{1}{u_1^{(s)} - u_2^{(s)}} \end{aligned} \quad (1.108)$$

(Horn, 1986). The difference $d = u_1^{(s)} - u_2^{(s)}$ occurring in Eq. (1.108) is termed disparity. The distance z is inversely proportional to the disparity, while the disparity is proportional to the baseline $\|\mathbf{t}\|$ for a given distance z . If we assume a fixed measurement error for the disparity, the accuracy of the determination of z increases with increasing baseline. On the other hand, the images become less similar with increasing baseline, and some parts of the scene imaged by camera 1 may even be invisible to camera 2, such that it becomes more difficult to reliably determine point correspondences between the images. Disparity also increases with increasing principal distance b_0 , corresponding to a magnification of the image, and also increases with decreasing pixel size d_p .

1.5.2 The Determination of Corresponding Points

In this section we assume that the regarded pair of stereo images has been rectified to standard geometry. Hence, the problem of three-dimensional scene reconstruction basically becomes a problem of establishing point correspondences along epipolar lines, i.e. corresponding image rows. Under these preconditions, the problem of three-dimensional scene reconstruction is solved as soon as a so-called disparity map has been generated, i.e. disparity values have been determined for all image pixels or a subset of them.

In order to minimise the number of false correspondences, several constraints can be imposed on the correspondence search. The uniqueness constraint (Marr and Poggio, 1979) requires that a given pixel from one image cannot correspond to more than one pixel from the other image. Here it is assumed that all objects in the scene are opaque. In the presence of occlusions, it may be impossible at all to find a corresponding point. The ordering constraint (Baker and Binford, 1981) requires that if a pixel is located to the left of another pixel in image 1, the corresponding pixels in image 2 must be ordered in the same manner, and vice versa, i.e. ordering of pixels is preserved across the images. The ordering constraint may be violated if an object in the scene is located much closer to the camera than the background, and if one pixel corresponds to a point on the object while the other pixel corresponds to a background point. A further constraint, which is valid only for scenarios in which smooth surfaces are reconstructed, is the continuity constraint (Marr and Poggio, 1979), requiring that the disparity map should vary smoothly almost everywhere in the image. This constraint becomes invalid at depth discontinuities in the scene.

We will now describe a variety of classical as well as recently developed methods for generating disparity maps. While these techniques are based on many different approaches to detect similarities between image regions, all of them exploit one or several of the described constraints on the correspondence search.

1.5.2.1 Correlation-based Blockmatching Stereo Vision Algorithms

Correlation-based blockmatching techniques directly compare small image windows along corresponding epipolar lines according to a pixel-based similarity measure S with

$$S(u_1, u_2) = C(I_1(V_s(u, v)), I_2(V_s(u - d(u, v), v))), \quad (1.109)$$

where I_1 is the intensity image 1, I_2 is the intensity image 2, $d(u, v)$ is the disparity, and V_s is a vector of pixels in a spatial neighbourhood of the pixel situated at (u, v) in image 1 and of the pixel situated at $(u - d(u, v), v)$ in image 2. An early description of the basic principle is provided by Horn (1986). Possible similarity measures defined by the function C in Eq. (1.109) are the cross-correlation coefficient, the sum of squared differences, or the sum of absolute differences (Franke and Joos, 2000). The image regions which display a sufficient amount of texture are extracted with an interest operator, e.g. a Sobel detector for vertical edges. In a second

step, point correspondences are established at pixel accuracy along corresponding epipolar lines by determining the optimum of the similarity measure. Hierarchical correspondence analysis at different resolution levels may significantly decrease the processing time. In a third step, the inferred integer disparity values can be refined to subpixel accuracy based on an interpolation of the measured similarity values by fitting a parabola to the local neighbourhood of the optimum (Franke and Joos, 2000). Correlation-based blockmatching algorithms are computationally efficient and thus favourably used in real-time vision systems.

An important drawback of correlation-based blockmatching approaches is the fact that uniform depth across a correlation window is assumed, which leads to depth errors near the borders of objects especially when large correlation windows are used. Hirschmüller (2001) introduces a framework which aims for correcting or decreasing the effects of depth discontinuities due to object borders. Instead of computing the similarity measure for a large window, it is inferred from several smaller neighbouring windows. Only those windows that contribute to the overall similarity measure in a consistent manner are taken into account. A left-right consistency check helps to invalidate uncertain correspondences. Accurate depth values at object borders are determined by splitting the corresponding correlation windows into two parts and searching for the optimum similarity measure separately on both sides of the object border. These improvements of the classical correlation-based blockmatching approach are demonstrated by Hirschmüller (2001) and in more detail by Hirschmüller et al. (2002) to significantly improve the overall performance of three-dimensional reconstruction.

1.5.2.2 Feature-based Stereo Vision Algorithms

General Overview

In the framework of feature-based stereo, for each pixel a vector of features is determined, and the similarity between two pixels located on corresponding epipolar lines is determined based on the similarity of the extracted feature vectors. Hence, the search for correspondences is applied to the attributes associated with the detected features. Edge elements, corners, line segments, and curve segments are quite robust against change of perspective and have therefore been used extensively in the domain of stereo vision.

Low-level features such as edge elements and corners are easy to extract but may suffer from occlusion. A well-known algorithm to extract edge elements is the Canny edge detector (Canny, 1986). Coordinates, local orientations, or the local intensity profile may be used as features for the correspondence search. For the detection of corners a variety of algorithms exists (Beaudet, 1978; Dreschler and Nagel, 1982; Kitchen and Rosenfeld, 1982; Zuniga and Haralick, 1983). Many state-of-the-art vision systems employ the Harris corner detector (Harris and Stephens, 1988). Apart from the coordinates of the extracted corners, the type of junction a corner corresponds to can be used for establishing corresponding points.

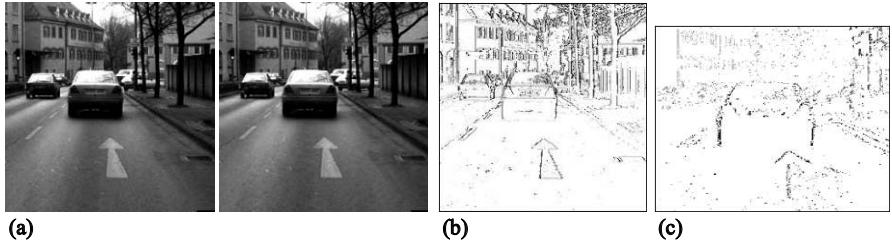


Fig. 1.17 (a) Pair of stereo images. (b) Classified image pixels. The pixel brightness encodes the class index. (c) Distance image. The pixel brightness is proportional to the distance. Images are taken from Franke et al. (1999).

Higher-level features such as line and curve segments may require a higher computational effort but tend to be more robust against occlusion, since they are spatially more extended and thus less likely to be completely occluded. Line segments are obtained by extracting edge elements in a first step with a suitable edge detector, which are then linked and merged based on criteria such as distance, orientation, similarity, or collinearity (Nevatia and Babu, 1980; Fischler and Bolles, 1983; Weiss and Boldt, 1986). The attributes that can be used for establishing correspondences include the coordinates of the end points, the centres, and the orientation of the line segments. Problems may occur in the presence of noise, leading to inaccuracies in the determined end points of the line segments, resulting in inaccurate disparity values, or to arbitrary splitting of line segments such that the uniqueness constraint remains no longer valid. Curve segments are still more difficult to match since it is usually not possible to establish unique correspondences between points on the curve. To circumvent these ambiguities, Deriche and Faugeras (1990) propose to utilise the turning points of curve segments as features.

Features on still higher levels, such as circles, ellipses, polygonal regions, or complete objects can also be utilised to establish correspondences between the images, but the applicability of such features tends to be restricted to indoor scenes containing objects of a limited range of shapes. The area of regions in the image, bounding line segments, and the centroids of the regions can be used as features to establish correspondences.

In many feature-based stereo vision systems it is possible to combine several features in order to reduce the number of false correspondences and to increase the accuracy inferred disparity values. For example, intensity, edges, and corners are combined in the system proposed by Weng et al. (1989) to form multiple attributes. A hierarchical approach regarding edges, curves, surfaces and two-dimensional regions is employed by Lim and Binford (1987) for establishing correspondences between the images on higher levels.

An example of a real-time feature-based stereo algorithm applied in the context of a vision-based driver assistance system is outlined by Franke and Kutzbach (1996) and by Franke et al. (1999). According to their approach, each pixel is classified according to the grey values of its four direct neighbours. It is determined

whether each neighbour is brighter or darker than the central pixel by a given amount or if it has a similar brightness. Accordingly, the pixel is assigned to one of $3^4 = 81$ classes. These pixel classes essentially encode edges and corners at different orientations. Two pixels on corresponding epipolar lines are regarded as potentially corresponding pixels when they belong to the same class (cf. Fig. 1.17). Since this assignment is not necessarily unique, in the case of ambiguities the pair with the smallest disparity, i.e. the largest distance, is chosen in order to avoid the formation of “phantom obstacles” close to the camera.

More recently, a similar assignment scheme has been proposed by Stein (2004) in the context of real-time optical flow analysis. In this approach, local neighbourhoods of arbitrary size can be regarded. The relation between the brightnesses of the central pixel and the neighbouring pixels is encoded as a chain of digits, where each digit may obtain the values 0 (darker than the central pixel), 1 (the absolute brightness difference does not exceed a predefined threshold), or 2 (brighter than the central pixel). This scheme can be regarded as an extension of the census transform (Zabih and Woodfill, 1994). Correspondences are established by searching for pixels in the two images with identical signatures. Some heuristic rules help to eliminate image regions that do not contain useful information, such as a region of uniform brightness. Furthermore, point correspondences are preferentially established based on signatures that do not occur frequently in the images.

A Contour-based Stereo Vision Algorithm

A feature-based stereo vision approach relying on the analysis of object contour segments has been introduced by Wöhler and Krüger (2003) in the context of surveillance of working areas in industrial production. This contour-based stereo vision (CBS) algorithm is based on the comparison of the current image pair with a pair of reference images. To detect changes between the current image and the reference image, we compute the absolute difference image. There are much more complex methods of change detection, cf. e.g. Durucan (2001) for an overview. Generally, however, these cannot guarantee that a zero image resulting from change detection is equivalent to the fact that current and reference image are identical. A zero difference image guarantees that current and reference image are exactly identical, which is of significant importance for the application scenario as a person in the surveillance area must not be missed under any circumstances.

The image pair is rectified to standard geometry. We transform the pair of difference images into binary images by thresholding with

$$\theta_0 = q\sigma_d \quad \text{with} \quad \sigma_d = \sqrt{2}\sigma_p, \quad (1.110)$$

where the pixel noise σ_p is the standard deviation of a camera pixel signal over time, given a constant input intensity, and the noise $\sigma_d = \sqrt{2}\sigma_p$ the resulting pixel noise of the difference image. In our experiments, we set $q = 3$ in order to detect only changes which are with 99 percent certainty significant with respect to pixel noise.

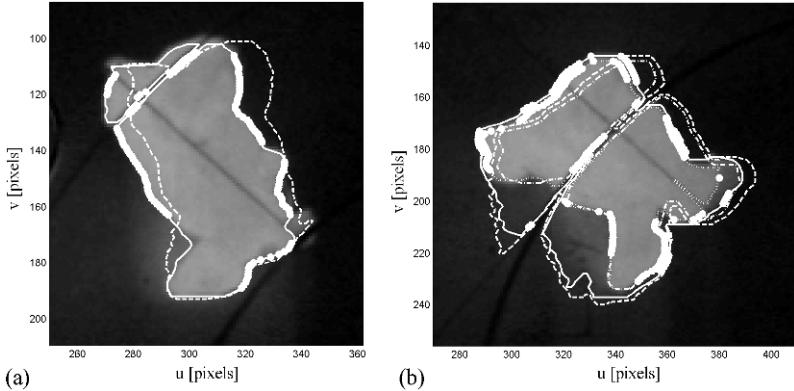


Fig. 1.18 (a) Contours extracted from absolute difference image with $\theta_0 = 100$. The solid contours have been extracted from the right image of the stereo pair, the dashed contours from the left image. Contour points on the solid contour for which a disparity value has been determined are marked by solid dots. Only the contours obtained with binary threshold θ_0 are shown. (b) Contours extracted from absolute difference image. The solid (binary threshold $\theta_0 = 50$) and the dotted (adaptive binary threshold derived from intensity histogram) contour have been extracted from the right image of the stereo pair, the dashed and the dash-dotted contour in a corresponding manner from the left image. The solid and the dashed contour are partially determined by shadow structures on the floor rather than by “real” object contours. Contour points for which a disparity value has been determined are marked by solid dots.

The image regions with pixel intensity above θ_0 are segmented using the binary connected components (BCC) algorithm (Mandler and Oberländer, 1990) which yields, among others, the properties area, centre coordinates and a contour description for each blob. This computation yields n_1 blobs on the right and n_2 blobs on the left image with centre coordinates $(U_1^{(i)}, V_1^{(i)})$, $i = 1, \dots, n_1$ and $(U_2^{(j)}, V_2^{(j)})$, $j = 1, \dots, n_2$. The extracted contours are smoothed by B-spline interpolation. Subsequently, we calculate the u (column) coordinates at which the interpolated contour intersects the rows in the v range covered by the corresponding B-spline in order to obtain subpixel accuracy. Hence, the contours are represented as sets $\{\mathbf{c}_{2,a}^{(m)}\}$ and $\{\mathbf{c}_{1,b}^{(n)}\}$ of points for the left and the right image, respectively:

$$\begin{aligned}\mathbf{c}_{1,b}^{(n)} &= (\bar{u}_{1,b}^{(n)}, \bar{v}_{1,b}^{(n)}) \\ \mathbf{c}_{2,a}^{(m)} &= (\bar{u}_{2,a}^{(m)}, \bar{v}_{2,a}^{(m)}),\end{aligned}\quad (1.111)$$

where $a = 1, \dots, n_1$ and $b = 1, \dots, n_2$ are the blob indices in the images and m and n the point indices in the respective contours. The values $\bar{u}_{1,b}^{(n)}$ and $\bar{u}_{2,a}^{(m)}$ are real numbers, while the values $\bar{v}_{1,b}^{(n)}$ and $\bar{v}_{2,a}^{(m)}$ denote image rows and thus are integer numbers.

In the following, a pair of blobs mutually assigned by the previously described procedure is regarded. For each epipolar line in the range covered by both blobs, the numbers of intersections between the contour and the epipolar line are calculated for both images. These intersection counts are denoted by $e_1(v)$ for the right and $e_2(v)$ for the left image of the stereo pair. Along with these values, the u coordinates of the epipolar intersections, $u_1^{(i)}(v)$, $i = 1, \dots, e_1(v)$ for the right image and $u_2^{(j)}(v)$, $j = 1, \dots, e_2(v)$ for the left image, are determined. They are known to subpixel accuracy due to the B-spline representation of the contours. The contour indices in the sets $\{\mathbf{c}_{1,b}^{(n)}\}$ and $\{\mathbf{c}_{2,a}^{(m)}\}$ corresponding to these epipolar intersections are denoted by $w_1^{(i)}$ and $w_2^{(j)}$, respectively. For each epipolar line v , the epipolar intersections are sorted in ascending order according to their respective $u_1^{(i)}$ and $u_2^{(j)}$ values. Assuming that the ordering constraint is valid, the following three cases have to be distinguished:

1. The contours in both images have an identical number of epipolar intersections, i.e. $e_1(v) = e_2(v)$. Then epipolar intersection \tilde{i} on the right image will be assigned to epipolar intersection \tilde{j} on the left image with $\tilde{i} = \tilde{j}$, respectively.
2. The contours on both images do not have an identical number of epipolar intersections, i.e. $e_1(v) \neq e_2(v)$, and either $e_1(v)$ or $e_2(v)$ is odd. In this case, the epipolar line is a tangent to the respective B-spline contour and therefore it is discarded.
3. The contours on both images do not have an identical number of epipolar intersections, i.e. $e_1(v) \neq e_2(v)$, and both $e_1(v)$ and $e_2(v)$ are even. Without loss of generality we assume $e_1 > e_2$. An even intersection index denotes an inward transition and an odd intersection index an outward transition. Hence, an intersection with even index j on the left image may only be assigned to an intersection with even index i on the right image, and analogously for odd indices, to account for the topology of the segmented blob features. According to the ordering constraint, we will always assign pairs of neighbouring intersections in the right image to pairs of neighbouring intersections in the left image, i.e. if intersection j is assigned to intersection i , intersection $j + 1$ will be assigned to intersection $i + 1$. According to these rules, $((e_1 - e_2)/2 + 1)$ assignments are allowed, for each of which we compute the sum of square differences:

$$S_j = \sum_{k=1}^{e_2} \left(u_1^{(k+2(j-1))} - u_2^{(k)} - d_{\min} \right)^2 \quad \text{for } j = 1, \dots, (e_1 - e_2)/2 + 1. \quad (1.112)$$

Epipolar intersection \tilde{i} on the right image will consequently be assigned to intersection $\tilde{j} = \arg \min_j \{S_j\}$ on the left image. This heuristic rule helps to avoid “phantom objects” situated near to the camera; similar heuristics are used in state-of-the-art correlation-based blockmatching algorithms (Franke and Joos, 2000).

The mutual assignment of contour points on epipolar line v results in pairs of indices of the form (\tilde{i}, \tilde{j}) . A disparity measure d_s that involves a single epipolar line v can

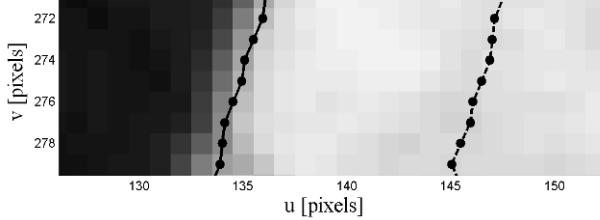


Fig. 1.19 Two contour segments of length $L_S = 8$ which are used to compute the disparity for the two contour points marked by black circles according to the contour segment based scheme given by Eq. (1.116). The solid contour has been extracted from the right image, the dashed contour from the left image of the stereo pair.

be obtained in a straightforward manner by

$$d_s = u_2^{(\tilde{j})}(v) - u_1^{(\tilde{i})}(v). \quad (1.113)$$

This disparity measure, however, may significantly change from one epipolar line to the next as the contours are often heavily influenced by pixel noise. Furthermore, d_s becomes inaccurate for nearly horizontal contour parts. An example of a disparity image obtained by using Eq. (1.113) is shown in Fig. 1.20.

To obtain a less noisy and more accurate disparity value, we define a contour segment based disparity measure which relies on an evaluation of L_S neighbouring epipolar lines. The two contour segments of length L_S are denoted by the sets $\{\mathbf{s}_1^{(i)}\}_{i=1,\dots,L_S}$ and $\{\mathbf{s}_2^{(j)}\}_{j=1,\dots,L_S}$ with

$$\mathbf{s}_1^{(i)} = \mathbf{c}_1^{(w_1^{(\tilde{i})} - L_S/2+i)} \quad \text{for } i = 1, \dots, L_S \quad (1.114)$$

$$\mathbf{s}_2^{(j)} = \mathbf{c}_2^{(w_2^{(\tilde{j})} - L_S/2+j)} \quad \text{for } j = 1, \dots, L_S \quad (1.115)$$

For an illustration of this contour segment extraction procedure, see Fig. 1.19. The contour segment based disparity d_c is then defined by

$$d_c = \frac{1}{L_S} \sum_{i=1}^{L_S} \left(\bar{u}_2^{(w_2^{(\tilde{j})} - L_S/2+i)} - \bar{u}_1^{(w_1^{(\tilde{i})} - L_S/2+i)} \right). \quad (1.116)$$

To avoid false correspondences or inaccurate disparities, the value of d_c computed according to Eq. (1.116) is only accepted if the four following conditions are fulfilled:

1. The i -th point of both contour segments is on the same epipolar line, respectively, for all values of i :

$$\bar{v}_1^{(w_1^{(\tilde{i})} - L_S/2+i)} = \bar{v}_2^{(w_2^{(\tilde{j})} - L_S/2+i)} \quad \text{for } i = 1, \dots, L_S. \quad (1.117)$$

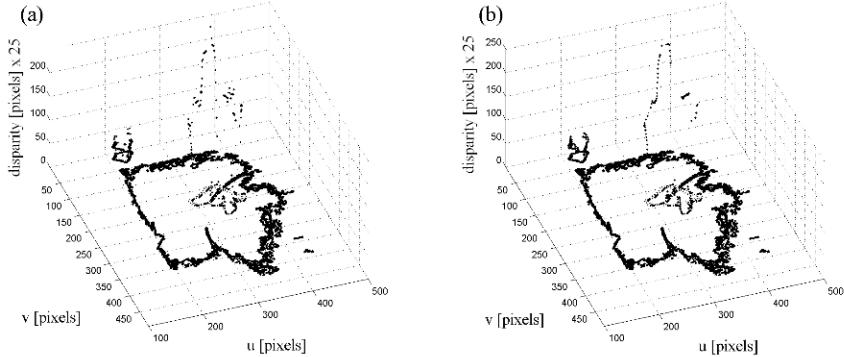


Fig. 1.20 (a) Disparity image obtained with binary threshold $\theta_0 = 7$ and $N_T = 3$ adaptive thresholds derived from the bounding box of each segmented blob, respectively. The contours are shown on the $d_s = 0$ plane of the three-dimensional plot. The disparities have been computed according to the contour point based scheme given by Eq. (1.113). (b) Disparity image of the same scene, computed according to the contour segment based scheme given by Eq. (1.116) with contour segment length $L_S = 8$, cross correlation threshold $\theta_{\text{corr}} = 0.7$, and absolute slope threshold $\theta_{\text{slope}} = 0.5$. It is obvious that the contour segment based method yields a significantly less noisy disparity image than the contour point based method.

2. The average absolute slopes of both contour segments are above a threshold θ_{slope} :

$$\frac{1}{L_S} \sum_{i=2}^{L_S} \left| \frac{\bar{v}_1^{(w_1^{(i)} - L_S/2+i)} - \bar{v}_1^{(w_1^{(i)} - L_S/2+i-1)}}{\bar{u}_1^{(w_1^{(i)} - L_S/2+i)} - \bar{u}_1^{(w_1^{(i)} - L_S/2+i-1)}} \right| > \theta_{\text{slope}} \quad (1.118)$$

and analogously for the left image.

3. The absolute cross-correlation coefficient of the contour segments (Heisele, 1998) exceeds a threshold θ_{corr} :

$$\left| \frac{1 - |\lambda_1 - \lambda_2|}{\lambda_1 + \lambda_2} \frac{\sum_{i=1}^{L_S} (\mathbf{s}_1^{(i)} - \langle \mathbf{s}_1 \rangle) \cdot (\mathbf{s}_2^{(i)} - \langle \mathbf{s}_2 \rangle)}{\sqrt{\sum_{i=1}^{L_S} |\mathbf{s}_1^{(i)} - \langle \mathbf{s}_1 \rangle|^2 \sum_{i=1}^{L_S} |\mathbf{s}_2^{(i)} - \langle \mathbf{s}_2 \rangle|^2}} \right| > \theta_{\text{corr}}, \quad (1.119)$$

where $\lambda_1 = \sum_{i=2}^{L_S} |\mathbf{s}_1^{(i)} - \mathbf{s}_1^{(i-1)}|$ and $\lambda_2 = \sum_{i=2}^{L_S} |\mathbf{s}_2^{(i)} - \mathbf{s}_2^{(i-1)}|$ denote the length of the contour segments and $\langle \mathbf{s}_1 \rangle^{(i)}$ and $\langle \mathbf{s}_2 \rangle^{(i)}$ the corresponding contour segment averages. This condition ensures that only contour segments which run approximately parallel are matched.

4. The intensity gradient in the difference image exceeds a threshold θ_{grad} :

$$\nabla_g \left(\bar{u}_1^{(w_1^{(\bar{i})})}, \bar{v}_1^{(w_1^{(\bar{i})})} \right) > \theta_{\text{grad}} \quad (1.120)$$

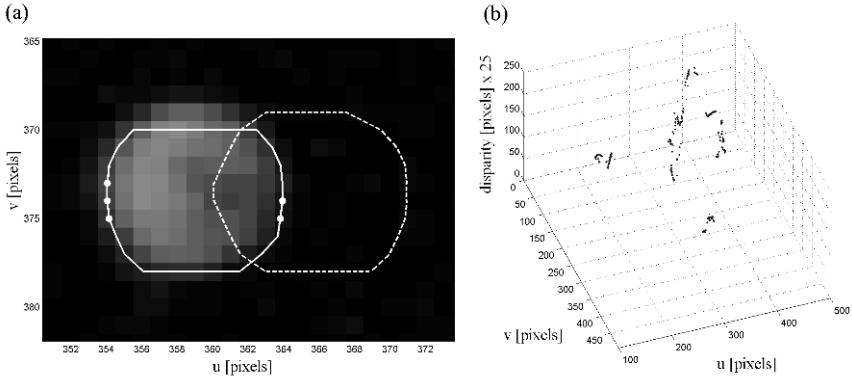


Fig. 1.21 (a) Contour segment based stereo analysis of a ball of 10 cm diameter, painted in the same colour as the floor. The solid contour is valid for the right image, the dashed contour for the left image of the stereo pair. Image points for which disparity values have been computed are denoted by white filled circles. (b) Disparity map of the same scene as shown in Fig. 1.20, obtained with the correlation based real-time stereo algorithm described in Franke and Joos (2000), which has been applied to the stereo pair of absolute difference images. Obviously, the disparity map is more noisy than the one obtained with the CBS algorithm (cf. Fig. 1.20b). This correlation based stereo vision algorithm, however, does not necessarily depend on reference images of the scene.

This condition ensures that only contour pixels are accepted the position of which is characterized by intensity changes, i.e. transitions from the background to an object, rather than by pixel noise in nearly uniform areas of the difference image. We assume that the surfaces of the objects that may enter the scene, e.g. persons, are arched rather than flat, such that an intensity difference between object and background is still perceivable due to shading even if the corresponding reflectivities are identical. This shading contrast determines the value of θ_{grad} .

The segmentation obtained with binary threshold θ_0 often does not yield contours around the objects in the scene, but around secondary structures such as shadows indirectly caused by the objects (Fig. 1.18b). Hence, for the bounding box of each blob segmented with binary threshold θ_0 , the described procedure is repeated with N_T different adaptive threshold values $\theta_a^{(q)}$, $q = 1, \dots, N_T$ which are derived from the intensity histogram of the corresponding bounding box. The values of the adaptive thresholds may be chosen to be equally spaced between θ_0 and the maximum possible grey value, or to lie within a suitable range around the minimum of the intensity histogram, which is in many cases a binary threshold value that is able to separate dark from bright image regions. Thus, disparity information is obtained from $N_T + 1$ isophotes in the pair of absolute difference images. These isophotes are determined by θ_0 and the chosen values of $\theta_a^{(q)}$.

Fig. 1.18a displays contours extracted from the stereo image pair with a relatively high binary threshold value of $\theta_0 = 100$. In Fig. 1.18b the contours extracted from the stereo image pair with $\theta_0 = 50$ are partially determined by shadows cast by the object rather than by the object itself (solid and dashed contour, respectively). For

these parts of the contours no disparity value is computed as the gradient at the corresponding positions in the image is too small to exceed the threshold θ_{grad} that depends on the minimum shading contrast; in our experiments we set $\theta_{\text{grad}} = 1$. Hence, the contour extraction procedure is repeated with an adaptive threshold derived from the intensity histogram of the bounding box of the solid contour, yielding the dotted and the dashed-dotted contour which are now fully characterized by the object boundary. Fig. 1.20 shows the resulting disparity map based on single contour points and based on contour segments with length $L_S = 8$. In the second example, we set the thresholds to the not very restrictive values $\theta_{\text{corr}} = 0.7$ and $\theta_{\text{slope}} = 0.5$, i.e. two corresponding contour segments may be oriented at an angle of up to 45° , provided that they are running at an angle of more than 30° with respect to the image rows. Note that changing these thresholds does not change the disparity value of a certain three-dimensional point but may only add it to or remove it from the disparity map. For this rather complex scene, the computation time of our C++ implementation of the CBS algorithm amounts to 80 ms on a 1.8 GHz Pentium IV processor. Fig. 1.20 illustrates that the contour segment based method yields a significantly less noisy disparity image than the contour point based method. Compared to the correlation based real-time stereo algorithm described by Franke and Joos (2000), which, on the other hand, does not make use of a reference image, the disparity values are more accurate and less noisy (Fig. 1.21b). In Fig. 1.21a the CBS algorithm is applied to a scene displaying a ball of 10 centimetres diameter and painted in the same colour as the floor, therefore appearing as a small circle of only eight pixels diameter in the stereo images. This very difficult object is detected by means of its shading contrast only.

1.5.2.3 Dense Stereo Vision Algorithms

Correlation-based blockmatching and feature-based stereo vision algorithms usually generate sparse depth maps, i.e. depth information is only derived for parts of the scene in which a sufficient amount of texture is observed, such that a sufficient amount of information is available to achieve a meaningful comparison between image windows along corresponding epipolar lines. In contrast, stereo vision algorithms which generate a depth value for each image pixel are termed dense stereo vision algorithms. An early dense stereo vision approach by Horn (1986) relies on the direct comparison of pixel intensities along corresponding epipolar lines instead of a comparison between image windows or local features. Since the intensity-based criterion alone leads to a highly ambiguous solution, a smooth depth map is assumed, i.e. there are no large differences between the disparity values assigned to neighbouring pixels. This line of thought leads to the minimisation of the error term

$$e = \sum_{u,v} \left[(\nabla^2 d(u,v))^2 + \lambda (I_1(u+d(u,v)/2, v) - I_2(u-d(u,v)/2, v))^2 \right], \quad (1.121)$$

where λ is a weight parameter and $\nabla^2 d(u, v)$ denotes the Laplacian of the disparity map—at this point, Horn (1986) states that omitting the square in the first term of Eq. (1.121) would lead to a strongly flattened solution. Based on the Euler equation of Eq. (1.121), a differential equation for $d(u, v)$ of fourth order in u and v is derived and solved numerically.

Another stereo vision algorithm that constructs dense depth maps is based on dynamic programming (Cox et al., 1996). It makes use of the ordering constraint which requires that for opaque surfaces the order of neighbouring point correspondences on two corresponding epipolar lines is always preserved. Cox et al. (1996) assume that if two pixels correspond to the same scene point, the distribution of the intensity difference with respect to all mutually corresponding pixels is Gaussian. A maximum likelihood criterion then defines an overall cost function which is minimised by the dynamic programming algorithm. While in the approach suggested by Cox et al. (1996) each epipolar line is processed independently, the graph cut or maximum flow method optimises the solution globally (Roy and Cox, 1998). Instead of the ordering constraint, a more general local coherence constraint is assumed which claims that disparities tend to be locally similar. The correspondence problem is then formulated as a maximum-flow problem in a graph.

A survey about dense stereo methods and an evaluation framework essentially based on synthetic images to assess their performance is provided by Scharstein and Szeliski (2002). Van der Mark and Gavrila (2006) examine dense stereo algorithms with respect to their suitability for real-time intelligent vehicle applications. Based on realistically looking synthetic image data and real image data from urban traffic scenes, they show that algorithms which involve global search optimisation are more affected by the variability of the encountered scene conditions than approaches relying on simpler selection criteria to establish point correspondences. Furthermore, it is demonstrated by van der Mark and Gavrila (2006) that the multiple window framework introduced by Hirschmüller et al. (2002) in the context of correlation-based blockmatching stereo vision, using local matching and a left-right consistency check, shows the best performance of all examined algorithms when applied in the context of real-time dense stereo.

A general drawback of dense stereo algorithms is the fact that the established depth values tend to be inaccurate for parts of the surface that do not show any surface texture at all, or for corresponding parts of the stereo image pair which do not display a similar structure. The latter behaviour may e.g. occur as a consequence of specular reflectance properties leading to a different appearance of the respective surface part in the stereo images. In such cases of missing or contradictory texture information, dense stereo algorithms usually interpolate the surface across the ambiguous image parts, leading to an inaccurate three-dimensional reconstruction result for the corresponding region. This problem is addressed explicitly by Hirschmüller (2006), who proposes an intensity consistent disparity selection scheme termed semi-global matching for dealing with untextured surface parts and suggests a discontinuity preserving interpolation approach for filling holes in the disparity map.

1.5.2.4 Spacetime Stereo Vision and Scene Flow Algorithms

General Overview

An extension of the classical pairwise frame-by-frame approach to stereo vision towards a spatio-temporal analysis of a sequence of image pairs has been introduced quite recently as spacetime stereo by several researchers (Zhang et al., 2003; Davis et al., 2005). Both studies present a framework that aims for a unification of stereo vision with active depth from triangulation methods such as laser scanning and coded structured light by generalising a spatial similarity measure according to Eq. (1.109) to the spatio-temporal domain. Zhang et al. (2003) incorporate temporal appearance variation to improve stereo matching and generate more accurate and reliable depth maps. They show that assuming a disparity value $d(u_c, v_c, t_c)$ which is constant throughout a spatio-temporal image window centred at (u_c, v_c, t_c) can only be assumed for a fronto-parallel surface. For static but oblique surfaces, a linear expansion for the disparity is introduced by Zhang et al. (2003), corresponding to

$$d(u, v, t) = d(u_c, v_c, t_c) + \frac{\partial d}{\partial u}(u - u_c) + \frac{\partial d}{\partial v}(v - v_c) + \dots \quad (1.122)$$

For moving scenes, this representation is extended according to

$$d(u, v, t) = d(u_c, v_c, t_c) + \frac{\partial d}{\partial u}(u - u_c) + \frac{\partial d}{\partial v}(v - v_c) + \frac{\partial d}{\partial t}(t - t_c) + \dots \quad (1.123)$$

since the disparity may change over time as a result of a radial velocity of the object. These expressions for the disparity are inserted into a similarity measure defined by Eq. (1.109). Dynamic programming followed by Lucas-Kanade flow (Lucas and Kanade, 1981) is utilised to establish point correspondences and estimate the disparities as well as their spatial and temporal first derivatives $\partial d / \partial u$, $\partial d / \partial v$, and $\partial d / \partial t$. A significant improvement over classical stereo analysis is achieved by Zhang et al. (2003) when a static scene is illuminated by a temporally variable illumination pattern which is not necessarily strictly controlled. In these cases, their three-dimensional reconstruction reveals details comparable to those obtained with a laser scanner. For moving objects under more natural lighting conditions, their spacetime stereo approach performs approximately as well as classical stereo vision with larger windows.

The spacetime stereo framework described by Davis et al. (2005) is fairly similar to the one presented by Zhang et al. (2003). However, the spatial and temporal derivatives of the disparity are not estimated. Davis et al. (2005) concentrate on the determination of the optimal spatio-temporal size of the matching window for static scenes and scenes with moving objects. For static scenes illuminated with temporally variable but not strictly controlled structured light patterns, they arrive at the conclusion that after acquiring only a short sequence of about 25 images, it is no longer necessary to use spatially extended matching windows, since a purely temporal matching vector turns out to yield the highest reconstruction accuracy. Scenes

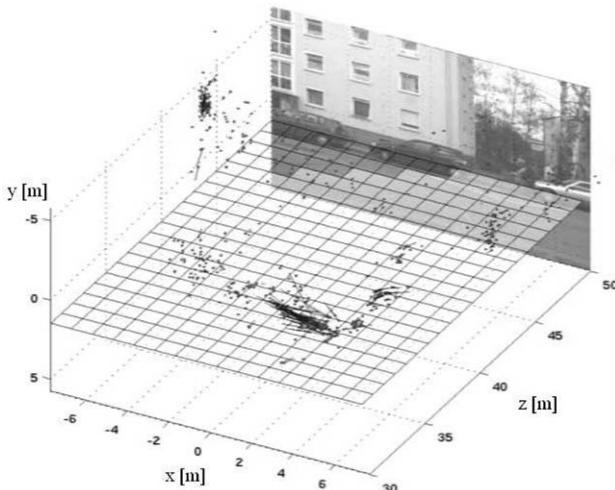


Fig. 1.22 Motion-attributed three-dimensional point cloud obtained with the 6D vision method (Franke et al., 2005) for a traffic scene.

with linearly moving and with rotating objects are illuminated with light patterns varying at high frequency, generated with an uncalibrated LCD projector. Davis et al. (2005) find that the optimum spatial and temporal window size is smaller for the scene with a linearly moving object than for the scene displaying a rotating object. The optimum temporal extension decreases with increasing speed of the motion. For illumination with a temporally variable light pattern, Davis et al. (2005) conclude that for scenes with fast motion, purely spatial stereo analysis is favourable, while static scenes should be analysed based on the purely temporal variations of the pixel brightnesses.

For obstacle avoidance in the context of driver assistance and mobile robotic systems, which require an extraction of depth information and the robust and fast detection of moving objects, Franke et al. (2005) introduce a framework termed 6D vision, addressing the fusion of stereo and optical flow information. The three-dimensional position and motion of each individual extracted depth point are estimated simultaneously based on Kalman filters. A typical motion-attributed three-dimensional point cloud obtained with this method for a traffic scene is shown in Fig. 1.22, where the motion vectors are indicated by lines. The bicyclist and the pedestrians near the right image border appear as clusters of three-dimensional points moving to the left and to the right, respectively.

Vedula and Baker (2005) introduce the concept of scene flow, which contains the three-dimensional positions of the scene points along with their three-dimensional displacement field. The classical two-dimensional optical flow thus corresponds to a projection of the scene flow into the image plane. Therefore, the 6D vision method by Franke et al. (2005) yields sparse scene flow information. Huguet and Devernay (2007) determine dense scene flow by coupling the estimation of optical flow in both

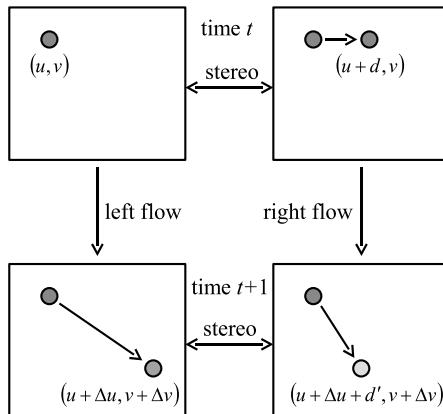


Fig. 1.23 Determination of scene flow information according to Huguet and Devernay (2007).

images with the computation of dense stereo (cf. Fig. 1.23). A variational framework yields a system of partial differential equations that simultaneously determine the disparities and the optical flow field, which allows to handle discontinuities of the disparities and the three-dimensional displacement field as well as occlusions. A numerical solution is obtained based on a multiresolution algorithm. While previous variational approaches e.g. by Pons et al. (2005) estimate the three-dimensional points and the scene flow field separately, they are computed simultaneously in a single adjustment stage by Huguet and Devernay (2007).

Local Intensity Modelling

In contrast to the previously described spacetime stereo approaches, which exploit a pixel-based similarity measure, the spacetime stereo algorithm introduced by Schmidt et al. (2007) relies on the fit of a parametric model to the spatio-temporal neighbourhood of each interest pixel in an image sequence. This method yields a cloud of three-dimensional points carrying additional information about the motion properties of the corresponding scene part. Hence, point correspondences may be resolved which would remain ambiguous without taking into account the temporal domain, thus reducing the rate of false correspondences. The additional motion cues may be used to support optional subsequent processing steps dealing with three-dimensional scene segmentation and object tracking (cf. Section 1.6.3).

Like in most stereo vision approaches that establish correspondences between small image regions, the first processing step of our algorithm consists of determining interest pixels in order to select the image regions for which three-dimensional information is computed in a later step. The interest operator may e.g. consist of the local grey value variance or of an edge detector. In this case, interest pixels correspond to image regions with small-scale intensity variations, implying the presence of image structures such as object boundaries upon which a correspondence analysis

can be based. A further possible interest operator in the context of spacetime stereo is a spatio-temporal feature detector. As an example, Wöhler and Anlauf (2001) define spatio-temporal features, which are optimised for the classification of certain object types in image sequences, by the weights of the spatio-temporal receptive fields of a time-delay neural network architecture. This kind of interest operator may extract spatio-temporal motion patterns, e.g. top-down moving corners.

The image sequence is defined in $(uvtg)$ space, where u and v denote the pixel coordinates, t the time coordinate, and g the pixel grey value. To the local spatio-temporal neighbourhood of each interest pixel a parameterised function $h(\mathbf{P}, u, v, t)$ is adapted, where the vector \mathbf{P} denotes the parameters of the function. The interest operator preferentially extracts image regions along the boundaries of objects in the scene.

Ideally, an object boundary is described by an abrupt intensity change. In real images, however, one does not observe such discontinuities since they are blurred by the point spread function of the optical system. Therefore, we model the intensity change at an object boundary by a “soft” function of sigmoidal shape like the hyperbolic tangent (cf. Section 1.4.8.2). Without loss of generality we will assume here that the epipolar lines are parallel to the image rows. As we cannot assume the image regions inside and outside the object to be of uniform intensity, we model the intensity distribution around an interest pixel by a combined sigmoid-polynomial approach:

$$h(\mathbf{P}, u, v, t) = p_1(v, t) \tanh[p_2(v, t)u + p_3(v, t)] + p_4(v, t). \quad (1.124)$$

The terms $p_1(v, t)$, $p_2(v, t)$, $p_3(v, t)$, and $p_4(v, t)$ denote polynomials in v and t . Here we assume that the stereo camera system is calibrated (Krüger et al., 2004) and the stereo image pairs are rectified to standard geometry (cf. Section 1.5).

The polynomial $p_1(v, t)$ describes the amplitude and $p_2(v, t)$ the steepness of the sigmoid function, which both depend on the image row v , while $p_3(v, t)$ accounts for the row-dependent position of the model boundary. The value of $p_2(v, t)$ is closely related to the sign of the intensity gradient and to how well it is focused, where large values describe sharp edges and small values blurred edges. The polynomial $p_4(v, t)$ is a spatially variable offset which models local intensity variations across the object and in the background, e.g. allowing the model to adapt to cluttered background. All described properties are assumed to be time-dependent. An interest pixel is rejected if the residual of the fit exceeds a given threshold.

The parametric model according to Eq. (1.124) in its general form requires that a nonlinear least-mean-squares optimisation procedure is applied to each interest pixel, which may lead to a prohibitively high computational cost of the method. It is possible, however, to transform the nonlinear optimisation problem into a linear problem by making the following simplifying assumptions:

1. The offset $p_4(v, t)$ is proportional to the average pixel intensity \bar{I} of the spatio-temporal matching window, i.e. $p_4(v, t) = w\bar{I}$.

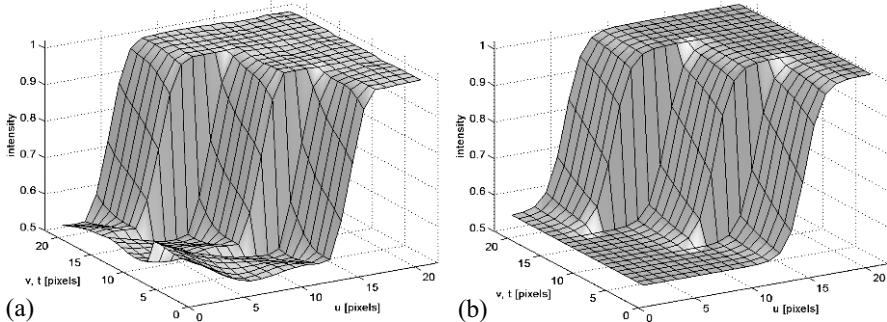


Fig. 1.24 (a) Spatio-temporal intensity profile of a moving object boundary, measured over a time interval of 3 time steps. The size of the spatio-temporal matching window is $21 \times 7 \times 3$ pixels. For visualisation, the (v, t) axis is divided such that each time step comprises an interval of 7 pixels. (b) Modelling result according to Eq. (1.125), with $p_2(v, t)$ of first order and $p_3(v, t)$ of second order in v and t .

2. The amplitude $p_1(v, t)$ of the sigmoid is proportional to the standard deviation σ_I of the pixel intensities in the spatio-temporal matching window with $p_1(v, t) = k\sigma_I$.

These simplifications yield the model equation

$$p_2(v, t)u + p_3(v, t) = \operatorname{artanh} \left[\frac{I(u, v, t) - w\bar{I}}{k\sigma_I} \right] \equiv \tilde{I}(u, v, t), \quad (1.125)$$

where the model parameters, i.e. the coefficients of the polynomials $p_2(v, t)$ and $p_3(v, t)$, can be determined by a linear fit to the transformed image data $\tilde{I}(u, v, t)$. Real-time processing speed is achieved by implementing the artanh function as a look-up table.

Pixels with $|[I(u, v, t) - w\bar{I}] / [k\sigma_I]| > \theta$ are excluded from the fit, where θ is a user-defined threshold with $\theta < 1$, since arguments of the artanh function close to 1 would lead to a strong amplification of noise in the original pixel intensities. The factors k and w are further user-defined parameters of the algorithm. A typical spatio-temporally local modelling result for a moving object boundary is shown in Fig. 1.24.

Eq. (1.124) allows for a direct computation of the location u_e of the epipolar intersection, i.e. the position of the intensity change at subpixel accuracy in u direction. This value is essential for a precise determination of disparity. The value of u_e is defined by the maximum gradient of the intensity profile in u direction, corresponding to the root of the hyperbolic tangent. This condition yields

$$u_e(v, t) = -p_3(v, t)/p_2(v, t), \quad (1.126)$$

where the value $u_e(v_c, t_c)$ with v_c and t_c denoting the centre of the matching window is used for the determination of disparity.

The direction δ of the intensity gradient at the location of the interest pixel, representing a feature that will be used for correspondence analysis later on, is given by

$$\delta = \left. \frac{\partial u_e}{\partial v} \right|_{v_c, t_c}. \quad (1.127)$$

The velocity μ of the intensity gradient along the epipolar line corresponds to the temporal derivative

$$\mu = \left. \frac{\partial u_e}{\partial t} \right|_{v_c, t_c} \quad (1.128)$$

of the location of the epipolar transection. Such explicit motion information is not revealed by the purely correlation-based spacetime approach described by Davis et al. (2005). The approach described by Franke et al. (2005) yields motion information for each three-dimensional point but requires a separate processing stage for individually tracking the corresponding positions and velocities.

For the purpose of correspondence analysis, a similarity measure between two interest pixels located on the same epipolar line v is determined based on the functions $h(\mathbf{P}_{\text{left}}, u, v, t)$ and $h(\mathbf{P}_{\text{right}}, u, v, t)$ fitted in the left and the right image to the spatio-temporal matching windows of the interest pixels, respectively, where the obtained function parameters are denoted by the vectors \mathbf{P}_{left} and $\mathbf{P}_{\text{right}}$.

A suitable similarity measure between two interest pixels analysed in the left and in the right image, respectively, corresponds to a weighted distance of the two interest pixels in the space spanned by the parameters according to

$$S_{\text{dist}} = \sum_i m_i \left(P_{\text{left}}^{(i)} - P_{\text{right}}^{(i)} \right)^2. \quad (1.129)$$

To account empirically for the significance of the individual function parameters, the weights m_i in Eq. (1.129) are chosen according to the distribution of the corresponding parameters $P^{(i)}$ over a representative set of interest pixels. Generally spoken, a broad distribution of $P^{(i)}$ favours a small weight m_i and vice versa.

Alternatively, the similarity measures well known from classical correlation-based stereo vision approaches (Franke and Joos, 2000), such as the sum of squared differences (SSD), the sum of absolute differences (SAD), or the cross-correlation coefficient, can be adapted to our algorithm. This is achieved by comparing the fitted functions $h(\mathbf{P}_{\text{left}}, u, v, t)$ and $h(\mathbf{P}_{\text{right}}, u, v, t)$ rather than the pixel intensities themselves (Davis et al., 2005). As an example, the SSD similarity measure then reads

$$S_{\text{SSD}} = \int \left[h(\mathbf{P}_l, u - u_e^{\text{left}}(v_c, t_c), v, t) - h(\mathbf{P}_r, u - u_e^{\text{right}}(v_c, t_c), v, t) \right]^2 du dv dt, \quad (1.130)$$

where u , v , and t traverse the spatio-temporal matching windows of the left and the right interest pixel, respectively. Analogous expressions are obtained for the SAD or cross-correlation similarity measure.

Once a correspondence between two interest pixels on the same epipolar line has been established by searching for the best similarity measure e.g. according to

Eqs. (1.129) or (1.130), the disparity d corresponds to the difference between the epipolar transections u_e^{left} and u_e^{right} computed according to Eq. (1.126) for the left and the right interest pixel, respectively:

$$d = u_e^{\text{left}}(v_c, t_c) - u_e^{\text{right}}(v_c, t_c). \quad (1.131)$$

To increase the accuracy of the determined disparity values, it is advantageous to establish the correspondences based on the spacetime approach by searching for the minimum value of S_{dist} or S_{SSD} along the epipolar line but to compute the corresponding disparities without utilising temporal information. This prevents the disparity value from becoming inaccurate when the true motion behaviour is not closely approximated by the model function.

Given the optical and geometrical parameters of the camera system, the velocity component $\bar{\mu}$ parallel to the epipolar lines (in pixels per time step) amounts to

$$\bar{\mu} = \frac{1}{2} (\mu^{\text{left}} + \mu^{\text{right}}). \quad (1.132)$$

In metric units, the epipolar velocity $U = \partial x / \partial t$ is given by

$$U = \frac{\partial x}{\partial t} = \|\mathbf{t}\| \frac{\bar{\mu} d - \frac{1}{2} (u_e^{\text{left}}(v_c, t_c) + u_e^{\text{right}}(v_c, t_c)) \frac{\partial d}{\partial t}}{d^2}. \quad (1.133)$$

The vertical velocity component $V = \partial y / \partial t$ cannot be inferred pointwise from the spacetime stereo data due to the aperture problem. The velocity component $\partial z / \partial t$ along the depth axis depends on the first temporal derivative of the disparity, which is obtained according to

$$\frac{\partial d}{\partial t} = \mu^{\text{left}} - \mu^{\text{right}} \quad (1.134)$$

(cf. Eqs. (1.128) and (1.131)). Inserting Eq. (1.134) into Eq. (1.108) yields for the velocity W along the z axis

$$W = \frac{\partial z}{\partial t} = - \frac{\|\mathbf{t}\| b_0 (\mu^{\text{left}} - \mu^{\text{right}})}{d_p d^2}. \quad (1.135)$$

Note, however, that small relative errors of μ^{left} and μ^{right} may lead to large relative errors of $\partial d / \partial t$ and $\partial z / \partial t$.

At this point we present first three-dimensional reconstruction results in order to illustrate how the described spacetime stereo approach can be used for the enhancement of correspondence analysis and for the low-level extraction of motion features. More detailed experimental results in the context of three-dimensional scene segmentation and object tracking will be described in Chapter 6. A Sobel edge filter of 7×7 pixels size is used as an interest operator. The matching window is chosen to be larger along the epipolar lines than in the perpendicular direction such that an object boundary moving at a vertical velocity of up to 10 pixels per time step is apparent at all three regarded moments in time. The polynomials $p_2(v, t)$ and $p_3(v, t)$

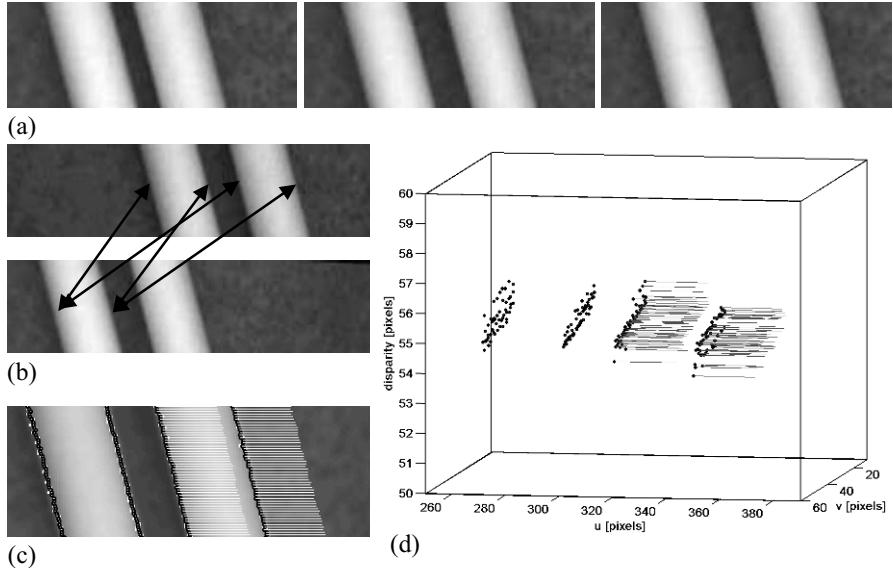


Fig. 1.25 (a) Three images of the “moving bar” sequence. (b) Ambiguities encountered during correspondence analysis based on a stereo pair of single images. (c) Interest pixels (white) and image locations for which stereo correspondences have been established (black). The horizontal velocity μ of each three-dimensional point in pixels per time step is indicated by the length of the associated line (four times exaggerated). (d) Oblique view into the point cloud.

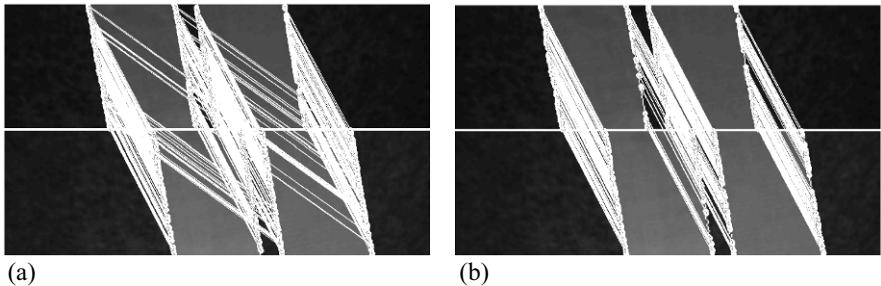


Fig. 1.26 Established correspondences in the “moving bar” scenario (cf. Fig. 1.25). (a) Correspondences established with a window size of $21 \times 7 \times 1$ pixels, thus neglecting motion information. (b) Correspondences established with the full spacetime stereo approach, using a window size of $21 \times 7 \times 3$ pixels.

in Eq. (1.125) are chosen to be of first and second order, respectively, leading to a left hand side of Eq. (1.125) which corresponds to an incomplete second-order polynomial in u , v , and t . Due to the limited extension of the matching window we assume that higher-order information about the shape of the object boundary or its motion behaviour cannot be reliably extracted. For each interest pixel, the model fit is computed for several combinations of θ , k , and w , choosing $\theta \in [0.4 \dots 0.9]$,

$k \in [1.0 \dots 1.5]$, and $w \in [0.8 \dots 1.2]$ at stepsizes of 0.1, respectively. The parameter vector \mathbf{P} obtained based on the configuration yielding the lowest residual is used for correspondence analysis.

The sequence shown in Fig. 1.25a shows two identical objects of cylindrical shape, situated at identical distance and appearing as two bars. One object is stationary, while the other one is moving towards the right hand side. The matching window to which a model function $h(\mathbf{P}, u, v, t)$ is fitted is of size $21 \times 7 \times 3$ pixels. The left object is stationary, while the right object is moving horizontally. Fig. 1.25b illustrates that in a stereo pair of single images, the mutual assignment of the object boundaries and thus the resulting disparities are ambiguous. It strongly depends on the assignment strategy between which object boundaries correspondences are established, if they are established correctly, or which possible correspondences are omitted. The result of spatio-temporal modelling is shown in Fig. 1.25c, where the lines associated with the three-dimensional points denote their horizontal velocity. An oblique view into the point cloud is shown in Fig. 1.25d.

For the similarity measure S_{dist} according to Eq. (1.129), instead of the fitted function parameters denoted by \mathbf{P} the physically more meaningful local direction δ of the intensity gradient and its horizontal velocity μ were used (cf. Eqs. (1.127) and (1.128)). Identical weights were chosen for these inferred parameters since both are of the order 1 in the example of Fig. 1.25 and display distributions of similar width.

Utilising the parameter distance similarity measure according to Eq. (1.129) and the SSD similarity measure according to Eq. (1.130), selecting the correspondence with the best similarity measure along each epipolar line, yields the same result in the scenario of Fig. 1.25. The disparity values in Fig. 1.25d indicate that the object boundaries are all situated at similar distances.

In Fig. 1.26, the correspondences established with the SSD similarity measure are shown for a “moving bar” scenario similar to Fig. 1.25. This example demonstrates that the spacetime approach is able to suppress the incorrect correspondences that tend to occur when only a single stereo pair is analysed as a result of the strong similarity of the two bars. Stereo algorithms that do not take into account the temporal domain would require additional ad-hoc constraints about the scene to reject such outliers, such as smoothness of the disparity map (Horn, 1986) or the assumption of locally similar disparities (Roy and Cox, 1998).

The objects in the scene shown in Fig. 1.26 are fronto-parallel, i.e. all points on the object boundaries have the same distance to the image plane. Hence, the standard deviation of the disparity values immediately yields their relative accuracy. It amounts to 0.1 pixels. Supposedly, however, this high relative accuracy is due to the fact that the object boundaries in this image sequence are well represented by the sigmoid-shaped model 1.124, and it will turn out in Chapter 6 that the relative accuracy is somewhat lower in more natural scenes where the sigmoid model is not always fully appropriate. In order to determine the absolute accuracy of the epipolar velocity μ computed according to Eq. (1.128), the horizontal motion of the object boundary is derived from the shift of the locations of the corresponding interesting pixels (which are given at pixel accuracy), averaged over 20 time steps, for the

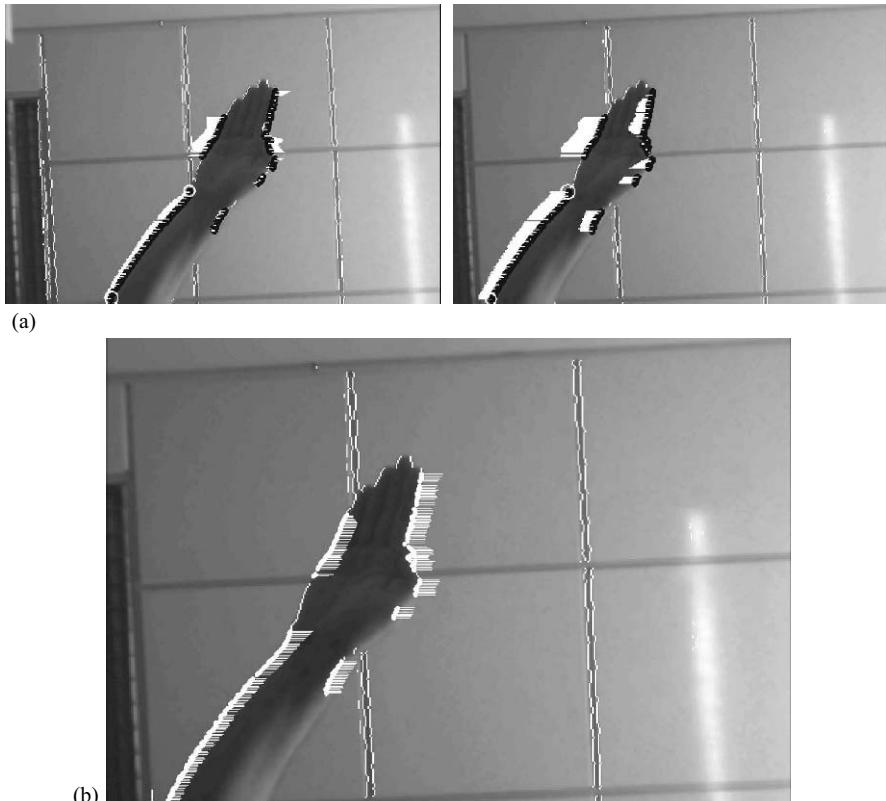


Fig. 1.27 Determination of the velocity along the z axis for an image sequence showing a hand moving towards the camera. (a) A typical image pair of the sequence. White lines indicate epipolar velocities. White circles indicate reference points for analysis of accuracy. (b) Temporal derivative $\partial d / \partial t$ determined according to Eq. (1.135) for the right stereo image. White lines indicate the value of $\partial d / \partial t$.

moving bar in the image sequence depicted in Fig. 1.26. This procedure yields an epipolar velocity of 8.34 pixels per time step. The spacetime stereo approach yields a very similar value of 8.39 pixels per time step, which differs by only 0.6 percent from the value derived from the motion of the interest pixels.

To illustrate the determination of the object velocity along the z axis according to Eqs. (1.134) and (1.135) we regard an image sequence showing a hand moving towards the camera as shown in Fig. 1.27 (Gövert, 2006). In Fig. 1.27a two stereo images of the sequence are shown along with the determined epipolar velocities. The velocity differences between the two images are clearly apparent. The white lines in Fig. 1.27b denote the temporal derivative $\partial d / \partial t$ of the disparity, computed according to Eq. (1.135). All determined values of $\partial d / \partial t$ are positive as object distance decreases while disparity increases. The velocity along the z axis is larger

for the hand than for the forearm since hand and forearm rotate around a point close to the elbow.

To examine the absolute accuracy of the temporal derivative $\partial d/\partial t$ of the disparity, computed according to Eq. (1.134), the disparities are derived from the horizontal positions of the corresponding interest pixels at time steps $t = -1$, $t = 0$, and $t = +1$ for the two reference locations marked as white circles in Fig. 1.27a. The average increase of disparity per time step is computed accordingly. For the upper reference point, the disparity increases by 3.5 pixels per time step, while a value of 2.0 pixels per time step is obtained for the lower reference point. The values for $\partial d/\partial t$ determined according to Eq. (1.134) amount to 3.89 and 1.94 pixels per time step, respectively, which is in reasonable accordance. Like the local spatio-temporal intensity modelling approach described in this section, the space-time stereo method by Zhang et al. (2003) estimates the temporal derivative $\partial d/\partial t$ of the disparity for each established point correspondence. However, no quantitative evaluation but merely a qualitative discussion is given in their study. Hence, a direct comparison between the accuracy of $\partial d/\partial t$ as determined with the method outlined in this section and the values obtained with the method by Zhang et al. (2003) is not possible. The spacetime stereo approach by Davis et al. (2005) does not take into account the temporal derivative of the disparity.

An important advantage of the spacetime stereo method described in this section is the fact that no explicit correspondences need to be established over time. Furthermore, the motion parameters are available nearly instantaneously (after acquisition of three images in the presented examples) since no tracking stage is involved—tracking systems usually require a certain settlement phase after initialisation before the estimated motion parameters become reliable. In Chapter 6 it is demonstrated in more detail that spacetime stereo is a useful technique in the context of three-dimensional scene segmentation and object tracking. Especially in the presence of several objects in the scene which all move in a different manner, it is often difficult to assign parts of the three-dimensional point cloud to specific objects when only spatial information is available. Adding motion cues to the point cloud introduces new information that may allow to distinguish unambiguously between the objects in the scene even when they come close to each other or mutually overlap.

1.6 Three-dimensional Pose Estimation and Segmentation Methods

In the previous sections it has been described how a three-dimensional reconstruction of a scene can be obtained from point correspondences between images, and how this problem is related with the problem of determining the intrinsic and extrinsic camera parameters. In this context, the reconstruction result is always obtained as a cloud of three-dimensional points. This point cloud contains information about the presence of matter at a certain position in space. However, no information is available about the presence or even the position and orientation of objects in the

scene. Hence, in this section we regard the problem of pose estimation of objects in the scene based on single or multiple calibrated images, where it is assumed that a (more or less detailed) model of the object is available.

For rigid objects, pose estimation refers to the estimation of the six degrees of freedom of the object, i.e. three rotation angles and three translation components (cf. Section 1.6.1). For objects with a limited number of internal degrees of freedom (so-called articulated objects) and objects with an infinite number of internal degrees of freedom (so-called non-rigid objects) pose estimation may correspond to a much higher-dimensional optimisation problem (cf. Section 1.6.2). Here, the three-dimensional object detection of objects is based directly on the acquired images of the scene or suitable features derived from them. A class of complementary approaches deals with the analysis of point clouds obtained e.g. by stereo image analysis with respect to their three-dimensional structure and motion behaviour, aiming for a segmentation into physically meaningful domains representing individual objects or object parts. Such methods are regarded in Section 1.6.3.

1.6.1 Pose Estimation of Rigid Objects

To estimate the pose of a rigid object, a set of three-dimensional scene points with coordinates given by a geometry model of the object is brought into correspondence with points in the image such that the rotation angles and the translation of the object with respect to the camera can be determined by minimising a suitable error measure, e.g. the backprojection error of the model points. Accordingly, pose estimation of rigid objects is equivalent to the problem of determining the extrinsic camera parameters (cf. Section 1.4). This is an important insight since all approaches described in Section 1.4 in the context of camera calibration can in principle be applied to the problem of pose estimation as well. Hence, an important issue in the context of pose estimation is to establish reliable correspondences between model points and points in the image. This section provides an overview of pose estimation techniques for rigid objects based on point features but also on edges or lines, and then regards in more detail the edge-based hierarchical template matching approach to pose estimation introduced by von Bank et al. (2003).

1.6.1.1 General Overview

Pose Estimation Methods Based on Explicit Feature Matching

An early survey of pose estimation methods based on the approach of bundle adjustment is provided by Szczepanski (1958). In the field of computer vision, a first description of the pose estimation problem is given by Fischler and Bolles (1981). Haralick et al. (1989) introduce several classes of pose estimation problems: 2D-2D pose estimation determines the pose of a two-dimensional object from an image,

while 3D-3D pose estimation deals with three-dimensional data points e.g. acquired with a range sensor from which the pose of a three-dimensional object is inferred. The term 2D-3D pose estimation denotes the estimation of the six pose parameters from a single two-dimensional view of the object, given a geometry model, or an estimation of the relative camera orientation from two views of the object without making use of a geometry model of the object. The solutions proposed by Haralick et al. (1989) are based on point correspondences between the images and the object model. They rely on a minimisation of the distances in three-dimensional space between the model points and the observed rays to the corresponding scene points. A proof of convergence is provided. A linear method based on singular value decomposition is proposed to determine the rotation angles, and it is demonstrated that robust M-estimation instead of least-mean-squares minimisation may increase the performance of the optimisation procedure, as a reasonable pose estimation result is still obtained when the rate of outliers is as high as 50 percent.

Image features beyond points are examined by Lowe (1987). Groupings and structures in the image which are likely to be invariant over a wide range of viewpoints are formed by perceptual organisation e.g. with respect to parallelism, collinearity, and end-point proximity. The search space during model based matching is reduced based on a probabilistic ranking method. The projections of three-dimensional models are brought into direct correspondence with features such as line segments in the image. A model matching step yields the unknown viewpoint parameters. An important edge-based approach to 2D-3D pose estimation is proposed by Lowe (1991). The object model is composed of polygonal surface patches as local approximations to the true surface shape. Object contours are obtained by projecting the visible outline into the image. Edge segments are extracted from the greyscale image with the Canny edge detector (Canny, 1986). The error function is given by the sum of the perpendicular distances in the image between the end points of the extracted edge segments and the closest projected object contour line. The pose parameters, including internal degrees of freedom of the object, are obtained by least-mean-square minimisation of the error function with the Gauß-Newton method or alternatively with the Levenberg-Marquardt algorithm (Press et al., 1992), where the latter generally displays a more robust convergence behaviour. A similar 2D-3D pose estimation approach based on point and line correspondences is described by Phong et al. (1996). They devise a quadratic error function by representing rotation and translation with a dual number quaternion. The determination of the pose parameters is performed with a trust-region optimisation method, which yields a superior performance compared to the Newton and the Levenberg-Marquardt method. In the framework developed by Grebner (1994), 2D-3D pose estimation of industrial parts is performed based on edges and corner points, where the parameter search, corresponding to the minimisation of an appropriately chosen cost function, is performed using the A* algorithm (cf. e.g. Sagerer (1985) for an overview).

A further important approach to 2D-3D pose estimation, introduced by Lamdan and Wolfson (1988), is geometric hashing. In a first step a model library is constructed by selecting n -tuples of model features lying in planar sections of a three-dimensional object model and using them as a basis for a coordinate system

into which the remaining model features are transformed as well. For point features, a basis is given by a point triple. A grid of cells, the so-called hash table, is defined, into which for each occupied cell the points are stored together with their corresponding base coordinate system. The model features are thus stored in the hash table in all possible coordinate representations. The matching procedure is performed by selecting a n -tuple of observed features and transforming all other features into the corresponding coordinate system. For each cell in the hash table which is sufficiently close to these transformed points an accumulator is incremented. One accumulator is defined per coordinate system basis, and all bases that exceed a given number of votes are accepted. The coordinate system corresponding to an accordingly selected accumulator cell allows to determine the object pose. This method is favourably used for estimating the pose of rigid objects as it relies on coordinate systems defined by subsets of fixed model and image features and does not explicitly take into account the occlusion between different parts of the objects.

Appearance-based Pose Estimation Methods

The pose estimation methods mentioned above have in common that they rely on explicitly established correspondences between model features and image features such as points and lines, which may be a difficult and error-prone task especially for scenes with complex background structure. A different class of methods are appearance-based approaches, which compare directly the observed image with the appearance of the object at different poses without explicitly establishing correspondences between model features and parts of the image.

An important appearance-based pose estimation method is the generalised Hough transform, also known as feature pose map (Westling and Davis, 1996; Blasko and Fua, 2001). A feature pose map is a table of the same size as the size of the image, containing one layer per feature type. Each cell in this three-dimensional table contains a list of the poses that create a feature of the appropriate type at the corresponding pixel position. For analysis of an image all features are computed, and an accumulator is incremented for each pose in a cell for which a feature is present. The accumulator architecture is inspired by the concept of the Hough transform (Jähne, 2005), which is a standard algorithm for the extraction of structural elements such as lines, curve segments, or circles from an image. The poses that belong to accumulators with a sufficiently large number of votes are accepted as recognition results.

In the context of industrial quality inspection, an appearance-based 2D-3D pose estimation approach is proposed by von Bank et al. (2003) which relies on the Chamfer matching technique to force convergence of a hierarchical template matching approach. To represent the object appropriately, a large number of two-dimensional edge templates covering the desired range of poses is generated based on a three-dimensional object model. This method is described in detail in Section 1.6.1.2.

A probabilistic approach to simultaneous pose estimation and object recognition is proposed by Niemann and Hornegger (2001). However, they only regard the problem of 2D-2D pose estimation. They treat object recognition as a problem of statistical learning theory and localisation as a parameter estimation problem. As an object model, a joint probability density function of the complete image showing an object of a certain class is introduced, relying on a mixture of Gaussians. The parameters of this multimodal distribution are estimated based on empirical data, where the structure of the object may be defined by the number of the mixture components while the parameters are given by the mixture parameters. The number of mixture components is determined by vector quantisation, while the parameters of the distribution are obtained with the expectation-maximisation algorithm. In this probabilistic framework, the localisation of objects corresponds to a maximum-likelihood estimation problem. In order to speed up the optimisation procedure, marginal distributions are used to decompose the search space into appropriate regions.

Pose Estimation Methods Concentrating on Specific Problem Aspects

Other contributions regard specific aspects of the problem of pose estimation. As an example, the problem of 2D-3D pose estimation of three-dimensional free-form surface models is discussed by Rosenhahn et al. (2003). The object is modelled as a two-parametric surface model represented by Fourier descriptors, and the pose estimation is solved in the framework of conformal geometric algebra. Low-pass surface information is used for approximation. The optimisation of the pose parameters is performed based on the appearance of the object contour projected into the image, applying the iterative closest point (ICP) algorithm by Zhang (1999b) for registration of free-form curves and surfaces (cf. Section 1.6.3). Further pose estimation algorithms rely on geometric (edges) and on photometric (surface brightness) information. In this context, Nayar and Bolle (1996) introduce an object representation based on reflectance ratios which is used to recognise objects from monocular brightness images of the scene. Pose estimation is performed relying on the reflectance ratio representation and the known geometric object properties, thus taking into account physical attributes of the object surface in addition to geometric information. Another technique which relies on the simultaneous extraction of edge and shading information for 2D-3D pose estimation is the appearance-based approach proposed by Nomura et al. (1996), who utilise synthetic edge and brightness images generated based on an object model. A nonlinear least-squares optimisation based on a comparison between the observed and the synthetic images yields the pose parameters. In Section 4.5 we regard a still more elaborate integrated 2D-3D pose estimation approach which in addition to edges and surface brightness also takes into account polarisation and defocus features.

Classical monocular pose estimation approaches have in common that they are not able to estimate the distance to the object at reasonable accuracy, since the only available information is the scale of a known object in the resulting image. Scale information yields no accurate results since for small distance variations the object

scale does not change significantly. In comparison, for a convergent stereo setup with a baseline similar to the object distance, for geometrical reasons a depth accuracy of the same order as the lateral translational accuracy is obtainable. For this reason, a variety of three-dimensional pose estimation methods relying on multiple images of the scene have been proposed more recently. For example, a fast tracking algorithm for estimating the pose of an automotive part from a pair of stereo images is presented by Yoon et al. (2003). Rosenhahn et al. (2006) compare the ICP algorithm for three-dimensional pose estimation in stereo image pairs with a numerical scheme which is introduced in the context of optical flow estimation. A quantitative evaluation of the two methods and their combination is performed, demonstrating that the highest stability and most favourable convergence behaviour is achieved with the combined approach. The method by von Bank et al. (2003), which is described in detail in Section 1.6.1.2, is extended by Krüger (2007) towards a multiocular setting characterised by three calibrated cameras. We refer to this work again in Section 5.1 in order to provide a quantitative comparison between monocular and multiocular methods for three-dimensional pose estimation.

1.6.1.2 Template-based Pose Estimation

Many industrial applications of pose estimation methods for quality inspection purposes impose severe constraints on the hardware to be used with respect to robustness and easy maintenance. Hence, it is often not possible to utilise multiocular camera systems since they have to be recalibrated regularly, especially when the sensor unit is mounted on an industrial robot. As a consequence, employing a monocular camera system may be favourable from the practical point of view while nevertheless a high pose estimation accuracy is required to detect subtle deviations between the true and the desired object pose. The appearance-based 2D-3D pose estimation method proposed by von Bank et al. (2003) involves a viewer-centred representation of the image data. The views are generated automatically from a three-dimensional object model by rendering, and the pose parameters of each view are stored in a table. Edge templates are computed for each view. For the input image, the best-fitting template and thus the corresponding pose parameters are determined by a template matching procedure. The difficult trade-off between the tesselation constant, i.e. the difference between the pose parameters of neighbouring views, and the accuracy of pose estimation is alleviated by a technique for hierarchical template matching (Gavrila and Philomin, 1999).

The input image first undergoes an edge detection procedure. A distance transform (DT) then converts the segmented binary edge image into a so-called distance image. The distance image encodes the distance in the image plane of each image point to its nearest edge point. If we denote the set of all points in the image as $A = \{^S\mathbf{a}_1, \dots, ^S\mathbf{a}_N\}$ and the set of all edge points as $B = \{^S\mathbf{b}_1, \dots, ^S\mathbf{b}_M\}$ with $B \subseteq A$, then the distance $d(^S\mathbf{a}_n, B)$ for point $^S\mathbf{a}_n$ is given by

$$d(^S\mathbf{a}_n, B) = \min_m (\| ^S\mathbf{a}_n - ^S\mathbf{b}_m \|), \quad (1.136)$$

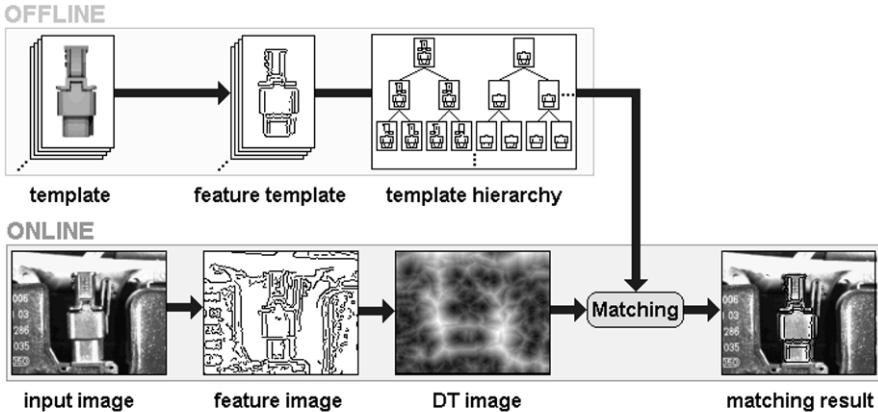


Fig. 1.28 Schematic description of the offline and the online part of the hierarchical template matching stage.

where $\| \dots \|$ is a norm on the points of A and B (e.g. the Euclidean norm). For numerical simplicity we use the so called chamfer-2-3 metric (Barrow, 1977) to approximate the Euclidean metric.

The chamfer distance $D_C(T, B)$ between an edge template consisting of a set of edge points $T = \{^S\mathbf{t}_1, \dots, ^S\mathbf{t}_Q\}$ with $T \subseteq A$ and the input edge image is given by

$$D_C(T, B) = \frac{1}{Q} \sum_{n=1}^Q d(^S\mathbf{t}_n, B). \quad (1.137)$$

In applications, a template is considered matched at locations where the distance measure (“dissimilarity”) $D(T, B)$ is below a user-supplied threshold θ . To reduce false detections, the distance measure was extended to include oriented edges (Gavrila and Philomin, 1999). A schematic description of the offline and the online part of the hierarchical template matching stage are shown in Fig. 1.28.

In order to recognize an object with unknown rotation and translation, a set of transformed templates must be correlated with the distance image. Each template is derived from a certain rotation of the three-dimensional object. In previous work, a uniform tesselation often involved the difficult choice for the value of the tesselation constant. If one chooses a relatively large value, the views that lie “in between” grid points on the viewing sphere are not properly represented in the regions where the aspect graph is undergoing rapid changes. This decreases the accuracy of the measured pose angles. On the other hand, if one chooses a relatively small value for the tesselation constant, this results in a large number of templates to be matched online; matching all these templates sequentially is computationally intensive and prohibitive to any real-time performance. The difficult trade-off regarding tesselation constant is alleviated by a technique for hierarchical template matching, introduced by Gavrila and Philomin (1999). That technique, designed for distance transform

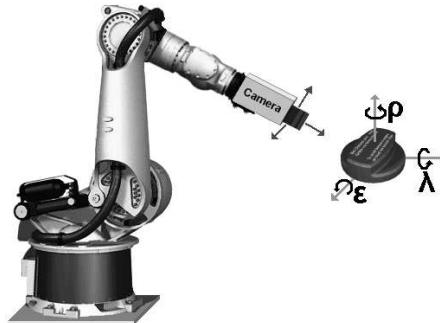


Fig. 1.29 Sketch of the robot-based inspection system with a definition of the pose angles ϵ (roll), λ (pitch), and ρ (yaw) in the camera coordinate system.



Fig. 1.30 Left: Best 30 matching solutions (drawn in black, best solution in white) for an automotive part (oil cap). Right: Matching results (best solution) for several example poses.

based matching, aims to derive a representation offline which exploits any structure in a particular template distribution, so that, on-line, matching can proceed optimized. This is done by grouping similar templates together and representing them by two entities: a “prototype” template and a distance parameter. When applied recursively, this grouping leads to a template hierarchy. It is built bottom-up, level by level using a partitional clustering algorithm based on simulated annealing.

Online, matching involves traversing the tree structure of templates. Each node corresponds to matching a (prototype) template p with the image at some particular locations. For the locations where the distance measure between template and image is below a user-supplied threshold θ_p , one computes new interest locations for the children nodes (generated by sampling the local neighborhood with a finer grid) and adds the children nodes to the list of nodes to be processed. For locations where the distance measure is above the threshold, search does not propagate to the sub-tree; it is this pruning capability that brings large efficiency gains. In our system, we do not need to estimate scale—the distance to the object is assumed to be known at an accuracy of better than 3 percent due to the fact that the system is designed for an industrial quality inspection scenario in which the approximate position of the parts is provided by CAD data. Template matching does not have to search all scales explicitly. Hence, the original pose estimation problem of determining six degrees of freedom can be reduced to a five degrees of freedom (three pose angles and two image position coordinates) problem.

For pose fine-tuning, the pose angles are interpolated between the n_b “best” template matching solutions, with $n_b = 30$ in our system. This is justified because in our pose estimation scenario the dissimilarity values of the 30 best solutions usually do not differ by more than about 20 percent and thus all these solutions contain a significant amount of information about the pose (Fig. 1.30).

In many applications, templates are generated from real-world image data (Demant, 1999). For inspection tasks, however, one can assume that a CAD model of the object to be inspected is available. We therefore generate realistic two-dimensional templates from CAD data using the public domain software POVRAY, simulating the properties of the surface material and the illumination conditions by employing raytracing techniques. The pose of the object is defined by the three angles ε (roll), λ (pitch), and ρ (yaw) as shown in Fig. 1.29. Typical matching results are shown for an automotive part in Fig. 1.30.

A quantitative evaluation of the described edge-based pose estimation technique is performed in the scenario of industrial quality inspection in Section 5.1.1. A multiocular extension is proposed by Krüger (2007), who simulates the appearance of the objects by varying the degrees of freedom in world coordinates and renders an image for each camera. This procedure results in a multiocular image corresponding to the concatenated images of the individual cameras. The template matching algorithm is similar to the monocular variant described above. The multiocular extension achieves a determination of all six degrees of freedom of a rigid object. To improve the robustness of monocular pose estimation in the presence of a cluttered background, the edge-based method is extended in Section 4.5.1 towards appearance-based monocular pose estimation based on geometric, photopolarimetric, and defocus cues (Barrois and Wöhler, 2007).

1.6.2 Pose Estimation of Non-rigid and Articulated Objects

In contrast to rigid objects, articulated objects consist of several rigid subparts which are able to move with respect to each other. Methods that aim for a pose estimation of such objects need to determine these internal degrees of freedom in addition to the six rotational and translational degrees of freedom encountered for rigid objects. Non-rigid objects have no rigid subparts at all and therefore have an infinite number of internal degrees of freedom. In this section we first give an overview of pose estimation methods for articulated and non-rigid objects. Then we regard the contour-based algorithm by d’Angelo et al. (2004) for three-dimensional reconstruction of non-rigid objects such as tubes and cables, which may be regarded as a multiocular extension of the concept of active contours (Blake and Isard, 1998). Subsequently, the multiocular contracting curve density algorithm (Hahn et al., 2007) is described which allows a three-dimensional reconstruction of non-rigid objects and a pose estimation of articulated objects in the presence of cluttered background.

1.6.2.1 General Overview

Non-rigid Objects

The extraction of two-dimensional object contours from images is an important problem in the field of image analysis. A classical approach to this problem are active contours or snakes. The original snake algorithm by Kass et al. (1988) attempts to determine a curve which has an internal energy depending on its shape, e.g. its length or curvature, and an external energy depending on the image information, e.g. denoting how closely the contour matches edges in the image, by minimising an overall energy term. Many variations and improvements of the original snake algorithm have been proposed, such as balloon snakes (Cohen, 1991), ziplock snakes (Neuenschwander et al., 1997), gradient vector field snakes (Xu and Prince, 1998), and implicit active contour models (Casselles et al., 1995; Sethian, 1999).

A different approach to two-dimensional curve fitting is the contracting curve density (CCD) algorithm introduced by Hanek (2004). The CCD algorithm employs a likelihood function as a quantitative measure of the fit between the curve model and the image data. This likelihood function depends on the local image grey value statistics determined based on the local vicinity of the expected curve. The posterior probability density is iteratively optimised with respect to the curve model parameters using blurred curve models, which enable the algorithm to trade-off the area of convergence and the accuracy of the fit. As a result, the CCD algorithm copes with highly inhomogeneous image regions and is capable of separating objects with ill-defined outlines from cluttered background. Later in this section we describe in detail a multiocular variant of the CCD algorithm which can be applied to the three-dimensional pose estimation of both non-rigid and articulated objects.

In many medical imaging applications, volumetric data need to be analysed, leading to the three-dimensional extension of the snake approach by Cohen and Cohen (1993). For pose estimation of non-rigid objects from multiple images, it is assumed by most approaches that the non-rigid object is adequately described by a one-dimensional curve in three-dimensional space. Such techniques are primarily useful for applications in medical imaging, e.g. for the extraction of blood vessels from orthogonal radiographs (Cañero et al., 2000) or for the inspection of bonding wires on microchips (Ye et al., 2001). A related method for extracting the three-dimensional pose of non-rigid objects such as tubes and cables from stereo image pairs of the scene based on three-dimensional ribbon snakes is described in detail later in this section.

Ellenrieder (2004) proposes a method similar to the shape from texture approach (Jiang and Bunke, 1997) which estimates the local surface normal of a textured object (a textile-coated tube) based on spatial variations of the amplitude spectrum of the surface texture. In the context of industrial quality inspection of tubes and cables, Ellenrieder (2005) introduces a method for three-dimensional pose estimation of non-rigid objects which is based on the evaluation of shadow contours on surfaces of arbitrary but known shape. A shadow image is obtained by computing the ratio image between two images of the scene, one of which is illuminated such that

the non-rigid object casts a shadow on the surface while no shadow of the object is visible in the second image (cf. also Section 2.1). The extracted shadow is rectified according to the known shape of the surface on which the shadow is cast, such that the object pose can be recovered based on the known light source position.

A novel approach to the computation of the derivatives of the bundle adjustment error function (1.11) for non-rigid objects is introduced by Krüger (2007). The underlying basic idea is to minimise the difference between the image and the model using a gradient descent scheme with precomputed gradient values which are reduced to their sign. This pose refinement algorithm is simple as it requires only one table look-up per feature for which a correspondence with the image is established. Using a quantised representation of the pose space leads to the necessity to store a single bit, denoting the sign of the gradient, per pixel, degree of freedom, and pose. The resulting bit matrices, which are termed gradient sign tables, have the same size as the image and can be computed offline. The computational complexity of this optimisation approach is quite low while its memory demand may become fairly high.

Articulated Objects

Most pose estimation approaches regarding articulated objects address the scenario of human body pose estimation. Moeslund et al. (2006) give a detailed introduction to and overview of the large field of pose estimation and tracking of the human body.

Many approaches, especially those aiming for gesture recognition in the context of human–robot interaction, rely on monocular image sequences. A more detailed overview of such techniques is thus given in Section 6.1.3. As an example, Schmidt et al. (2006) adapt an articulated body model consisting of chains of cylinders to monocular colour images of the person, where the optimisation of the pose parameters basically relies on skin colour detection as well as on intensity, edges, and the spatially varying statistical distribution of colour cues. Sminchisescu (2008) provides a broad discussion of the advantages and limitations resulting from monocular body pose estimation. Specifically, the problem of 3D–2D projection ambiguities is addressed, which may lead to reconstruction errors especially when partial self-occlusions of the body occur. Body pose estimation methods are divided into generative algorithms, relying on a model of the observation likelihood with maxima ideally centred at correct pose hypotheses, and discriminative algorithms, which learn the state conditional probabilities from examples and directly predict them using Bayesian inference.

An early approach by Gavrila and Davis (1996) to full body pose estimation involves template matching in several distance-transformed images acquired from different viewpoints around the person. Plänkers and Fua (2003) and Rosenhahn et al. (2005) apply multiple-view three-dimensional pose estimation algorithms which are based on silhouette information. Plänkers and Fua (2003) make use of three-dimensional data generated by a stereo camera to obtain a pose estimation and tracking of the human upper body. The upper body is modelled with implicit surfaces,

and silhouettes are used in addition to the depth data to fit the surfaces. Lange et al. (2004) propose a method to track a moving body in a sequence of camera images by adaptation of a stick figure model. A stochastic search algorithm is used to infer the model parameters. A comparison between rendered model images and the acquired images of the user yields an appropriate error function. Stereo views and relevance maps are used to refine the inferred joint angles of the model. Ziegler et al. (2006) use an ICP-based approach to estimate the pose of the upper human body, relying on a comparison between a three-dimensional point cloud obtained by stereo image analysis and a synthetically rendered depth map, obtained with a polygonal model of the upper body using the z-buffer technique. The body pose is tracked using an unscented Kalman filter. The system determines the torso position along with the joint angles of the shoulders and elbows.

Rosenhahn et al. (2005) track a three-dimensional upper body model with 21 degrees of freedom using a four-camera setup. The pose estimation is based on silhouettes which are extracted using level set functions. Tracking is performed by using the pose in the last frame as initial pose in the current frame. Under laboratory conditions with no cluttered background they achieve a high reconstruction accuracy of about 2° for the joint angles, which is demonstrated by comparison to a commercial marker-based tracking system with eight cameras. An extension of this method by Brox et al. (2008) uses the silhouette of a person extracted from one or multiple images for fitting a three-dimensional body surface model to it. The procedures of pose estimation and contour extraction based on level sets (Sethian, 1999) are coupled in order to allow tracking in scenes with a non-uniform and non-static background. In this context, Bayesian inference involving the local probability density models of image regions with respect to different features such as grey value, RGB colour values, or texture is used for simultaneously extracting a contour and a set of pose parameters. For large pose differences between successive images, prediction of the pose is achieved based on the optical flow. Since the pose estimation yields the 2D-3D correspondences in the current time step and the optical flow provides 2D-2D correspondences between image points in the current and the subsequent time step, a set of 2D-3D point correspondences can be computed for the subsequent time step, which then allows to determine the corresponding pose. A priori knowledge on joint angle configurations is used to impose constraints on the pose by learning the probability distribution of the joint angles from examples and incorporating it as a prior into the Bayesian inference scheme.

A method for tracking the motion of dressed people based on embedding a cloth draping technique into a silhouette-based system for body pose estimation is proposed by Rosenhahn et al. (2008). The body pose is determined by minimising an error functional with respect to silhouettes, pose and kinematic chain parameters, cloth draping components, and external forces. A fairly detailed modelling is performed, since parameters of the clothes of the person such as the length of a skirt are extracted during the optimisation process, and external physical forces on the clothes resulting e.g. from wind are explicitly modelled. A quantitative evaluation demonstrates that the error of the proposed method is less than one degree higher

than the error range of marker-based tracking systems despite the fact that parts of the tracked body are occluded by clothes.

Grest and Koch (2008) adapt a three-dimensional body model consisting of rigid fixed body parts of arbitrary shape to a three-dimensional point cloud extracted from a pair of stereo images with a dynamic programming based technique for dense disparity estimation. A maximum number of 28 pose parameters is estimated for the human body model. An ICP algorithm is used to perform a 3D-3D pose estimation of the human body. The Gauß-Newton, gradient descent, and stochastic meta descent optimisation methods are compared with respect to their convergence behaviour. The Gauß-Newton method turns out to be the superior approach for the regarded problem of ICP-based body pose estimation from three-dimensional point clouds.

A markerless system for three-dimensional body pose estimation specifically designed for biomechanical applications such as the investigation of normal and pathological human movement is described by Mündermann et al. (2008). It is based on an ICP technique involving an articulated surface model of the body with soft-joint constraints, such that the positions of the functional joints centres can be fine-tuned (within certain limits) during the model adaptation procedure. The person is segmented from the background by comparison to background images based on intensity and colour thresholding. The articulated model is adapted to the visual hull of the person, which is constructed from multiple (between 4 and 64) images of the scene. A direct comparison to a marker-based body pose estimation system yields accuracies of 10.6 ± 7.8 mm, 11.3 ± 6.3 mm, and 35.6 ± 67.0 mm for the full body and 8.7 ± 2.2 mm, 10.8 ± 3.4 mm, and 14.3 ± 7.6 mm for the lower limbs for 64, 8, and 4 cameras, respectively.

The approaches by Plänkers and Fua (2003), Rosenhahn et al. (2005), and Brox et al. (2008) determine a single pose which is updated at every time step. To increase the robustness of tracking, other approaches (Deutscher et al., 2001; Schmidt et al., 2006) use a particle filter according to Blake and Isard (1998). This probabilistic framework employs the Monte Carlo technique of factored sampling to propagate a set of samples (“particles”) through state space in an efficient manner. A problem in its application to tracking body parts is the high number of degrees of freedom, since the required number of particles rises exponentially with the dimension of the state space. Deutscher et al. (2001) extend the particle filtering scheme. To avoid local minima of the state probability they use additional resampling steps in a manner similar to simulated annealing. With this modification, they are able to track full body motion with 100 particles. They use edge detection and background subtraction to weight the particles. For three-dimensional body tracking, Schmidt et al. (2006) employ a kernel particle filter, which approximates the probability density in state space by a superposition of Gaussian kernels. They use 150 particles to track a three-dimensional model of the upper body with 14 degrees of freedom, relying on monocular colour images. The particles are weighted by the use of colour cues which are combined with ridge and edge cues. Due to the monocular approach, pose ambiguities may be encountered in this framework.

At this point it is useful to mention methods for three-dimensional pose estimation and tracking of the human hand, which also represents a complex articulated object with a large number of degrees of freedom. The survey by Erol et al. (2007) provides an overview of hand pose estimation techniques. As in the case of full body pose estimation, many approaches are based on monocular images and are thus two-dimensional, especially those aiming for recognising gestures in the context of human–robot interaction. Hence, an overview of such techniques is provided in Section 6.1.3. A three-dimensional method for hand pose estimation is introduced by Stenger et al. (2001), who utilise a detailed three-dimensional hand model with six degrees of freedom, which is composed of quadrics. The hand is segmented and tracked by minimising the geometric error between the model projection and the edges of the human hand in the image. An unscented Kalman filter is used to track the hand across an image sequence. Other three-dimensional approaches rely on range data, colour, and edge features (Jennings, 1999), point correspondences between model contours and edges in the image (cf. Erol et al. (2007) for details), or disparity information (Bray et al., 2004). These methods for hand pose estimation, which are actually similar to many of the previously described full body pose estimation approaches from the methodological point of view, are embedded into the context of human–robot interaction in Section 6.1.3.

The work by Stößel (2007) is one of the few studies that examine the problem of three-dimensional pose estimation of articulated objects in the context of industrial quality inspection. The described system achieves a pose estimation of objects composed of multiple rigid parts (“multi-part assemblies”) based on a single image of the object. The employed object models have up to about 30 degrees of freedom. The pose parameters are determined by a minimisation of the Hausdorff distance (Ruckridge, 1996) between edges in the image and lines inferred from the articulated object model projected into the image. Mutual occlusions of different parts of the assembly are explicitly taken into account. For minimisation of the error function, the extended kernel particle filter technique is developed which is used as a stochastic optimisation technique.

1.6.2.2 Three-dimensional Active Contours

In this section we describe a parametric active contour framework to recover the three-dimensional contours of rotationally symmetric objects such as tubes and cables. The proposed algorithm is a three-dimensional ziplock ribbon active contour algorithm based on multiple views (d’Angelo et al., 2004).

Active Contours

In the snake approach by Kass et al. (1988), the basic snake is a parametric function \mathbf{p} representing a contour curve or model:

$$\mathbf{p} = \mathbf{v}(s) \quad \text{for } s \in [0, l], \quad (1.138)$$

where \mathbf{p} is a contour point for a certain value of the length parameter s . An energy function E_C is minimised over the contour $\mathbf{v}(s)$ according to

$$E_C = \int_0^l E_{\text{snake}}(\mathbf{v}(s)) ds. \quad (1.139)$$

The snake energy E_{snake} is separated into four terms:

$$E_{\text{snake}}(\mathbf{v}(s)) = \alpha E_{\text{cont}}(\mathbf{v}(s)) + \beta E_{\text{curv}}(\mathbf{v}(s)) + \gamma E_{\text{ext}}(\mathbf{v}(s)) + \delta E_{\text{con}}(\mathbf{v}(s)). \quad (1.140)$$

The so-called internal energy $E_{\text{int}} = \alpha E_{\text{cont}}(\mathbf{v}(s)) + \beta E_{\text{curv}}(\mathbf{v}(s))$ regularises the problem by favouring a continuous and smooth contour. The external energy E_{ext} depends on the image at the curve point $\mathbf{v}(s)$ and thus links the contour with the image. Here we use the negative gradient magnitude of the image as the external energy, which then becomes $E_{\text{ext}} = -\|\nabla I(\mathbf{v}(s))\|$. The term E_{con} is used in the original snake approach by Kass et al. (1988) to introduce constraints, e.g. linking of point of the active contour to other contours or springs. User interaction can also be cast into E_{con} . Balloon snake techniques (Cohen, 1991) use E_{con} to “inflate” the active contour in order to counteract the shrinking introduced by the internal energy E_{int} . The weight factors α , β , γ , and δ can be chosen based on the application.

Disadvantages of the parametric model include its dependence on parameterisation, which may lead to numerical instabilities and self-intersection problems when applied to complex segmentation tasks. Implicit active contours models avoid these problems and handle topological changes automatically (Casselles et al., 1995). However, they are not used here due to their higher computational complexity. A contour does not need to be a single curve, and modifications to delineate ribbon structures, like roads in aerial images (Fua and Leclerc, 1990) or blood vessels in angiographic images (Hinz et al., 2001) have been proposed in the literature. In many applications, snakes are used as an interactive segmentation tool. A human operator can specify the initial contour roughly and can move the snake around possible local minima, for example with a force field attached to the mouse cursor (Kass et al., 1988).

We use the greedy active contours approach introduced by Williams and Shah (1992) as the basis for the three-dimensional snake framework. The contour is modelled by a polyline consisting of n points, and finite differences are used to approximate the energy terms E_{cont} and E_{curv} at each point \mathbf{p}_s , $s = 1, \dots, n$ according to

$$\begin{aligned} E_{\text{cont}}(\mathbf{v}(s)) &\approx \left| \|\mathbf{p}_s - \mathbf{p}_{s-1}\| - h \right| \\ E_{\text{curv}}(\mathbf{v}(s)) &\approx \|\mathbf{p}_{s-1} - 2\mathbf{p}_s + \mathbf{p}_{s+1}\|, \end{aligned} \quad (1.141)$$

where h is the mean distance between the polyline points. The greedy minimisation algorithm is an iterative algorithm which selects the point of minimal energy inside a local neighbourhood. The greedy optimisation is applied separately to each point, from the first point at $s = 0$ to the last point at $s = l$. The energy $E_C(\mathbf{v}(s))$

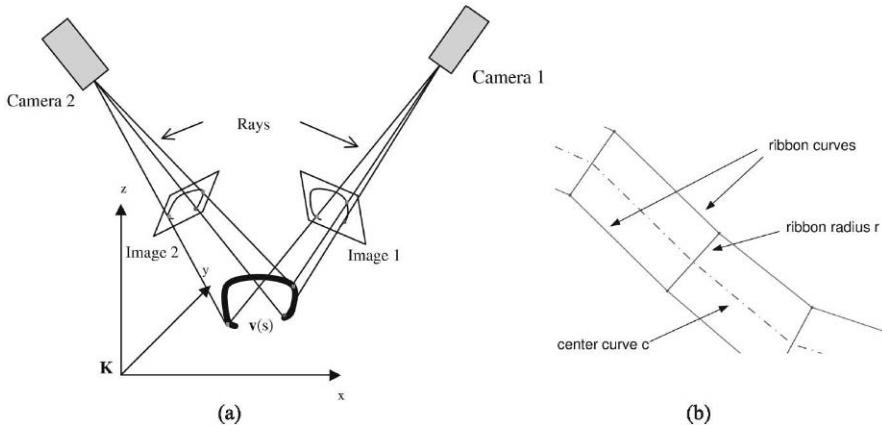


Fig. 1.31 (a) Projection of the three-dimensional curve model $v(s)$ into multiple images. Two cameras are observing the scene. Calculation of the external energy requires the contour curves in the image planes of all cameras. (b) Sketch of a ribbon snake.

is calculated for each candidate point \mathbf{p} in a neighbourhood grid $H \in \mathbb{R}^d$, where d is the dimensionality of the curve. The point \mathbf{p}_{\min} of minimum energy inside H is selected as the new curve point at $v(s)$. This procedure is repeated until all points have reached a stable position or a maximum number of iterations has been reached.

Since the greedy optimisation algorithm does not necessarily find a global minimum, it needs a good initialisation to segment the correct contours. Especially for the case of segmenting non-rigid objects, however, providing a suitable initialisation along the whole contour might not always be feasible. In such cases, the ziplock snake algorithm (Neuenschwander et al., 1997) is used. Ziplock snakes are initialised with the contour end points and their tangents. The contour is divided into active parts to which the full energy term E_{snake} applies and inactive parts which are only influenced by the internal regularisation energy E_{int} . The force boundaries start close to the end points and travel towards each other once the boundary point has converged to a stable position.

Three-dimensional Multiple-view Active Contours

If volumetric images are available, the image energy E_{ext} of a three-dimensional contour can be calculated directly from the volumetric data. In industrial quality inspection applications, volumetric data are not available, but it is usually possible to obtain images acquired from multiple viewpoints. In that case E_{ext} can be calculated by projecting the contour into the image planes of the cameras (Fig. 1.31a). The camera system is calibrated with the method by Krüger et al. (2004) (cf. Section 1.4). An arbitrary number N of images ($i = 1, \dots, N$) can be used for this projection. The intrinsic and extrinsic parameters of each camera are assumed to be known. Each model point ${}^W\mathbf{p}$ in the scene defined in the world coordinate system is projected

to the corresponding image point $S_i \mathbf{p}$ define in the sensor coordinate system. The energy term E_{ext} that connects the contour with the images is calculated based on the projection of the contour into each image plane according to $E_{\text{ext}}^* = \sum_{i=1}^N E_{\text{ext}}(S_i \mathbf{p})$. Hence, generating different active contours separately for each image or a depth map determined e.g. by stereo image analysis is therefore not necessary. All image and constraint information is used by the energy minimisation step in the model parameter space. In our examples, we use the image gradient as the external energy term, i.e. $E_{\text{ext}}(S_i \mathbf{p}) = -\|\nabla I_i(S_i \mathbf{p})\|$. Occlusions of the object are not considered during modelling, but the algorithm copes with partial occlusions and holes in the contour if they do not appear excessively. The energy terms E_{int} and E_{con} are independent of occlusions that may occur in some views.

The described approach requires that either the points of the extracted three-dimensional snake correspond to the same scene points, respectively, or that the object displays certain symmetries. In our scenario of inspection of tubes and cables, the objects to be segmented are rotationally symmetric with respect to a centreline, such that their silhouette is similar from multiple viewpoints and can be described by a centre curve $(x(s), y(s), z(s))^T$ and a radius $r(s)$, leading to $\mathbf{v}(s) = (x, y, z, r)^T$ (Fig. 1.31b). In this case, the contour model is not a simple curve and cannot be projected into the images as described above. Instead, the centre line $(x, y, z)^T$ is projected into the images while the corresponding ribbon curves are calculated by projecting the radius r into the image planes.

If prior knowledge about the contour is available, it can be integrated into the snake optimisation algorithm to improve its convergence. In an industrial production environment, CAD models of the parts that are subject to quality inspection are available and can be used to provide prior knowledge to the active contour segmentation. If the shape of objects should be recovered, only constraints that are invariant to the possible shapes of the object should be used. For quality inspection of tubes, this includes elasticity, length, radius and sometimes the mounting position, depending on the application. The model information is introduced into the optimisation process in two ways. The first is by adding additional energy terms, the second by using a constrained optimisation algorithm. Additional model-based energy terms that favour a certain shape of the contour can be added to E_{con} . For example, the approximate radius of a cable is known, but it may vary at some places due to labels or constructive changes.

As a first approach, we utilise a three-dimensional ribbon snake is used to detect the cable shape and position. Hence, $E_{\text{con}} = E_{\text{rib}} = [r(s) - r_{\text{model}}(s)]^2$ can be used as a “spring energy” to favour contours with a radius r close to a model radius given by the function $r_{\text{model}}(s)$. However, adding constraint terms to the objective function may result in an ill-posed problem with poor convergence properties and adds more weight factors that need to be tuned for a given scenario (Fua and Brechbühler, 1996). The second approach is to enforce model constraints through optimiser constraints. In the greedy algorithm this is achieved by intersecting the parameter search region H and the region C permitted by the constraints to obtain the allowed search region $H_c = H \cap C$. This ensures that these constraints cannot be violated. They are also called hard constraints. For some applications like glue line detection, the sur-

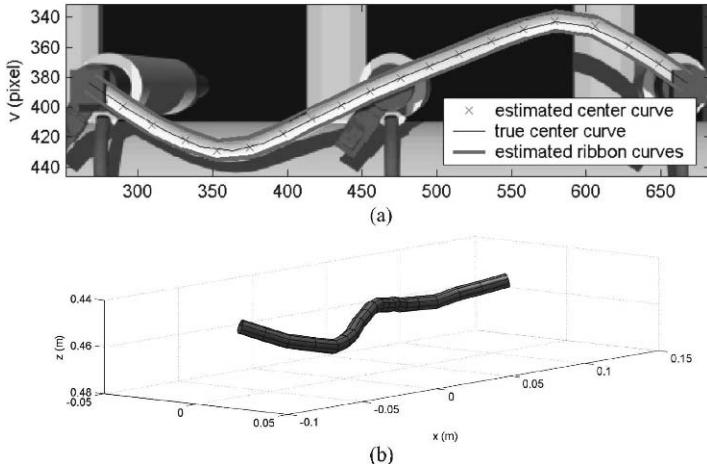


Fig. 1.32 (a) One of the three artificially rendered input images with the overlaid reprojections of the three-dimensional ribbon snake. A virtual trinocular camera with a resolution of 1024×768 pixels and a base distance of 100 mm has been used to generate the images. (b) Reconstructed tube, shown from a different viewpoint. The RMSE between ground truth and reconstruction is 1.5 mm.

face in which the contour is located is known, for example from CAD data. In other cases, the bounding box of the object can be given. This knowledge can be exploited by a constraint that restricts the optimisation to the corresponding surface or volume. Model information can also be used to create suitable initial contours—for example, tubes are often fixed with brackets to other parts. The pose of these brackets is usually given a priori when repeated quality inspection tasks are performed or can be determined by using pose estimation algorithms for rigid objects (cf. Section 1.6.1). These points can be used as starting and end points, i.e. boundary conditions, for three-dimensional ziplock ribbon snakes.

Experimental Results on Synthetic Image Data

As a first test, the described algorithm has been applied to synthetically generated image data, for which the ground truth is inherently available. For all examples, a three-dimensional ribbon snake was used. The weight factors of Eq. 1.140 were set to $\alpha = 1$, $\beta = 1$, $\gamma = 3$, and $\delta = 0$. Additionally, a hard constraint has been placed on the minimum and maximum ribbon width, which avoids solutions with negative or unrealistically large width. To estimate the reconstruction quality, a synthetically rendered scene was used as a test case as this allows a direct comparison to the known ground truth. Fig. 1.32a and b show the example scene and its three-dimensional reconstruction result. The start and end points of the object and their tangents were specified. In a real-world application these could either be taken from a CAD model, or estimated by pose estimation of the anchoring brackets. The ground truth used to produce the image is known and is compared to the segmented

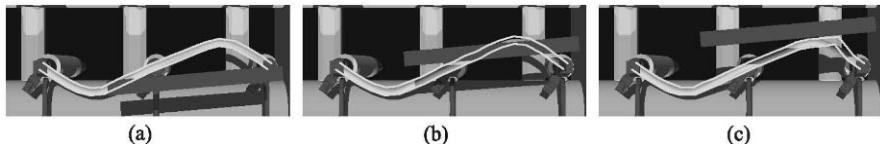


Fig. 1.33 Behaviour of the three-dimensional ziplock ribbon snake algorithm with respect to partial occlusion of the object. A virtual rod is moved across the rendered scene of Fig. 1.32.

contour. The utilised error measure is the root-mean-square error (RMSE) of the centre curve of the estimated ribbon snake with respect to the model centre curve. In the example shown in Fig. 1.32, the RMSE amounts to 1.5 mm, which roughly corresponds to 1 pixel disparity error. Subpixel greedy stepwidths and interpolated image energy calculation were used to obtain this result.

The behaviour of the algorithm in the presence of partial occlusions has been tested by moving a virtual rod over the scene as shown in Fig. 1.33. In Fig. 1.33a, where only a small part of the object is occluded, the RMSE with respect to the ground truth amounts to 1.1 mm, while for stronger occlusions as in Fig. 1.33b and c the RMSE corresponds to 3.1 mm and 3.8 mm, respectively. All described test cases were run on a 1.7 GHz Pentium Mobile Processor. The computation time of the optimisation procedure amounts to between 1 and 23 seconds, depending on the complexity of the scene and the parameters chosen for the optimisation procedure.

The experimental results show that the proposed three-dimensional ziplock ribbon snake algorithm is able to perform a fairly accurate three-dimensional contour segmentation. The stability of the algorithm is mainly due to the usage of model-based constraints and initial contour curves based on model information. However, the proposed method is sensitive to occlusions and self-intersections if they appear excessively. It is limited to lines and tube shaped objects, which is useful for a variety of applications e.g. in the field of industrial quality inspection. As real-world applications of the three-dimensional ziplock ribbon snake method, the three-dimensional reconstruction of a cable and of a glue line on a car body part are addressed in Section 5.2.

1.6.2.3 The Multiocular CCD Algorithm

As an example of three-dimensional pose estimation of articulated objects, this section addresses the problem of marker-less pose estimation and tracking of the motion of human body parts in front of a cluttered background. The multiocular contracting curve density (MOCCD) algorithm inspired by Hanek (2004) is introduced to determine the three-dimensional pose of the hand-forearm limb (Hahn et al., 2007). The object pose is tracked over time with a fairly traditional multiple Kalman filter framework. Due to the limited resolution of our trinocular greyscale camera setup it is infeasible in our system to model each finger of the hand as it is possible e.g. in the work by Stenger et al. (2001). Furthermore, a cylindrical model

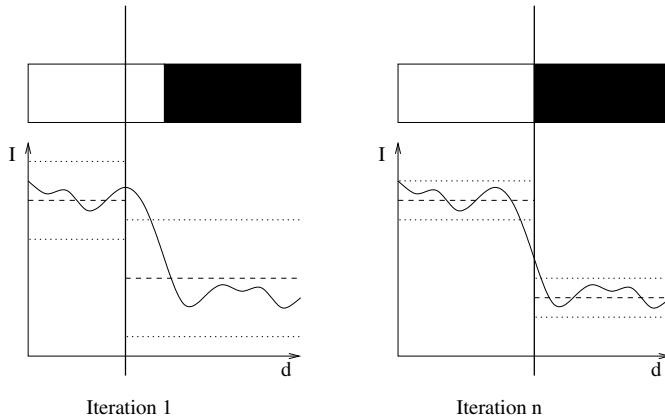


Fig. 1.34 The principle of the CCD algorithm. Fitting the segmentation boundary (bold line) to the image grey values (solid line) by estimating the mean (dashed line) and standard deviation (dotted line) on either side of the assumed boundary. The boundary is moved such that the image grey values have the highest probability according to the means and standard deviations.

of the forearm (Schmidt et al., 2006) is too coarse due to the variability of human appearance, e.g. clothes. Our approach is therefore based on a three-dimensional hand-forearm model which represents the three-dimensional contour by an Akima spline (Akima, 1970) using control points defined by a parameter vector. The MOCCD algorithm is computationally too expensive to use it in a particle filter framework. Hence we integrate the MOCCD algorithm in a traditional Kalman filter based tracking framework which estimates more than one pose hypothesis at a single time step.

The CCD Algorithm

The real-time CCD algorithm (Hanek, 2004) fits a parametric curve $c(\alpha, \mathbf{T})$ to an image I , e.g. marking the border between an object and the background. The parameter $\alpha \in [0, 1]$ increases monotonically along the curve, and \mathbf{T} denotes a vector containing the curve parameters to be optimised. The principle of the CCD algorithm is depicted in Fig. 1.34. The input values of the CCD are an image I and the Gaussian a priori distribution $p(\mathbf{T}) = p(\mathbf{T}|\hat{\mathbf{m}}_{\mathbf{T}}, \hat{\Sigma}_{\mathbf{T}})$ of the model parameters \mathbf{T} , defined by the mean $\hat{\mathbf{m}}_{\mathbf{T}}$ and the covariance $\hat{\Sigma}_{\mathbf{T}}$. The CCD algorithm estimates a model pose by computing the maximum of the a posteriori probability according to

$$p(\mathbf{T}|I) = p(I|\mathbf{T}) \cdot p(\mathbf{T}). \quad (1.142)$$

In Eq. (1.142) the value of $p(I|\mathbf{T})$ is approximated by $p(I|S(\mathbf{m}_{\mathbf{T}}, \Sigma_{\mathbf{T}}))$, with $S(\mathbf{m}_{\mathbf{T}}, \Sigma_{\mathbf{T}})$ representing the pixel value statistics close to the curve. The maximisation of Eq. (1.142) is performed by iterating the following two steps until the changes of $\mathbf{m}_{\mathbf{T}}$ and $\Sigma_{\mathbf{T}}$ fall below a threshold or a fixed number of iterations is

completed. The procedure starts from the user supplied initial density parameters $(\hat{\mathbf{m}}_T, \hat{\Sigma}_T)$.

1. Compute the pixel value statistics $S(\mathbf{m}_T, \Sigma_T)$ on both sides of the curve. For grey scale images this procedure amounts to computing a mean and a standard deviation of the pixel grey values on either side of the curve.
2. Refine the curve density parameters (\mathbf{m}_T, Σ_T) towards the maximum of Eq. (1.142) by performing one step of a Newton-Raphson optimisation procedure. This step moves the segmentation boundary such that the image content (grey values) conforms better with the pixel statistics, i.e. towards an edge.

A numerically favourable form of Eq. (1.142) is obtained by computing the log-likelihood

$$X = -2 \ln [p(I|S(\mathbf{m}_T, \Sigma_T)) \cdot p(\mathbf{T}|\hat{\mathbf{m}}_T, \hat{\Sigma}_T)]. \quad (1.143)$$

In terms of image processing this procedure can be seen as follows: The sum of Gaussian probability densities $p(I|S(\mathbf{m}_T, \Sigma_T))$ is an edge detector along the curve normal, i.e. if the curve is at the edge, the function value is maximal. In contrast to classical edge detectors (e.g. Sobel, Prewitt) the kernel size is adaptive and the function is spatially differentiable. These properties are the main reasons for the robustness and accuracy of the CCD algorithm.

Extension to Multiple Cameras

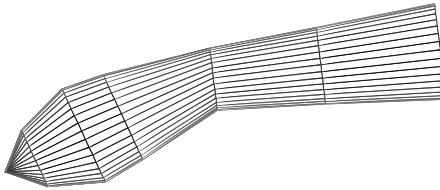
We extend the CCD algorithm to multiple calibrated cameras by projecting the boundary of a three-dimensional contour model into each image. The intrinsic and extrinsic parameters of the camera model by Bouguet (1999) are obtained by multiocular camera calibration (Section 1.4, cf. also Krüger et al. (2004)). An arbitrary number N_c of images can be used for this projection. We maximise the joint probability

$$p(\mathbf{T}|I_1, \dots, I_{N_c}) = \left[\prod_{c=1}^{N_c} p(I_c|S_c(\mathbf{m}_T, \Sigma_T)) \right] \cdot p(\mathbf{T}|\hat{\mathbf{m}}_T, \hat{\Sigma}_T) \quad (1.144)$$

with $S_c(\mathbf{m}_T, \Sigma_T)$ representing the grey value statistics close to the projected curve in image I_c . The underlying assumption is that images are independent random variables. Like in the original CCD framework, there is a numerically favourable form of Eq. (1.144) obtained by computing the log-likelihood. The MOCCD performs an implicit triangulation and is summarised in Algorithm 1. This evaluation scheme is an improvement over the approach by Krüger and Ellenrieder (2005), where the three-dimensional contour model is projected to N_c individual two-dimensional curves which are matched individually in two dimensions. At a later stage the individual curves are integrated to an update of the three-dimensional model by means of a 2D-3D pose estimation. The advantage of the improvement presented here is that spatial constraints are handled directly in the optimisation procedure instead of the 2D-3D pose estimation. Hence, the model shape is optimised to fit all images

Algorithm 1 Pseudocode of the MOCCD algorithm.**Input:** images I_1, \dots, I_{N_c} , a priori density $(\hat{\mathbf{m}}_T, \hat{\Sigma}_T)$ **Output:** refined model parameters (\mathbf{m}_T, Σ_T) **for** $iter = 1$ to $maxIter$ **do** **for** $c = 1$ to N_c **do**

Project the three-dimensional contour to its two-dimensional representation.

 Compute the statistics $S_c(\mathbf{m}_T, \Sigma_T)$ in image I_c . **end for** Refine the parameters (\mathbf{m}_T, Σ_T) towards the maximum of Eq. (1.144) by one Newton-Raphson step.**end for****Fig. 1.35** Articulated three-dimensional hand-forearm model.

equally well instead of allowing arbitrary two-dimensional deformations, as it is the case in the approach proposed by Krüger and Ellenrieder (2005).

Modelling the Hand-forearm Limb as an Articulated object

We use an analytical three-dimensional model of the human hand-forearm limb, a kinematic chain connecting the two rigid elements forearm and hand. The model consists of five truncated cones and one complete cone (Fig. 1.35). Fig. 1.36 (left) depicts the definition of the cones by the following nine parameters:

$$\mathbf{T} = (p_{1x}, p_{1y}, p_{1z}, \alpha_1, \beta_1, \alpha_2, \beta_2, r_1, r_4)^T. \quad (1.145)$$

The point ${}^W\mathbf{p}_1$ in three-dimensional space defines the beginning of the forearm and is part of the parameter vector \mathbf{T} . The wrist and fingertip positions ${}^W\mathbf{p}_2$ and ${}^W\mathbf{p}_3$ are computed by

$${}^W\mathbf{p}_2 = {}^W\mathbf{p}_1 + R_Z(\beta_1) \cdot R_Y(\alpha_1) \cdot l_{\text{forearm}} \cdot (1, 0, 0)^T \quad (1.146)$$

$${}^W\mathbf{p}_3 = {}^W\mathbf{p}_2 + R_Z(\beta_2) \cdot R_Y(\alpha_2) \cdot l_{\text{hand}} \cdot (1, 0, 0)^T, \quad (1.147)$$

where l_{forearm} and l_{hand} are the defined lengths of the human hand-forearm limb. The matrix $R_Y(\alpha)$ represents the rotation around the Y axis by the angle α and $R_Z(\beta)$ the corresponding rotation around the Z axis. The radii of the cones are computed by

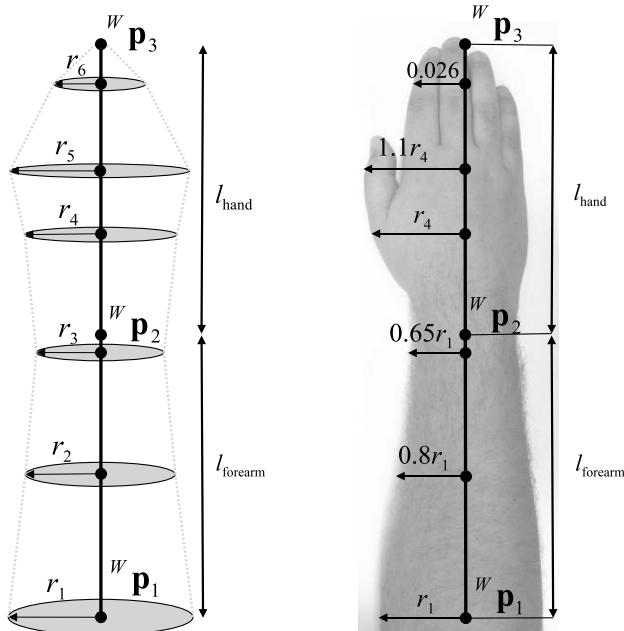


Fig. 1.36 Left: Definition of the cones. Right: Dependencies of the radii derived from human anatomy.

$$\begin{aligned}
 r_2 &= 0.8 \cdot r_1 \\
 r_3 &= 0.65 \cdot r_1 \\
 r_5 &= 1.1 \cdot r_4 \\
 r_6 &= 26 \text{ mm} = \text{const.}
 \end{aligned} \tag{1.148}$$

The dependencies of the radii are derived from human anatomy, see Fig. 1.36 (right). Only r_1 and r_4 are part of the parameter vector \mathbf{T} . As the MOCCD algorithm adapts a model curve to the image, the silhouette of the three-dimensional model in each camera coordinate system has to be extracted. Thus a vector from the origin of each camera coordinate system to the point in the wrist is computed, e.g. ${}^{C_1}\mathbf{p}_2$ for camera 1. This vector and the direction vector ${}^W\mathbf{p}_1 - {}^W\mathbf{p}_2$ of the forearm span a plane which is intersected with the three-dimensional model to yield the three-dimensional outline. The extracted three-dimensional contour model for the given camera—consisting of 13 points—is projected in the pixel coordinate system of the camera. The two-dimensional contour model is computed by an Akima interpolation (Akima, 1970) along the curve distance with the 13 projected points as control points. Fig. 1.37 depicts the extraction and projection of the three-dimensional contour model for camera 1.

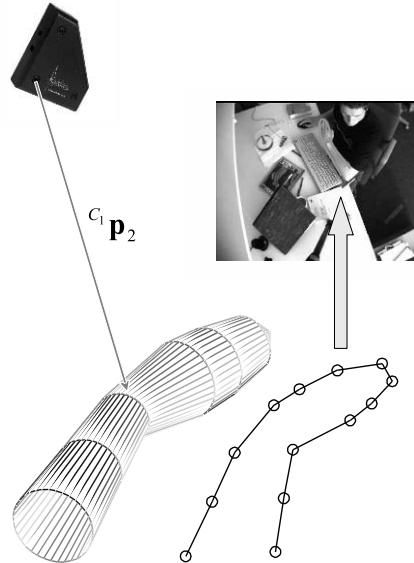


Fig. 1.37 Extraction and projection of the three-dimensional contour model.

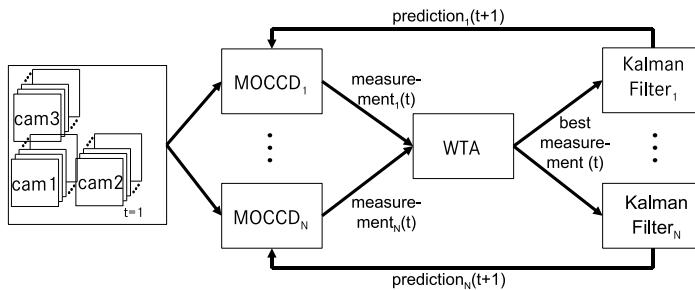


Fig. 1.38 Generic architecture of the recognition system.

System Overview

We have developed a generic recognition system (Fig. 1.38) consisting of three components: N instances of the MOCCD algorithm, a subsystem to find the best measurement of the different MOCCD instances (winner takes all) and N Kalman filters associated with the MOCCD instances. The input images are obtained with the PointGrey Digiclops trinocular greyscale camera system. The following algorithm outlines the interaction of the different components:

1. Initialisation.
2. For all MOCCDs: compute the measurement at time step t .
3. Find the best of these measurements in t .
4. For all Kalman filters: compute their prediction using the best measurement.

In the initialisation step the parameter vector $\mathbf{T}(t = 1)$, which describes the position of the hand-forearm limb at the first time step $t = 1$, is defined. After initialisation the developed system is able to track the human hand-forearm limb in three-dimensional space. The tracking is performed by steps 2 through 4. In the second step the measurements of all N MOCCDs in time step t are computed. The measurement of MOCCD i for the parameter vector \mathbf{T} consists of $(\mathbf{m}_{\mathbf{T},i}(t), \Sigma_{\mathbf{T},i}(t))$. The measurement of a MOCCD is based on the Gaussian a priori distribution $p(\mathbf{T}) = p(\mathbf{T}|\hat{\mathbf{m}}_{\mathbf{T}}, \hat{\Sigma}_{\mathbf{T}})$ of the model parameters \mathbf{T} , which are refined as described above. At time step $t = 1$ the Gaussian a priori distribution is user defined and in all other time steps it is computed by the prediction of the tracker associated with the MOCCD. To select starting parameters for Algorithm 1 the mean $\hat{\mathbf{m}}_{\mathbf{T},i}(t)$ of MOCCD i at time step t is obtained from the predicted parameter vector $\hat{\mathbf{T}}_i(t)$. The covariance $\hat{\Sigma}_{\mathbf{T},i}(t)$ is assumed to be a constant matrix.

After all MOCCD measurements are computed, the best measurement $\mathbf{T}(t)$ of the parameter vector \mathbf{T} at time step t is extracted based on a winner-takes-all approach with respect to $(\mathbf{m}_{\mathbf{T},1}(t), \Sigma_{\mathbf{T},1}(t), \dots, \mathbf{m}_{\mathbf{T},N}(t), \Sigma_{\mathbf{T},N}(t))$. As criteria for the measurement quality, we utilise the confirmation measurement of the MOCCD, the quality of the prediction, and the difference of the grey value statistics along the model curve. The confirmation measurement is introduced by Hanek (2004) and is an indicator of the convergence of the MOCCD. The second criterion describes how similar the prediction of the tracker and the measurement of the MOCCD are. With the third criterion it is ensured that the MOCCD separates grey value statistics along the projected curve. A measurement that is better than any other in at least two criteria is deemed the winner. In the fourth step of the algorithm the best measurement $\mathbf{T}(t)$ is used in each Kalman filter to produce the prediction of the parameter vector \mathbf{T} . These N predictions $\hat{\mathbf{T}}_1(t+1), \dots, \hat{\mathbf{T}}_N(t+1)$ are used to produce the measurements for $t+1$.

Kinematic Models

We investigate three different kinematic models (Bar-Shalom and Li, 1993) to find the best compromise between computational effort and tracking capabilities: A single Kalman filter with a constant-velocity model, two Kalman filters with constant-acceleration and constant-velocity models, and three Kalman filters with constant-acceleration, constant-velocity, and constant-position models. Each of the Kalman filters implements a different kinematic model, assuming a different object motion. Thus, once three Kalman filters are used, three different motion models are evaluated simultaneously. The winner-takes-all component then selects the best-fitting model in the next time step. The idea behind this kinematic modelling is to provide a sufficient amount of flexibility for changing hand-forearm motion. It is required for correctly tracking reversing motion, e.g. occurring during tightening of a screw.



Fig. 1.39 Typical result of pose estimation and tracking of the hand-forearm limb while performing a pointing gesture.

Results and Applications

A typical result of pose estimation and tracking of the hand-forearm limb while performing a pointing gesture is shown in Fig. 1.39. This example already indicates that the object pose is tracked at reasonable accuracy in the presence of a fairly cluttered background. Section 6.4 describes an extensive experimental evaluation of the described method in consideration of the scenario of interaction between humans and industrial robots. This quantitative evaluation is performed on a large set of real-world image sequences in direct comparison with the previously described three-dimensional active contour approach and a particle filter framework.

Spatio-temporal Extension

The shape flow method introduced by Hahn et al. (2008b) provides a spatio-temporal extension of the MOCCD framework. Effectively, it yields scene flow information for all points on the model surface. A three-dimensional spatio-temporal contour model is adapted to N_c multiocular images acquired at N_t subsequent time steps. The joint probability

$$p(\mathbf{T}|\{I_{c,t}\}) = [\prod_c \prod_t p(I_{c,t}|S_{c,t}(\mathbf{m}_T, \Sigma_T))] p(\mathbf{T}|\hat{\mathbf{m}}_T, \hat{\Sigma}_T) \quad (1.149)$$

is maximised with respect to the pose parameters \mathbf{T} and their temporal derivatives $\dot{\mathbf{T}}$, where $S_{c,t}(\mathbf{m}_T, \Sigma_T)$ represents the image statistics close to the projected curve in image $I_{c,t}$ taken by camera c at time step t . The underlying assumption is that the images are independent random variables. A numerically favourable form of Eq. (1.149) is again obtained by computing the log-likelihood.

A hierarchical approach is used to determine the three-dimensional pose \mathbf{T} and its temporal derivative $\dot{\mathbf{T}}$. The temporal derivatives of the model radii r_1 and r_4 (cf. Fig. 1.36) are not part of $\dot{\mathbf{T}}$ as they are constant. The three-dimensional pose $\mathbf{T}(t)$ at time step t is computed using the MOCCD algorithm based on the image

triple at time step t , while the spatio-temporal three-dimensional curve model at time steps $(t \pm \Delta t)$ required to evaluate Eq. (1.149) is computed according to $\mathbf{T}(t \pm \Delta t) = \mathbf{T}(t) \pm \dot{\mathbf{T}}(t) \cdot \Delta t$. Tracking is performed based on a winner-takes-all approach relying on two pose hypotheses, the first one assuming constant position and the second one constant velocity.

1.6.3 Point Cloud Segmentation Approaches

For the point-based three-dimensional pose estimation methods outlined in Section 1.6.1 explicit knowledge about correspondences between three-dimensional model points and two-dimensional image points is required. The problem of pose estimation is then equivalent to that of determining exterior camera orientation (cf. Section 1.4). In contrast, appearance-based pose estimation approaches like those described in Sections 1.6.1 and 1.6.2 do not rely on explicit correspondences but minimise the difference between the expected appearance of the object according to the estimated pose and the true object appearance.

In many scenarios, a three-dimensional description of the scene is given as a point cloud obtained e.g. by stereo image analysis (cf. Section 1.3) or with active sensors such as laser scanning devices. Initially, this point cloud contains no information about objects in the scene. In such cases, an important task is the segmentation of the point cloud into objects, either without using a-priori information or based on (weak or strong) model assumptions about the objects found in the scene. A scene segmentation without a-priori knowledge can be achieved by clustering methods (Duda and Hart, 1973), while an important approach to model-based segmentation of point clouds is the iterative closest point (ICP) algorithm (Besl and McKay, 1992; Zhang, 1992). Similar methods have been developed in the domain of photogrammetry e.g. to extract human-made objects such as buildings from topographic maps or terrestrial laser scanner data (Brenner, 2005; Rottensteiner et al., 2005; Rottensteiner, 2006). We regard in detail a method introduced by Schmidt et al. (2007) for detection and tracking of objects in a three-dimensional point cloud with motion attributes generated with the spacetime stereo approach described in Section 1.5.2.4. This method involves a clustering step relying on the spatial distribution and motion behaviour of the scene points, a subsequent model fitting stage, and a kernel particle filter for tracking the detected objects.

1.6.3.1 General Overview

Segmentation of a point cloud corresponds to the subdivision of the cloud into subparts that likely represent different objects. Without a-priori knowloedge about the sizes and shapes of the objects encountered in the scene, i.e. in the absence of model information, an appropriate approach to scene segmentation is clustering (Duda and Hart, 1973). The general definition of clustering is the partitioning of a data set into

subsets (clusters) such that each subset shares common properties. An intuitive similarity criterion is the distance in terms of an appropriately defined metric, such as the Euclidean distance of scene points in three-dimensional space—naturally, clustering techniques can also be applied in spaces with more than three dimensions. Hierarchical clustering techniques establish clusters successively, relying on clusters determined during previous steps, while partitional clustering techniques determine all clusters simultaneously. Hierarchical clustering algorithms are either agglomerative or divisive. Agglomerative clustering starts with each data point representing a cluster and merges small clusters into larger ones until a dissimilarity criterion between the clusters is met. Divisive clustering algorithms start with the complete data set and subdivide it into smaller clusters until the elements within each cluster are sufficiently similar. The representation of the cluster hierarchy is a dendrogram, i.e. a tree with individual data points at one end and one single cluster containing all data points element on the other end.

Classical Agglomerative and Divisive Clustering Methods

Agglomerative clustering algorithms start at the top of the tree and divisive algorithms at its bottom. The amount of detail of the clustering increases with increasing level of the tree. Useful measures for the distance between clusters are the maximum distance between elements of each cluster (complete linkage clustering), the minimum distance (single linkage clustering), and the average distance (average linkage clustering). Other distance definitions are the sum of the intra-cluster variances, the increase in variance for the merged cluster, and the probability that the clusters are generated by the same distribution function.

A classical divisive clustering algorithm is k -means clustering. Each data point is assigned to the closest cluster centroid, where the centroid is the average of all points in the cluster. According to MacQueen (1967), the algorithm consists of four steps: The desired number k of clusters is defined *a priori*. Then k random points are generated as cluster centroids. Each data point is assigned to the nearest cluster centroid, and the new cluster centroids are computed. These two steps are repeated until a suitable convergence criterion is fulfilled, e.g. that the assignment of the data points to the established cluster centroids does not change during an iteration. This clustering algorithm is simple and fast, but since the result of the k -means algorithm depends on the random initialisation, it is noteworthy that it does not yield identical results after different runs. A variant of the k -means algorithm is the fuzzy c -means algorithm. In principle, it is analogous to the k -means algorithm but differs in that to each data point a degree of belonging to each cluster is assigned according to the inverse distance to each cluster, and that the centroid of a cluster is the correspondingly weighted mean of all data points.

The EM Algorithm

Another important approach to clustering of data points is the expectation-maximisation (EM) algorithm (Dempster et al., 1977). This algorithm consists of an expectation (E) step by computing the expected likelihood according to the current values of the parameters to be estimated (e.g. the cluster indices of the data points) and a subsequent maximisation (M) step, which consists of determining a maximum likelihood estimate of the parameters by maximising the expected likelihood obtained during the E step. The E and the M step are performed in a alternating manner until a maximum of the likelihood function has been reached.

Unsupervised Neural Network Training

A different class of clustering methods is made up by unsupervised training approaches e.g. based on neural network architectures (Hertz et al., 1991). A classical example is the self-organising map, which is a single layer feedforward network in which the output neurons are arranged in a low-dimensional (usually two-dimensional or three-dimensional) lattice. Each input neuron is connected to every output neuron, such that each weight vector has the same dimension as the input vector, which is usually much higher than the dimension of the output lattice. Learning in the self-organizing map corresponds to associating different parts of the output lattice to respond in a similar manner to a certain class of input patterns, corresponding to an unsupervised partitioning of the high-dimensional input space in several clusters.

Graph Cut and Spectral Clustering

A more recent approach to the clustering problem is the normalized graph cuts method introduced by Shi and Malik (1997). In the context of image segmentation, the image is modelled as a weighted undirected graph, where each image pixel corresponds to a node in the graph while a graph edge is defined by every pair of pixels. The similarity between the pixel pair determines the weight of an edge, which may be based e.g. on the position in the image plane, grey value, gradient magnitude, gradient direction, and texture of the pixel and its local neighbourhood. For segmentation of three-dimensional point clouds, depth data can be regarded additionally. Removing the edges connecting the segments out of which the image consists leads to a partitioning of the image into disjoint sets. The optimal partitioning is the configuration with the minimal removed edge weights, which are termed the cut. Hence, such techniques are also known as graph cut methods. They have become fairly popular in the domain of image segmentation.

The graph cut approach is closely related to the method of spectral clustering (Shi and Malik, 2000; Fowlkes et al., 2004). It is assumed that a number of N image pixels is given by their previously determined attributes such as position, grey value,

texture, or depth. The basis of spectral clustering is the $N \times N$ symmetric similarity matrix W for the graph $G = (V, E)$ with nodes V representing pixels and edges E whose weights denote the pairwise similarities between pixels. If we assume that a bipartition of V is given by A and B with $A \cup B = V$ and $A \cap B = \emptyset$, $\text{cut}(A, B)$ denotes the sum of the weights between A and B according to

$$\text{cut}(A, B) = \sum_{i \in A} \sum_{j \in B} W_{ij}. \quad (1.150)$$

The degree d_i of node i is given by $d_i = \sum_j W_{ij}$, the volume of a set as the sum of the degrees within the set, i.e. $\text{vol}(A) = \sum_{i \in A} d_i$ and $\text{vol}(B) = \sum_{i \in B} d_i$. The normalised cut between the sets A and B is then given by

$$\text{ncut}(A, B) = \frac{2 \text{cut}(A, B)}{\text{vol}(A)\|\text{vol}(B)} \quad (1.151)$$

with the harmonic mean $a\|b = 2ab/(a+b)$. At this point it is desired to determine the graph bipartition given by A and B such that $\text{ncut}(A, B)$ is minimised. Shi and Malik (2000) show that an approximate solution is obtained by thresholding the eigenvector corresponding to the second smallest eigenvalue λ_2 of the normalised Laplacian \mathcal{L} given by

$$\mathcal{L} = D^{-1/2}(D - W)D^{-1/2} = I - D^{-1/2}WD^{-1/2} \quad (1.152)$$

with D as a diagonal matrix with elements $D_{ii} = d_i$. The matrix \mathcal{L} is positive semidefinite and its eigenvalues lie within the interval $[0, 2]$.

This framework is extended by Fowlkes et al. (2004) to more than two groups by regarding multiple eigenvectors to transform each pixel into an Euclidean space of dimension N_E , where N_E is much smaller than the original dimension N . This transformation aims for preserving significant differences between the pixels and at suppressing noise. In the N_E -dimensional space, groups of pixels are determined based on the k -means algorithm. The transformation from N -dimensional to N_E -dimensional space is obtained by computing the $N \times N_E$ matrix V of the largest eigenvectors and the $N_E \times N_E$ diagonal matrix Λ of the system $(D^{-1/2}WD^{-1/2})V = V\Lambda$. The i th coordinate E_{ij} of pixel j in the N_E -dimensional space is then given by $E_{ij} = V_{i+1,j} D_{jj}^{-1/2}$ with $i = 1, \dots, N_E$ and $j = 1, \dots, N$, where the eigenvectors are sorted in increasing order according to their corresponding eigenvalues.

The computational effort of this approach may become considerable as it increases quadratically with the number of pixels. Hence, Shi and Malik (1998) suggest an approximate version of the similarity matrix W in which each pixel is only connected to its nearest neighbours in the image plane while the other elements of W are set to zero, such that numerically efficient methods to determine the eigen-system of sparse matrices can be employed. Fowlkes et al. (2004) suggest to utilise the Nyström method (Press et al., 1992) to obtain an approximate solution of the eigensystem.

The ICP Algorithm

A method to register a point cloud to a geometric object model is the iterative closest point (ICP) algorithm introduced by Besl and McKay (1992). Given an initial estimate of the object pose, the pose parameters are updated by minimising the mean squared distance between the scene points and the model, and the point assignment is repeated accordingly. This procedure is applied in an iterative manner. As a result, the algorithm yields the three-dimensional object pose. Besl and McKay (1992) apply this method to the registration of point sets, curves, and surfaces. However, their approach can only be applied in situations where all scene points belong to the object. Hence, it is a pose estimation rather than a scene segmentation technique. In the ICP algorithm proposed by Zhang (1992), the scene points as well as the object model are represented as sets of chained points (similar e.g. to the three-dimensional object contours extracted by the contour-based stereo algorithm by Wöhler and Krüger (2003) described in Section 1.5.2.2). During each iteration step the pose parameters are updated while at the same time some scene points are assigned to the object model while others are rejected, based on the distance to the object and the similarity of the tangent directions of the scene and model curves. Thus, outliers in the point cloud data are automatically rejected, and the algorithm is robust with respect to disappearing and re-appearing object parts as well as partial occlusions. As a result, the algorithm yields the subset of scene points belonging to the object, i.e. a scene segmentation, along with the three-dimensional object pose.

Photogrammetric Approaches

In the domain of photogrammetry, an important application which requires the segmentation of three-dimensional point clouds is the reconstruction of buildings from images but also from airborne or terrestrial laser scanner data. As an example, Vögtle and Steinle (2000) propose a method for building extraction from airborne laser scanner and spectral data based on estimating the building vertices by local adjustment without incorporating geometric constraints. Rottensteiner (2006) states that at scales typical for topographic mapping, buildings can usually be modelled by polyhedra. Further geometrical regularities which can be exploited are the perpendicularity of walls, horizontally running roof edges, or symmetry constraints. Bottom-up reconstruction involves the segmentation of the sensor data in order to obtain three-dimensional features such as edges and planes, which are combined in a subsequent step to form a polyhedral building model (Rottensteiner et al., 2005). Assumptions about geometric regularities are considered as additional information during the final parameter estimation process. Top-down reconstruction (Brenner, 2000) corresponds to adapting parametric primitives to parts of the three-dimensional point cloud. Geometric regularities can either be taken into account by constraint equations (“hard constraints”) or by introducing “soft constraints” which allow residual deviations of the fitted building model from the observations.

Rottensteiner et al. (2005) describe a method for automatic reconstruction of buildings from airborne laser scanner data which is based on the extraction and subsequent combination of roof planes. Coplanar roof planes are merged, and the neighbourhood relations of the roof planes are used to generate hypotheses for intersecting lines and step edges. Geometric constraints are exploited in a subsequent step in order to obtain a consistent estimation of the building parameters. Rottensteiner (2006) suggests a refinement of the final parameter estimation step based on so-called Gestalt parameters describing a point situated on a polynomial surface. Roof models are simultaneously fitted to image and airborne laser scanner data.

Haala et al. (2006) construct topologically correct representations of building polyhedra from planar surface patches obtained from airborne laser scanner data, incorporating geometric constraints such as coplanarity or orthogonality. Besides the three-dimensional point cloud, their approach requires a two-dimensional ground plan. The building extraction method is based on the technique of cell decomposition, implying a reconstruction of polyhedral building models by subdivision of space into three-dimensional primitives.

1.6.3.2 Segmentation and Spatio-temporal Pose Estimation

The problem of three-dimensional scene segmentation along with the detection and pose estimation of articulated objects is primarily addressed in the context of human motion capture (cf. Section 1.6.2 for an overview). A technique for model-based three-dimensional human body tracking based on the ICP algorithm is presented by Knoop et al. (2005). Normal optical flow is used by Duric et al. (2002) to predict the location of a moving object in a “tracking by detection” framework. Translational and rotational components of camera motion are estimated by Gonçalves and Araújo (2002) based on a combined analysis of stereo correspondences and optical flow.

This section describes a vision system for model-based three-dimensional detection and spatio-temporal pose estimation of objects in cluttered scenes proposed by Barrois and Wöhler (2008). As low-level features, this approach requires a cloud of three-dimensional points attributed with information about their motion and the direction of the local intensity gradient. We extract these features by spacetime stereo based on local image intensity modelling according to Section 1.5.2.4. After applying a graph-based clustering approach to obtain an initial separation between the background and the object, a three-dimensional model is adapted to the point cloud based on an ICP-like optimisation technique, yielding the translational, rotational, and internal degrees of freedom of the object. An extended constraint line approach is introduced which allows to estimate the temporal derivatives of the translational and rotational pose parameters directly from the low-level motion information provided by the spacetime stereo data.

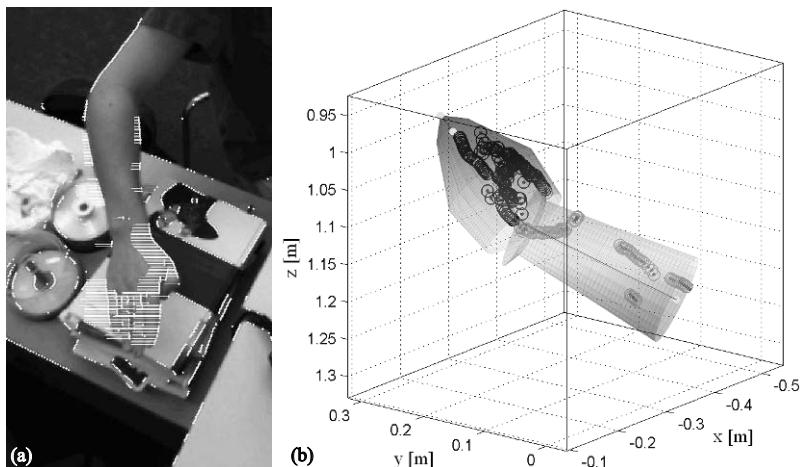


Fig. 1.40 (a) Image from a test sequence. Interest pixels for which stereo correspondences are established are shown as bright dots. Epipolar velocities are indicated as white lines. (b) Three-dimensional model adapted to the point cloud.

Scene Clustering and Model-based Pose Estimation

An initial segmentation of the attributed three-dimensional point cloud extracted with the spacetime stereo technique is obtained by means of a graph-based unsupervised clustering technique (Bock, 1974) in a four-dimensional space spanned by the spatial coordinates and the epipolar velocity of the three-dimensional points. This clustering stage generates a scene-dependent number of clusters, essentially separating the moving object from the (stationary or differently moving) background. For the first image of a sequence, the approximate position and orientation of the object are estimated based on a principal component analysis of the corresponding cluster points and used as initial values for the model adaptation procedure. For the subsequent images, the initial pose parameters are inferred for the current time step from the previous spatio-temporal pose estimation result as described later on.

We follow the ICP approach according to Zhang (1992) in order to fit a three-dimensional model of the hand-forearm limb (which does not necessarily represent the object at high accuracy) to the three-dimensional points determined to belong to the moving foreground object by the preceding clustering stage. We utilise the hand-forearm model introduced in Section 1.6.2 (Hahn et al., 2007), made up by a kinematic chain connecting the two rigid elements forearm and hand. The model consists of five truncated cones and one complete cone (cf. Fig. 1.40b). The cone radii corresponding to the hand and the upper end of the forearm are both set to 60 mm, and the lengths of the forearm and the hand are fixed to 220 mm and 180 mm, respectively. The other radii are inferred from human anatomy as described in Section 1.6.2. For each of the two rotationally symmetric model parts, the 5-dimensional vector \mathbf{T} of translational and rotational pose parameters is determined. The relative orientation

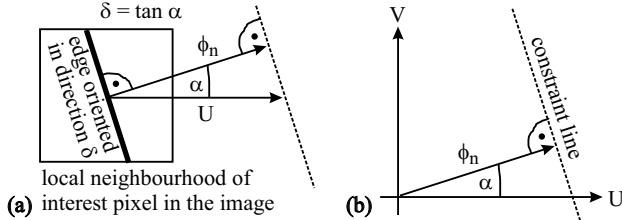


Fig. 1.41 (a) Relation between edge direction δ and normal velocity ϕ_n . (b) Definition of the constraint line in UV space representing the configurations (U, V) which are consistent with the observed normal velocity ϕ_n .

between forearm and hand is described by two angles, which are included into the model as internal degrees of freedom. In the course of the adaptation process, three-dimensional points not previously determined to belong to the object may be added to it while others may be rejected, resulting in a robust behaviour with respect to errors of the preceding clustering stage. The optimisation procedure is implemented as an M-estimator (Rey, 1983) with the “fair” weighting function. It is straightforward to utilise the result of this three-dimensional pose estimation procedure as an initialisation for the appearance-based MOCCD technique described in Section 1.6.2.

Estimation of the Temporal Pose Derivatives

Both motion components of a scene point parallel to the image plane can only be recovered from the corresponding local pixel neighbourhood if the intensity pattern around the pixel is corner-like. Edge-like intensity patterns only allow the determination of one velocity component, such as the component parallel to the epipolar lines computed by the spacetime stereo algorithm (cf. Section 1.5.2.4). This ambiguity is a consequence of the well-known aperture problem (Horn, 1986). Restricting the stereo and motion analysis to corner-like image features (Franke et al., 2005) may result in fairly sparse depth maps. If edge-like image features are evaluated, as it is the case in all image sequences regarded in this study, projecting the determined velocity component onto a line orthogonal to the local edge direction yields the normal velocity ϕ_n as depicted in Fig. 1.41a. The angle α between the direction of the horizontal epipolar lines and the direction of the normal velocity is given by $\delta = \tan \alpha$ with δ as defined by Eq. (1.127) in Section 1.5.2.4.

In the following, the translational velocity components of the object parallel to the x , y , and z axis are denoted by U_{obj} , V_{obj} , and W_{obj} , respectively, and expressed in metres per second. A two-dimensional space is spanned by the horizontal and vertical velocity components U and V measured in the scene and expressed in metres per second. This space is termed UV space. Given the observed normal velocity ϕ_n , all consistent configurations (U, V) are represented by the corresponding constraint line in UV space as defined by Schunck (1989) (cf. Fig. 1.41b). Fermüller and Aloimonos (1997) extend the concept of constraint lines towards the analy-

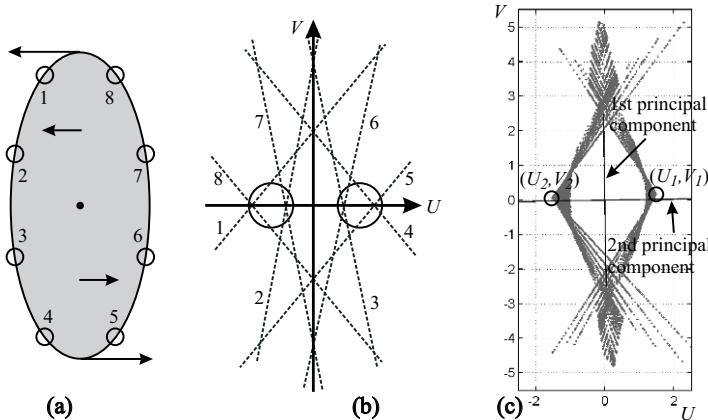


Fig. 1.42 (a) Rotating ellipse with reference points marked on its boundary. (b) Constraint lines resulting from the rotation of the ellipse. (c) Typical distribution of constraint line intersections in UV space for a real-world image from the first test sequence (cf. Fig. 1.40). The mean of the distribution has been subtracted from all points, the principal components are drawn as solid black lines.

sis of image displacement fields, i.e. optical flow fields (Horn and Schunck, 1981), stereo disparity fields, and normal flow fields, due to rigid motion. Rigid motion is e.g. present when the camera is moving through a static scene. It is shown that the image displacement field displays a global, scene-independent geometric structure, such that motion vectors of certain lengths and directions are constrained to lie on the image surface (specifically, the cases of spherical projection and projection on a planar surface are regarded) at particular locations. The locations of equal vectors, termed iso-motion contours by Fermüller and Aloimonos (1997), can be described by conic sections that only depend on the parameters of the three-dimensional rigid motion. The global structure of rigid motion fields is characterised by studying the properties of these constraint curves and regions and their relationships.

In the spatio-temporal pose estimation scenario, for an object performing a purely translational motion parallel to the image plane all constraint lines belonging to pixels on the object intersect in a single point in UV space. Both components of the translational motion are thus uniquely recovered. For objects with a rotational motion component in the image plane, a case which is not addressed by Horn and Schunck (1981) and Schunck (1989), the intersection points between constraint lines are distributed across an extended region in UV space. This situation is illustrated in Fig. 1.42a for an ellipse rotating counterclockwise. The constraint lines belonging to the indicated contour points are shown in Fig. 1.42b. In this example, the U coordinates of the constraint line intersection points are a measure for the mean horizontal velocities of the corresponding pairs of image points. The V coordinates have no such physical meaning. The distribution of intersection points is elongated in vertical direction due to the fact that a vertical edge detector is used for interest pixel extraction and because only image points with associated values of

$|\delta| < \delta_{\max}$ with δ_{\max} typically chosen between 1 and 2 (cf. Section 6.3) are selected by the spacetime stereo approach. Hence, constraint lines running at small angles to the U axis do not exist.

Fig. 1.42c shows a distribution of constraint line intersection points obtained from the scene shown in Fig. 1.40a, being typical of a rotationally symmetric and elongated object like the forearm partial model used in this study. The points in UV space are weighted according to the spatial density of the corresponding three-dimensional points along the longitudinal object axis. The mean $(U_{\text{obj}}, V_{\text{obj}})$ of the intersection point distribution, corresponding to the translational motion component of the object, has already been subtracted from the intersection points in Fig. 1.42c. The translational motion component W_{obj} parallel to the z axis is given by the median of the (fairly noisy) values of $\partial z / \partial t$ for all three-dimensional points assigned to the object or object part.

In the example regarded in Fig. 1.42c, scene points near the wrist are moving faster in the image plane than scene points near the elbow. The resulting intersection points are strongly concentrated near the points (U_1, V_1) and (U_2, V_2) depicted in Fig. 1.42c, which represent the motion of the scene points near the elbow and near the wrist. In this scenario, two circular markers attached to the upper and the lower end of the forearm, respectively, would yield two narrow clusters of intersection points in UV space at (U_1, V_1) and (U_2, V_2) . Regarding scene points at arbitrary positions on the forearm instead of well-localised markers yields a distribution which is largely symmetric with respect to the line connecting the points (U_1, V_1) and (U_2, V_2) . The information about the rotational motion of the object is thus contained in the range Δv covered by the projections of the intersection points on the principal component of the distribution which is oriented perpendicular to the longitudinal axis of the object (the second principal component in Fig. 1.42c). The value of Δv then corresponds to the velocity dispersion across the object caused by rotational motion in the image plane. In our system, we robustly estimate Δv based on the 10 percent and 90 percent quantiles of the distribution of the projection values. The angular velocity ω_p of the object rotation parallel to the image plane is then obtained by $\omega_p = \Delta v / \Delta l$ with Δl as the length interval parallel to the longitudinal object axis covered by the assigned three-dimensional points.

The rotation orthogonal to the image plane is determined based on the values of $\partial z / \partial t$ determined in Section 1.5.2.4 for the extracted three-dimensional points. For each model part, the projections $p^{(i)}$ of the assigned three-dimensional points on the longitudinal object axis are computed, and a regression line is fitted to the $(p^{(i)}, \partial z / \partial t^{(i)})$ data points. The slope of the regression line directly yields the velocity dispersion Δw in z direction and thus the angular velocity ω_o of the object rotation orthogonal to the image plane. Due to the rotational symmetry of the object models regarded in this study, the rotational motion of the object is already fully determined by the two components ω_p and ω_o of the angular velocity.

The technique described in this section allows to extend the determination of the vector \mathbf{T} of pose parameters by a direct estimation of the temporal pose derivative $\dot{\mathbf{T}}$ without the need for an object tracking stage.

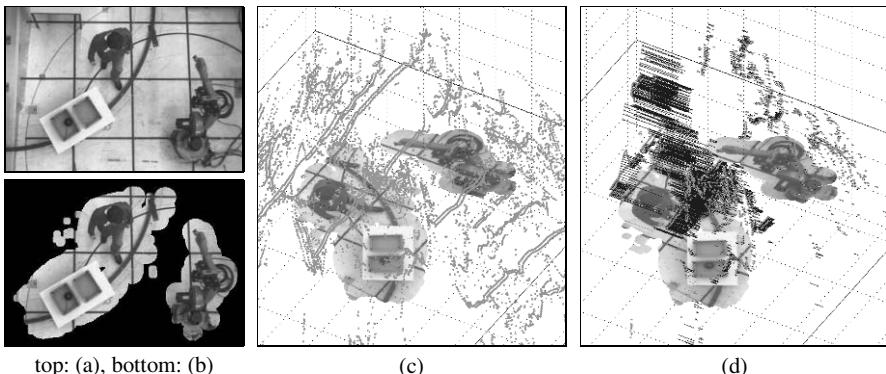


Fig. 1.43 (a) Original image (left camera). (b) Background subtracted image. (c) Full point cloud obtained with the correlation-based stereo vision technique. (d) Reduced motion-attributed point cloud.

1.6.3.3 Object Detection and Tracking in Point Clouds

This section describes the approach to the detection and tracking of objects in three-dimensional point clouds suggested by Schmidt et al. (2007), which relies on a motion-attributed point cloud obtained with the spacetime stereo approach described in Section 1.5.2.4. In a subsequent step, motion-attributed clusters are formed which are then used for generating and tracking object hypotheses.

Motion-attributed Point Cloud

A three-dimensional representation of the scene is generated with the correlation-based stereo vision algorithm by Franke and Joos (2000) and with the spacetime stereo algorithm described in Section 1.5.2.4 (cf. also Schmidt et al. (2007)). Both stereo techniques generate three-dimensional points based on edges in the image, especially object boundaries. Due to the local approach they are independent of the object appearance. While correlation-based stereo has the advantage of higher spatial accuracy and is capable of generating more point correspondences, spacetime stereo provides a velocity value for each stereo point. However, it generates a smaller number of points and is spatially less accurate, since not all edges are necessarily well described by the model defined in Eq. (1.124). Taking into account these properties of the algorithms, the results are merged into a single motion-attributed three-dimensional point cloud. For each extracted three-dimensional point c_k an average velocity $\bar{v}(c_k)$ is calculated, using all spacetime points s_j , $j \in (1, \dots, J)$ in an ellipsoid neighbourhood defined by $\delta_S(s_j, c_k) < 1$ around c_k . To take into account the spatial uncertainty in depth direction of the spacetime data, $\delta_S(s_j, c_k)$ defines a Mahalanobis distance whose correlation matrix Σ contains an entry $\Sigma_z \neq 1$ for the depth coordinate which can be derived from the recorded data, leading to

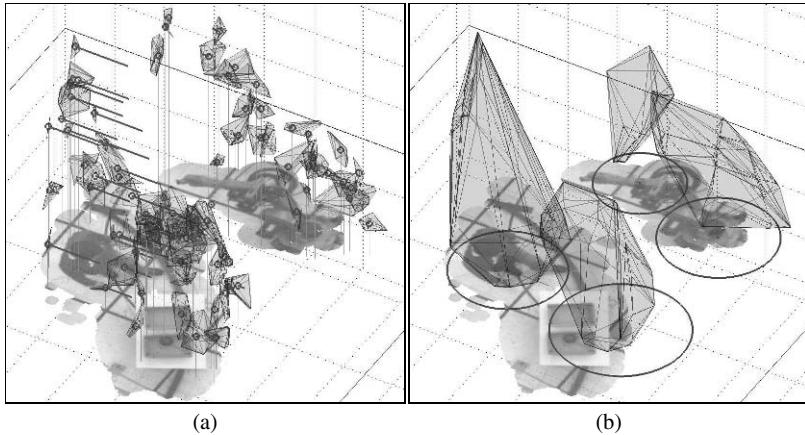


Fig. 1.44 (a) Over-segmentation and cluster velocities. (b) Objects with convex hull.

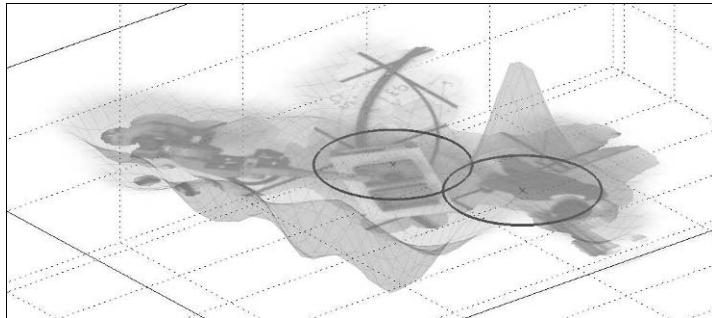


Fig. 1.45 Error function plot for minimisation, showing the error surface for $v = 0.26 \text{ m s}^{-1}$, $r = 0.53 \text{ m}$ (lower surface), and $v = -0.79 \text{ m s}^{-1}$, $r = 0.53 \text{ m}$ (upper surface, values mirrored for clearer display).

$$\bar{v}(c_k) = \frac{\rho}{J} \sum_{j=1}^J v(s_j) \quad \forall \quad s_j : \delta_S(s_j, c_k) < 1. \quad (1.153)$$

The factor ρ denotes the relative scaling of the velocities with respect to the spatial coordinates. It is adapted empirically depending on the speed of the observed objects. This results in a four-dimensional point cloud, where each three-dimensional point is attributed with an additional one-dimensional velocity component parallel to the epipolar lines, see Fig. 1.43d.

A reference image of the observed scene is used to reduce the amount of data to be processed by masking out three-dimensional points that emerge from static parts of the scene, as shown in Fig. 1.43a and b. Furthermore, only points within a given interval above the ground plane are used, as we intend to localise objects and humans and thus always assume a maximum height for objects above the ground.



Fig. 1.46 Trajectory of the tracked object (dark grey) with annotated ground truth (light grey) for a tabletop sequence.

Over-segmentation for Motion-attributed Clusters

To simplify the scene representation, we apply a hierarchical clustering algorithm, recognising small contiguous regions in the cloud, based on features like spatial proximity or homogeneity of the velocities. This procedure deliberately over-segments the scene, generating motion-attributed clusters. By incorporating velocity information for clustering, we expect an improvement in segmentation at these early stages of the algorithm, without needing strong models to ensure separation of neighbouring objects. For clustering, we apply the complete linkage algorithm to describe the distance between two clusters. The resulting hierarchical tree is partitioned by selecting a clustering threshold and addressing each subtree as an individual cluster (cf. Fig. 1.44a). The criterion for selecting the threshold is the increase in distance between two adjacent nodes in the tree, for which a maximum allowed value is determined empirically. For each resulting cluster l , the weight $w(l)$ is set according to the number of points P belonging to l as $w(l) = \sqrt{P}$. The square root is used to constrain the weight for clusters consisting of many points. For each cluster the mean velocity of all points belonging to it is determined.

Generation and Tracking of Object Hypotheses

From here on, persons and objects can be represented as a collection of clusters of similar velocity within an upright cylinder of variable radius. An object hypothesis $R(a)$ is represented by a four-dimensional parameter vector $\mathbf{a} = (x, y, v, r)^T$, with x and y being the centre position of the cylinder on the ground plane, v denoting the

velocity of the object and r the radius. This weak model is suitable for persons and most encountered objects.

To extract the correct object positions, we utilise a combination of parameter optimisation and tracking. We first generate a number of initial hypotheses, optimise the location in parameter space, and then utilise the tracking algorithm to select hypotheses which form consistent trajectories. Initial object hypotheses are created at each time step by partitioning the observed scene with cylinders and by including the tracking results from the previous frame, respectively. Multidimensional unconstrained nonlinear minimisation (Nelder and Mead, 1965), also known as the simplex algorithm, is applied to refine the position and size of the cylinders in the scene, so that as many as possible neighbouring clusters with similar velocity values can be grouped together to form compact objects, as shown in Fig. 1.44b. An error function $f(\mathbf{a})$ used for optimisation denotes the quality of the grouping process for a given hypothesis. Each hypothesis is weighted based on the relative position, relative velocity, and weight of all clusters l within the cylinder $R(\mathbf{a})$ using Gaussian kernels according to

$$f(\mathbf{a}) = f_r(\mathbf{a}) \sum_{l \in R(\mathbf{a})} w(l) f_d(l, \mathbf{a}) f_v(l, \mathbf{a}) \quad (1.154)$$

with $f_r(\mathbf{a}) = \exp\left(-\frac{r(\mathbf{a})^2}{2H_{r,\min}^2}\right) - \exp\left(-\frac{r(\mathbf{a})^2}{2H_{r,\max}^2}\right)$ keeping the radius in a realistic range, $f_d(l) = \exp\left(-\frac{|s(l)-s(\mathbf{a})|^2}{2H_d^2}\right)$ reducing the importance of clusters far away from the cylinder centre, and $f_v(l, \mathbf{a}) = \exp\left(-\frac{|v(l)-v(\mathbf{a})|^2}{2H_v^2}\right)$ masking out clusters having differing velocities. The functions $r(\mathbf{a})$, $s(\mathbf{a})$, and $v(\mathbf{a})$ extract the radius, the two-dimensional position on the ground plane, and the velocity of the hypothesis \mathbf{a} , respectively. The kernel widths H are determined empirically. Fig. 1.45 shows the error function from Eq. (1.154), parameterised for opposing velocities. Local minima are centred on top of the objects of interest.

After optimisation, hypotheses with identical parameterisation are merged and those without any clusters within $R(\mathbf{a})$ are removed. The remaining hypotheses are tracked over time using a particle filter, keeping only object hypotheses forming a consistent trajectory. The trajectory of a moving object in a fairly simple tabletop sequence as determined with the proposed algorithm is shown in Fig. 1.46 in comparison with the manually determined ground truth. This example demonstrates that a moving object can be separated from stationary background objects. In Section 6.2 we provide a detailed quantitative evaluation of our method in a complex industrial production scenario.

Chapter 2

Photometric Approaches to Three-dimensional Scene Reconstruction

In contrast to geometric methods, photometric approaches to three-dimensional scene reconstruction exploit the pixel intensities in the image to recover the shape of a surface or an object. Shadows in multiple images of a surface acquired under different illumination conditions may contain a significant amount of information about the surface shape (shape from shadow, cf. Section 2.1). The pixel intensities can be used to determine the position-dependent surface orientation, which by integration yields a depth map of the surface (shape from shading, cf. Section 2.2). Making use of the intensity information contained in multiple images leads to the photometric stereo approach described in Section 2.3. Utilising the polarisation state of light reflected from the surface instead of its intensity under certain circumstances allows a more accurate determination of surface orientation (shape from polarisation, Section 2.4).

2.1 Shape from Shadow

A shadow in the image of a surface contains information about the depth difference on the surface between the starting point of the shadow, e.g. the summit of a mountain, and its end point situated downslope and corresponding to the shadow tip, given that the direction of incident light is known. To explore the unknown properties of the surface of the Moon, the measurement of shadow lengths by visual observation was already utilised by seventeenth-century astronomers to estimate the heights of mountains and the depths of craters. Determination of a height difference Δz by determination of the shadow length l relies on the simple geometric relation

$$\Delta z = l \tan \mu, \quad (2.1)$$

where μ is the elevation angle of the light source and l has to be measured along the azimuthal direction of the incident light. Here, the geometric information l is extracted by regarding photometric information, in this case the absence of light

in certain parts of the image of the surface. Hence, shadow information is readily obtained from an image by a thresholding operation.

2.1.1 Extraction of Shadows from Image Pairs

A contour-based method for the extraction of shadows from image pairs at subpixel accuracy and the subsequent reconstruction of small-scale surface features is proposed by Hafezi and Wöhler (2004). Two images are required which show the same surface region illuminated from opposite directions, respectively, and under oblique illumination. Pixels with identical coordinates in these two images must correspond to the same physical points on the surface. One possibility to achieve this is to neither move the camera nor the object to be reconstructed during image acquisition, or to apply image registration techniques (Gottesfeld Brown, 1992). However, the same approach can be used to extract shadow regions from single images. This section describes how shadow regions can be determined from images at subpixel accuracy. Without loss of generality, it will always be assumed that the scene is illuminated exactly from the left or the right hand side.

Fig. 2.1a shows two images of the same part of the lateral surface of a friction plate. The scene is illuminated from the left and from the right hand side (images I_1 and I_2 , respectively). As neither the part nor the camera is moved, no image registration step is necessary. In the ratio images I_1/I_2 and I_2/I_1 , the shadows cast by the respective light source appear as bright regions as shown in Fig. 2.1b. This approach has the advantage to be insensitive to albedo variations of the surface because regions of low albedo that might be erroneously taken for shadows appear dark in both images, thus producing no bright region in one of the ratio images. A disadvantage is that surface parts situated in a shadow in both images are discarded—the illumination conditions should therefore be chosen such that this situation does not occur. The bright regions in the ratio images are segmented and analysed using the binary connected component (BCC) analysis algorithm (Mandler and Oberländer, 1990) which yields—among others—the properties area, centre coordinates, and a pixel contour description for each segmented region. This computation yields n_1 regions in the first and n_2 regions in the second image. The extracted contours are smoothed by B-spline interpolation (Rogers, 2000). Without loss of generality, it is assumed that the scene is illuminated along the image rows. We thus calculate the u (column) coordinates at which the interpolated contour intersects the rows in the v (row) range covered by the corresponding B-spline in order to obtain subpixel accuracy (cf. Section 1.5.2.2). Optionally, the B-spline contours can be used as an initialization to active contour techniques such as the one introduced by Williams and Shah (1992) for further refinement. The contours are represented as sets $\{\mathbf{c}_{1,a}^{(m)}\}$ and $\{\mathbf{c}_{2,b}^{(n)}\}$ of points for the first and the second image, respectively:

$$\mathbf{c}_{1,a}^{(m)} = \left(\hat{u}_{1,a}^{(m)}, \hat{v}_{1,a}^{(m)} \right), \quad a = 1, \dots, n_1 \quad (2.2)$$

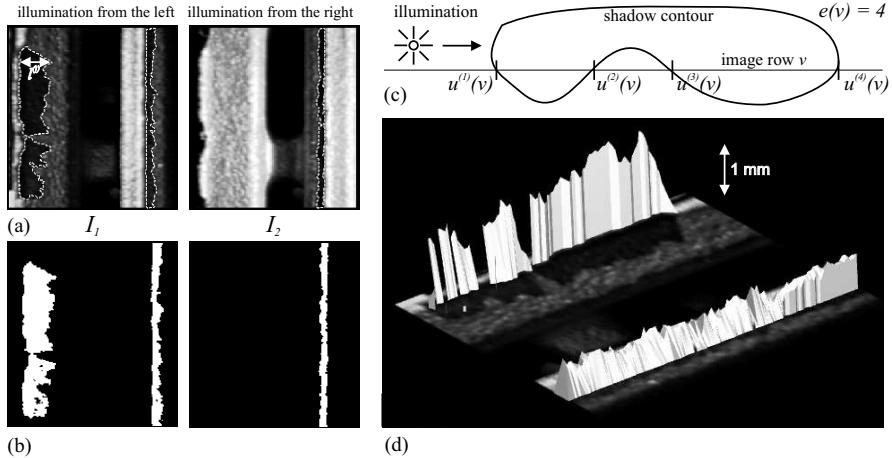


Fig. 2.1 (a) Side view of an industrial part (friction plate) displaying two ridges due to technical failure. The ridges are perpendicular to the surface such that they do not appear in perpendicular (top) view, as it is the case here. Their shadows on the surface, however, can be clearly distinguished (dashed contour lines). (b) Binarised ratio images. (c) Analysis of the shadow contour. (d) Three-dimensional reconstruction of the ridges based on their shadows.

$$\mathbf{c}_{2,b}^{(n)} = \left(\hat{u}_{2,b}^{(n)}, \hat{v}_{2,b}^{(n)} \right), \quad b = 1, \dots, n_2 \quad (2.3)$$

where a and b are the region indices in the images and m and n the point indices in the respective contours. The values $\hat{u}_{1,a}^{(m)}$ and $\hat{u}_{2,b}^{(n)}$ are real numbers, while the values $\hat{v}_{1,a}^{(m)}$ and $\hat{v}_{2,b}^{(n)}$ denote image rows and therefore are integer numbers.

For each image row v in the range covered by a contour, the number of intersections between the contour and the image row is calculated. These intersection counts are denoted by $e(v)$. Along with these values, the u coordinates of the intersections $u^{(i)}(v)$ with $i = 1, \dots, e(v)$ are determined. They are known at subpixel accuracy due to the B-spline representation of the contours. For each image row v , the intersections are sorted in ascending order according to their respective $u^{(i)}(v)$ values (Fig. 2.1c). For extraction of ridges of surface features, image rows v with an odd number $e(v)$ of intersections are discarded. For the shadows in image 1 with illumination from the left hand side, intersections $u^{(i)}(v)$ with even indices i are points at which a shadow ends, while intersections $u^{(i)}(v)$ with odd indices i are ridges that cast a shadow (Fig. 2.1c). The situation is just inverse in image 2 illuminated from the right hand side. Hence, the shadow lengths for image row v are

$$l^{(j)}(v) = \left| u^{(j)}(v) - u^{(j-1)}(v) \right| \quad \text{with } j \text{ even.} \quad (2.4)$$

Given the elevation angle μ of the light source with respect to the ground plane, the shadow length $l^{(j)}(v)$ yields information about the difference in depth (in the following denoted by z) at the corresponding pixel positions:

$$(\Delta z)_{\text{shadow}} = \left| z(u^{(j)}(v), v) - z(u^{(j-1)}(v), v) \right| = l^{(j)}(v) \tan \mu \quad \text{with } j \text{ even.} \quad (2.5)$$

The sign of the depth difference in Eq. (2.5) depends on whether the scene is illuminated from the right or from the left hand side. We assume that the scene is illuminated by parallel light and that the elevation angle μ of the light source with respect to the ground plane is known. For geometrical reasons an “aperture problem” occurs since structures that run parallel to the direction of incident light cannot be evaluated.

To extract the bottom of a surface feature, we make use of the fact that a surface region with a slope towards the light source appears brighter than a flat surface region. We thus segment and analyse with the BCC analysis algorithm all regions from image I_1 (I_2) which are brighter than a given threshold θ_1 (θ_2) and which are illuminated in one image while unlighted in the other. For the pixels of these regions, the relations $I_1 > \theta_1$ ($I_2 > \theta_2$) and $I_1/I_2 > \theta_0$ ($I_2/I_1 > \theta_0$) must hold, respectively. The contour lines of these image regions are obtained by BCC analysis and are used as an initialisation to the active contour algorithm described by Williams and Shah (1992) which then adapts them more accurately to the outline of the bottom of the surface feature.

The ridges and bottoms in the two images are merged based on the mutual distance of their centres, measured in pixel coordinates in image I_1 and I_2 , respectively, to form complete surface features. As it is not always evident which object in image I_1 belongs to which object in image I_2 , the algorithm is implemented such that it may suggest combinations based on their mutual distance that can either be accepted or rejected by the user. Fig. 2.1d shows the result of three-dimensional reconstruction by shadow analysis.

2.1.2 *Shadow-based Surface Reconstruction from Dense Sets of Images*

Kender and Smith (1987) present a method for extracting surface shape information based on self-shadowing under moving light sources, implicitly assuming an infinite number of images acquired under different light source elevation angles. The distance of the light source is assumed to be infinite, resulting in parallel incident light. The azimuthal direction of illumination is parallel to the image rows. The method regards the two-dimensional version of the shape from shadow problem, i.e. it reconstructs intersections of the surface with planes perpendicular to the xy plane, which are of the form $z(x)$. Thus, shadows are extracted in object coordinates (x, y, z) and not in pixel coordinates (u, v) . If a shadow begins at horizontal position x_i , for the surface slope the relation $\frac{\partial z}{\partial x} \Big|_{x_i} = \tan \mu$ is valid. At the same time, for the surface parts between the beginning x_i and the end x_e of a shadow the surface is located below the straight line through the points $z(x_i)$ and $z(x_e)$, which has the slope $\tan \mu$. Kender and Smith (1987) show that in the two-dimensional case the surface shape

can be fully recovered when the shadows are observed for the continuous range of all possible elevation angles μ . Also in the discrete case, i.e. for a finite number of observations, many images are necessary to obtain a good reconstruction accuracy. To extend the reconstruction to the full surface $z(x,y)$, Kender and Smith (1987) propose an iterative relaxation method.

This algorithm is extended by Hatzitheodorou (1989) by introducing a spline-based approach to recover the shape of a surface from a set of images displaying shadows. It is shown that the spline algorithm is the best possible approximation to the original function $z(x,y)$. The problem is formulated and solved in a Hilbert space setting, which allows to determine the spline parameters according to the observed shadows. The spline approach ensures a certain smoothness of the reconstructed surface, such that reasonable results can be obtained without the need for a continuously moving light source. However, it requires images acquired under several elevation and azimuth angles.

The previously described shadow constraints are formulated as graphs by Yu and Chang (2002), providing a systematic approach to represent and integrate shadow constraints from multiple images. They show that the shadow graph alone is sufficient to solve the shape from shadow problem from an infinite number of images. To recover the surface shape from a sparse set of images (utilising, however, a still considerable number of about 10 images), they propose a method to integrate shadow and shading constraints which is based on acquiring images of the scene from a fixed viewpoint under a variety of known illumination conditions. The height values of a small number of pixels are initialised appropriately, and an upper bound for the height value of each pixel is derived. A constrained optimisation procedure makes the height results obtained by shape from shading consistent with the shadow-based upper bounds.

Schlüns (1997) uses shadow information in the context of photometric stereo with multiple light sources to recover locally unique surface normals from two image intensities and a zero intensity caused by shadow. This approach incorporates shadow information qualitatively rather than quantitatively in order to remove the ambiguity of a surface reconstruction result obtained by photometric stereo (cf. Section 2.3) resulting from the missing intensity information.

The shape from shadow approach is combined with silhouette information by Savarese et al. (2002) to estimate the three-dimensional shape of an object. A space carving technique is employed to obtain an initial estimate of the object shape based on a set of silhouette images of the objects. Relying on a further set of images in which the object is illuminated by point light sources, the areas of self-shadowing on the object surface are compared to those expected from the initial shape estimate. The object shape is then refined in a shadow carving step that adjusts the current shape estimate to resolve contradictions with respect to the acquired images. This combined method reveals the object shape much more accurately than shape from silhouettes alone, but still does not yield a finely detailed object description unless a very large number of shadow images is taken into account.

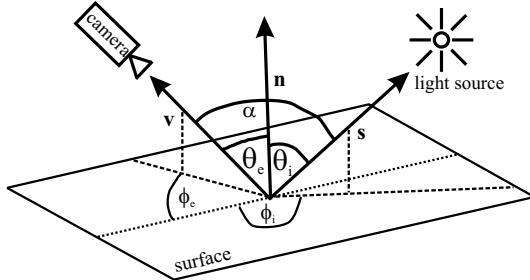


Fig. 2.2 Definition of the surface normal \mathbf{n} , the illumination direction \mathbf{s} , the viewing direction \mathbf{v} , the phase angle α , the polar angle θ_i and azimuth angle ϕ_i of incidence, and the polar angle θ_e and azimuth angle ϕ_e of emission. The vectors \mathbf{n} , \mathbf{s} , and \mathbf{v} are generally not coplanar, such that $\alpha \leq \theta_i + \theta_e$.

2.2 Shape from Shading

This section describes methods for three-dimensional reconstruction of object surfaces which are based on the pixel intensities of a single image. Photometric information does not directly allow a determination of height but merely of the surface gradients (Section 2.2.2). Early photometric approaches aiming for a reconstruction of height profiles along image rows (photoclinometry) emerged in the domain of remote sensing (Wilhelms, 1964). Under certain assumptions about the surface it is also possible to determine a full height map of the surface visible in the image (Section 2.2). This approach is termed shape from shading (Horn, 1986, 1989; Horn and Brooks, 1989). An overview of the variety of techniques to infer a height map from the determined surface gradients is provided in Section 2.2.3.

2.2.1 The Bidirectional Reflectance Distribution Function (BRDF)

Prior to the detailed description of photometric three-dimensional surface reconstruction methods we introduce the basic radiometric quantities. The light energy per unit time incident on the surface, measured in watts per square metre (W m^{-2}), is termed irradiance. The amount of light radiated away from a surface is defined as the power per unit area per unit solid angle, measured in watts per square metre per steradian ($\text{W m}^{-2} \text{ sr}^{-1}$), and is termed radiance. The normalisation to unit solid angle is necessary due to the fact that the surface may radiate different amounts of energy in different directions. For an image of a surface acquired by means of a lens, Horn (1986) demonstrates that if L_{surf} is the radiance of the surface in the direction towards the lens and E_{image} is the irradiance of the image at the regarded image position, the relation

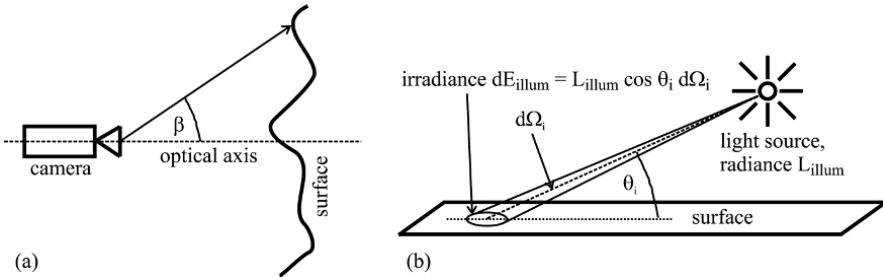


Fig. 2.3 (a) Definition of the angle β in Eq. (2.6). (b) Irradiance dE_{illum} due to illumination by a light source with radiance L_{illum} on a surface element covering a solid angle element $d\Omega_i$.

$$E_{\text{image}} = L_{\text{surf}} \frac{\pi}{4} \left(\frac{d}{f} \right)^2 \cos^4 \beta \quad (2.6)$$

is valid, where d denotes the lens diameter, f the focal length, and β the angle between the optical axis and the ray from the centre of the lens to the regarded scene point (cf. Fig. 2.3a). Hence, image irradiance E_{image} is proportional to surface radiance L_{surf} . The decrease of E_{image} with $\cos^4 \beta$ is termed vignetting effect. It occurs independent of how the lens is constructed. For most lenses, however, the observed vignetting is stronger than the falloff proportional to $\cos^4 \beta$ due to multiple apertures introduced to minimise distortions, because these apertures tend to cut off light rays inclined to the optical axis.

Scene radiance depends on the amount of light falling on a surface and on the fraction of the incident light that is reflected into the observing direction. Hence, the radiance of a surface depends on the direction from which it is viewed and on the direction of the incident light. The direction from which the surface is illuminated is given by the polar angle θ_i and the azimuth angle ϕ_i of incidence (cf. Fig. 2.2), while the direction into which the reflected light is emitted is denoted by the polar angle θ_e and the azimuth angle ϕ_e of emission. At this point we introduce the bidirectional reflectance distribution function (BRDF) $f(\theta_i, \phi_i, \theta_e, \phi_e)$, which denotes the radiance (brightness) L_{surf} of a surface viewed from the direction given by (θ_e, ϕ_e) , given an irradiance E_{illum} due to illumination from the direction given by (θ_i, ϕ_i) . Accordingly, the BRDF is defined as

$$f(\theta_i, \phi_i, \theta_e, \phi_e) = \frac{dL_{\text{surf}}(\theta_e, \phi_e)}{dE_{\text{illum}}(\theta_i, \phi_i)} = \frac{dL_{\text{surf}}(\theta_e, \phi_e)}{L_{\text{illum}}(\theta_i, \phi_i) \cos \theta_i d\Omega_i}. \quad (2.7)$$

In Eq. (2.7), $dL_{\text{surf}}(\theta_e, \phi_e)$ denotes the differential radiance in viewing direction, $dE_{\text{illum}}(\theta_i, \phi_i) = L_{\text{illum}} \cos \theta_i d\Omega_i$ the differential irradiance from the light source situated in direction (θ_i, ϕ_i) onto the surface (cf. Fig. 2.3b), $L_{\text{illum}}(\theta_i, \phi_i)$ the total radiance of the light source, and $d\Omega_i$ the differential solid angle under which the incident light falls on the regarded surface element. The factor $\cos \theta_i$ originates from the foreshortening of the surface as seen from the light source. The physical unit of the BRDF is sr^{-1} .

For many materials encountered in the real world, the BRDF does not depend separately on the angles ϕ_e and ϕ_i but only on the difference $(\phi_e - \phi_i)$. This is true for diffusely and specularly reflecting surfaces but not for surfaces with oriented microstructures. In the case of dependence of the BRDF on the difference $(\phi_e - \phi_i)$ it is often convenient to express the BRDF in terms of the incidence angle θ_i , the emission angle θ_e , and the so-called phase angle α between the illumination direction \mathbf{s} and the viewing direction \mathbf{v} (cf. Fig. 2.2 and Section 2.2.2.1).

For two surfaces in thermal equilibrium, the second law of thermodynamics implies that radiation energy emitted from one surface to the other must be compensated by an identical amount of radiation energy in the opposite direction, since otherwise the thermal equilibrium would be disturbed, resulting in one surface heating up and the other one cooling down. This equilibrium assumption implies that a physically motivated BRDF fulfills the so-called Helmholtz reciprocity condition

$$f(\theta_i, \phi_i, \theta_e, \phi_e) = f(\theta_e, \phi_e, \theta_i, \phi_i). \quad (2.8)$$

In this context, an important special case is the ideal Lambertian surface, which appears equally bright from all directions while reflecting all incident light. For such a surface, the BRDF must be constant, i.e. independent of the angles θ_i , ϕ_i , θ_e , and ϕ_e . To determine the value of the constant, we follow the derivation outlined by Horn (1986). The integral of the radiance of the surface over all directions must be equal to the total irradiance, leading to

$$\int_{-\pi}^{\pi} \int_0^{\pi/2} f(\theta_i, \phi_i, \theta_e, \phi_e) E(\theta_i, \phi_i) \cos \theta_i \sin \theta_e \cos \theta_e d\theta_e d\phi_e = E \cos \theta_i. \quad (2.9)$$

By taking into account that $2 \sin \theta_e \cos \theta_e = \sin 2\theta_e$, we obtain $\pi f = 1$ for an ideal Lambertian surface, corresponding to

$$f_{\text{Lambert}}(\theta_i, \phi_i, \theta_e, \phi_e) = \frac{1}{\pi}. \quad (2.10)$$

Hence, the radiance L is obtained from the irradiance E_0 by $L = E_0/\pi$.

A BRDF model which is widely used in the domain of computer graphics to represent smooth surfaces with a diffuse and a specular reflectance component has been introduced by Phong (1975). The Phong BRDF is given by the sum of a Lambertian BRDF component and a specular BRDF component according to

$$f_{\text{Phong}}^{\text{spec}}(\theta_i, \theta_r) = \sigma \frac{(\cos \theta_r)^m}{\cos \theta_i} \quad (2.11)$$

with σ and m as constant parameters and θ_r as the angle between the viewing direction \mathbf{v} and the specular direction into which the light would be reflected if the surface were an ideal mirror. The parameter σ denotes the strength of specular reflection relative to diffuse reflection, while the parameter m governs the width of the specular reflection. Setting $m \rightarrow \infty$ corresponds to an ideal mirror. The angle θ_r can be expressed in terms of the angles θ_i , θ_e , and α by

$$\cos \theta_r = 2 \cos \theta_i \cos \theta_e - \cos \alpha. \quad (2.12)$$

It is important to note that the Phong BRDF is a purely phenomenological model, i.e. it is not motivated by the physical processes governing the reflection of light at real surface materials. As a consequence, it does not fulfill the Helmholtz reciprocity condition (2.8), but it is nevertheless a useful representation of the reflectance behaviour of specular surfaces.

A physically motivated BRDF model for rough surfaces is introduced by Torrance and Sparrow (1967). For many rough surfaces the specular reflection component does not reach its peak exactly in the forward scattering direction $\theta_e = \theta_i$, as it would be expected from an ideal mirror, but the maximum is shifted towards higher values of the emission angle θ_e . The Torrance-Sparrow model assumes that the surface is composed of a large number of microfacets. The corresponding BRDF is given by the sum of a Lambertian component and a specular component f_{TS}^{spec} according to

$$f_{TS}^{\text{spec}}(\theta_i, \theta_e, \phi, n, k, w, \delta) = \frac{F(\theta_i, \theta_e, \phi, n, k)G(\theta_i, \theta_e, \phi) \exp(-w^2\delta^2)}{\cos \theta_i \cos \theta_e}. \quad (2.13)$$

In Eq. (2.13), the angle ϕ denotes the azimuth difference ($\phi_e - \phi_i$) and δ the angle of the surface normal of the microfacet with respect to the macroscopic surface normal. The refraction index of the surface is assumed to be complex and amounts to $n + ik$ with k as the attenuation coefficient (cf. also Section 2.4.2). The expression $F(\theta_i, \theta_e, \phi, n, k)$ is the Fresnel reflectance as given e.g. by Hapke (1993). $G(\theta_i, \theta_e, \phi)$ is the geometric attenuation factor which describes the effects of masking and shadowing and obtains values between 0 and 1. Setting $G(\theta_i, \theta_e, \phi) = 1$ is a good approximation in many scenarios (Meister, 2000). Nayar et al. (1991) give the analytic expression

$$G(\theta_i, \theta_e, \phi) = \min \left(1, \frac{2 \cos \delta \cos \theta_e}{\cos \theta'_i}, \frac{2 \cos \delta \cos \theta_i}{\cos \theta'_e} \right) \quad (2.14)$$

with θ'_i as the local illumination angle of the surface facet. Meister (2000) shows that $\cos 2\theta'_i = \cos \theta_i \cos \theta_e - \sin \theta_i \sin \theta_e \cos \phi$. The Torrance-Sparrow model assumes that the orientations δ of the normals of the surface facets have a Gaussian probability distribution proportional to $\exp(-w^2\delta^2)$, where w denotes a width parameter which may e.g. be determined empirically. An analytic expression for δ as a function of θ_i , θ_e , and ϕ is given by Meister (2000).

Simple BRDF models like the Lambertian or the Phong BRDF are generally of limited accuracy when used in the context of three-dimensional surface reconstruction. On the other hand, determination of the parameters of physically motivated BRDFs such as the Torrance-Sparrow model is not always possible or requires considerable experimental efforts. Hence, in the application scenario of industrial quality inspection described in detail in Chapter 5 we make use of a phenomenological BRDF model for rough metallic surfaces which is an extension of the model by Phong (1975) and has been chosen such that the empirically determined reflectance

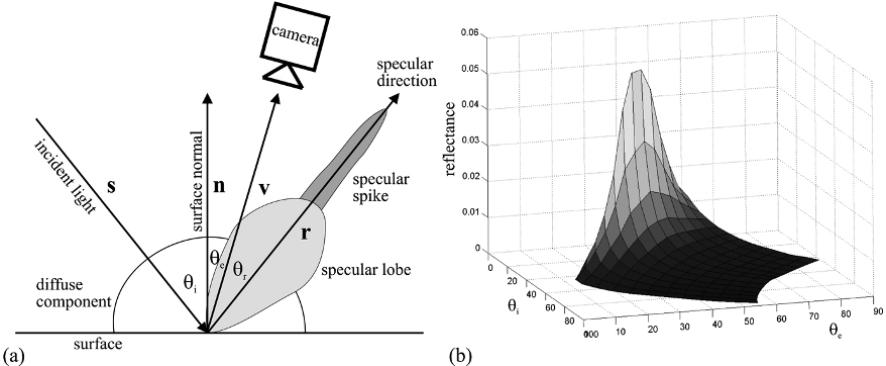


Fig. 2.4 Sketch of the three reflectance components according to Eq. (2.15). (b) Measured reflectance of a raw forged iron surface for $\alpha = 75^\circ$.

properties of the material are well represented. Here, the reflectance of a typical rough metallic surface is assumed to consist of three components: a diffuse (Lambertian) component, the specular lobe, and the specular spike (Nayar et al., 1991). The diffuse component is generated by internal multiple scattering processes. The specular lobe, which is caused by single reflection at the surface, is distributed around the specular direction and may be rather broad. The specular spike is concentrated in a small region around the specular direction and represents mirror-like reflection, which is dominant in the case of smooth surfaces. Fig. 2.4a illustrates the three components of the reflectance function. For illustration, Fig. 2.4b shows a reflectance function measured for raw forged iron at a phase angle of $\alpha = 75^\circ$ (as shown in Section 2.2.2, for the materials regarded here the reflectance function corresponds to the BRDF multiplied by $\cos \theta_i$). We define an analytical form for the reflectance for which we perform a least-mean-squares fit to the measured reflectance values, depending on the incidence angle θ_i and the angle θ_r between the specular direction \mathbf{r} and the viewing direction \mathbf{v} (cf. Fig. 2.4a):

$$f_N^{\text{spec}}(\rho, \theta_i, \theta_r, \alpha) = \rho \left[1 + \sum_{n=1}^N \sigma_n \cdot \frac{(\cos \theta_r)^{m_n}}{\cos \theta_i} \right]. \quad (2.15)$$

For $\theta_r > 90^\circ$ only the diffuse component is considered. The reflectance measurement is performed for a small part of the surface, for which the albedo ρ can be assumed to be constant. The shapes of the specular components of the reflectance function are approximated by $N = 2$ terms proportional to powers of $\cos \theta_r$. The coefficients σ_n denote the strength of the specular components relative to the diffuse component, while the exponents m_n denote their widths. Generally, all introduced phenomenological parameters depend on the phase angle α . The angle θ_r is defined according to Eq. (2.12), such that our phenomenological reflectance model only depends on the incidence angle θ_i , the emission angle θ_e , and the phase angle α . Like the Phong model, the specular BRDF according to Eq. (2.15) does not fulfill

the Helmholtz reciprocity condition (2.8), but it provides a useful and numerically well-behaved description of the reflectance behaviour of the kind of specularly reflecting rough metallic surfaces regarded in Chapter 5 in the context of industrial quality inspection.

In the scenario of remote sensing (cf. Chapter 7), a very powerful physically motivated BRDF model for planetary regolith surfaces has been introduced by Hapke (1981, 1984, 1986, 1993). For the purpose of three-dimensional surface reconstruction, however, it has been shown by McEwen (1991) that a very good approximation of the true reflectance behaviour over a wide range of incidence and emission angles and surface orientations is obtained by the phenomenological Lunar-Lambert law

$$f_{LL}(\rho, \theta_i, \theta_r, \alpha) = \rho \left[2L(\alpha) \frac{1}{\cos \theta_i + \cos \theta_e} + (1 - L(\alpha)) \right] \quad (2.16)$$

with $L(\alpha)$ as a phase angle dependent empirical factor tabulated by McEwen (1991). A more detailed discussion of the Hapke model and the Lunar-Lambert BRDF is given in Chapter 7.

2.2.2 Determination of Surface Gradients

2.2.2.1 Photoclinometry

Photoclinometric and shape from shading techniques take into account the geometric configuration of camera, light source, and the object itself, as well as the reflectance properties of the surface to be reconstructed. In some early work on photoclinometry the reconstruction problem is explored in the context of full perspective projection (Rindfleisch, 1966). This kind of presentation, however, tends to obscure the underlying principles of the reconstruction problem. Hence, we always assume parallel incident light and an infinite distance between camera and object (Horn, 1986), which is a good approximation in the application scenario of industrial quality inspection (cf. Chapter 5) and nearly exactly fulfilled in the remote sensing scenario (cf. Chapter 7). Under these conditions, the intensity I_{uv} of the image pixel located at (u, v) amounts to

$$I_{uv} = R_I(\rho, \mathbf{n}, \mathbf{s}, \mathbf{v}). \quad (2.17)$$

Eq. (2.17) is termed image irradiance equation. Here, ρ is the surface albedo, \mathbf{n} the surface normal, \mathbf{v} the direction to the camera, \mathbf{s} the direction of incident light, and R_I the so-called reflectance map. The reflectance map indicates the relationship between surface orientation and brightness (Horn, 1986), based on information about surface reflectance properties and the distribution of the light sources that illuminate the surface. In the example of a Lambertian surface illuminated by a light source of radiance E , the observed scene radiance is $L = (E/\pi) \cos \theta_i$. The factor $\cos \theta_i$ is of purely geometrical nature and originates from the foreshortening of the surface as

seen from the light source (cf. Eq. (2.7)). The factor ρ is a normalisation constant and absorbs several quantities such as the amount of light not reflected by the surface, the absolute irradiance of the light source, and the sensitivity of the camera sensor used to acquire the image of the surface. Hence, the Lambertian reflectance map is given by

$$R_I^{\text{Lambert}}(\rho, \mathbf{n}, \mathbf{s}) = \rho \cos \theta_i, \quad (2.18)$$

where the incidence angle θ_i corresponds to the angle between \mathbf{n} and \mathbf{s} with $\cos \theta_i = \mathbf{n} \cdot \mathbf{s} / (|\mathbf{n}| |\mathbf{s}|)$. In many practical applications, the Lambert model does not correspond very well to the true reflectance behaviour of the regarded surface. Hence, we will show in Sections 5.3 and 7.1 how realistic reflectance maps are obtained in the regarded application scenarios.

The surface is described by the function $z(x, y)$; in practice, however, it turns out to be advantageous to utilise the formulation z_{uv} defined at the discrete pixel positions (u, v) . In the following, the surface normal \mathbf{n} is represented in gradient space by the directional derivatives $p = \partial z / \partial x$ and $q = \partial z / \partial y$ of the surface function $z(x, y)$ with $\mathbf{n} = (-p, -q, 1)^T$. In an analogous manner we define $\mathbf{s} = (-p_s, -q_s, 1)^T$ and $\mathbf{v} = (-p_v, -q_v, 1)^T$. In the following we assume an infinite distance between the surface, the light source, and the camera, respectively. Accordingly, the reflectance map R_I can be expressed in terms of the surface gradients p_{uv} and q_{uv} at each pixel position and the constant vectors \mathbf{s} and \mathbf{v} , leading to the formulation $I_{uv} = R_I(\rho_{uv}, p_{uv}, q_{uv}, \mathbf{s}, \mathbf{v})$.

The angle α between the vectors \mathbf{s} and \mathbf{v} is termed phase angle. The angle between the vectors \mathbf{n} and \mathbf{s} is given by θ_i and the angle between \mathbf{n} and \mathbf{v} by θ_e , such that these angles can be expressed as

$$\begin{aligned} \cos \theta_i &= \frac{\mathbf{n} \cdot \mathbf{s}}{|\mathbf{n}| \cdot |\mathbf{s}|} = \frac{1 + p_s p_{uv} + q_s q_{uv}}{\sqrt{1 + p_s^2 + q_s^2} \sqrt{1 + p_{uv}^2 + q_{uv}^2}} \\ \cos \theta_e &= \frac{\mathbf{n} \cdot \mathbf{v}}{|\mathbf{n}| \cdot |\mathbf{v}|} = \frac{1 + p_v p_{uv} + q_v q_{uv}}{\sqrt{1 + p_v^2 + q_v^2} \sqrt{1 + p_{uv}^2 + q_{uv}^2}}. \end{aligned} \quad (2.19)$$

According to Eq. (2.17), we attempt to determine two variables p_{uv} and q_{uv} for each pixel from one single measurement, the pixel intensity I_{uv} . Without further assumptions about the surface, this is an ill-posed problem. The photoclinometric approach consists of computing height profiles along image rows under the assumptions that the terrain is gently sloping ($|p|, |q| \ll 1$), the illumination is highly oblique, and the scene is illuminated along the image rows, corresponding to $q_s = 0$. An early description of the principle of photoclinometry is given by Wilhelms (1964). As long as the reflectance is similar to the Lambert law (2.18), it depends much stronger on p than on q , such that we may set $q = 0$. This is a reasonable assumption especially when cross-sections of a linearly extended feature like a ridge, a graben, or the central section of a rotationally symmetric feature like a crater are regarded (Horn, 1989).

Often a constant albedo ρ is assumed, and the value of ρ is chosen such that the average surface slope over the region of interest is zero (Wöhler et al., 2006b)

or corresponds to a given nonzero value. Alternatively, a non-uniform albedo ρ_{uv} is determined based on a pair of images acquired under different illumination conditions according to Lena et al. (2006) (cf. Section 2.3.2). Eq. (2.17) is then solved for the surface gradient p_{uv} for each pixel with intensity I_{uv} . For each image row v , a height profile z_{uv} can be readily obtained by integration of the surface gradients p_{uv} .

2.2.2.2 Single-image Approaches with Regularisation Constraints

The classical shape from shading approach regarded in this section is based on the global optimisation of an energy function (Horn, 1986, 1989; Horn and Brooks, 1989). This method involves searching for two functions $p(x,y)$ and $q(x,y)$ which imply a surface that generates the observed image intensity $I(x,y)$. The original problem formulation is expressed in the continuous variables x and y , resulting in a variational framework. Here we will, however, immediately deal with finite sums over the image pixels, the positions of which are denoted by the discrete variables u and v , and rewrite the error integrals introduced by Horn (1986) accordingly. Hence, the intensity constraint can be expressed by the minimisation of an intensity error term e_i with

$$e_i = \sum_{u,v} [I_{uv} - R_I(p_{uv}, q_{uv})]^2 \quad (2.20)$$

with $R_I(p_{uv}, q_{uv})$ as the reflectance function of the regarded surface. It is straightforward to extend this error term to two or more light sources (Section 2.3). This section, however, concentrates on the single light source scenario. As the correspondingly defined reconstruction problem is ill-posed, we furthermore introduce a regularisation constraint e_s which requires local continuity of the surface. Such a smooth surface implies that the absolute values of the derivatives $\partial p / \partial x$, $\partial p / \partial y$, $\partial q / \partial x$, and $\partial q / \partial y$ are small, which results in an error term e_s with

$$e_s = \sum_{u,v} \left[\left\{ \frac{\partial p}{\partial x} \right\}_{uv}^2 + \left\{ \frac{\partial p}{\partial y} \right\}_{uv}^2 + \left\{ \frac{\partial q}{\partial x} \right\}_{uv}^2 + \left\{ \frac{\partial q}{\partial y} \right\}_{uv}^2 \right]. \quad (2.21)$$

This leads to a minimisation of the overall error

$$e = e_s + \lambda e_i, \quad (2.22)$$

where the Lagrange multiplier λ denotes the relative weight of the two error terms e_i and e_s . With the approximations

$$\begin{aligned} \left\{ \frac{\partial p}{\partial x} \right\}_{uv} &= \frac{1}{2} (p_{u+1,v} - p_{u-1,v}) \\ \left\{ \frac{\partial p}{\partial y} \right\}_{uv} &= \frac{1}{2} (p_{u,v+1} - p_{u,v-1}) \end{aligned}$$

$$\begin{aligned}\left\{\frac{\partial q}{\partial x}\right\}_{uv} &= \frac{1}{2}(q_{u+1,v} - q_{u-1,v}) \\ \left\{\frac{\partial q}{\partial y}\right\}_{uv} &= \frac{1}{2}(q_{u,v+1} - q_{u,v-1})\end{aligned}\quad (2.23)$$

and the average values

$$\begin{aligned}\bar{p}_{uv} &= (p_{u+1,v} + p_{u-1,v} + p_{u,v+1} + p_{u,v-1})/4 \\ \bar{q}_{uv} &= (q_{u+1,v} + q_{u-1,v} + q_{u,v+1} + q_{u,v-1})/4\end{aligned}\quad (2.24)$$

we obtain an iterative update rule for p_{uv} and q_{uv} by setting the derivatives of e with respect to p_{uv} and q_{uv} to zero (Horn, 1986):

$$\begin{aligned}p_{uv}^{(n+1)} &= \bar{p}_{uv}^{(n)} + \lambda \left(I_{uv} - R_I(\bar{p}_{uv}^{(n)}, \bar{q}_{uv}^{(n)}) \right) \frac{\partial R_I}{\partial p} \Big|_{\bar{p}_{uv}^{(n)}, \bar{q}_{uv}^{(n)}} \\ q_{uv}^{(n+1)} &= \bar{q}_{uv}^{(n)} + \lambda \left(I_{uv} - R_I(\bar{p}_{uv}^{(n)}, \bar{q}_{uv}^{(n)}) \right) \frac{\partial R_I}{\partial q} \Big|_{\bar{p}_{uv}^{(n)}, \bar{q}_{uv}^{(n)}}\end{aligned}\quad (2.25)$$

The initial values $p_{uv}^{(0)}$ and $q_{uv}^{(0)}$ must be provided based on a-priori knowledge about the surface. The surface profile z_{uv} is then derived from the slopes p_{uv} and q_{uv} by means of numerical integration as outlined in detail in Section 2.2.3. The albedo ρ_{uv} is assumed to be uniform over the image and updated using Eqs. (2.17) and (2.19) in each iteration step based on a certain number of selected pixels (e.g. all pixels of a certain image column)—hence, the iterative update rule (2.25) not only determines the surface gradients p_{uv} and q_{uv} but also the albedo ρ by minimisation of error function (2.22). Section 2.3 describes how a non-uniform albedo ρ_{uv} is taken into account.

A reconstruction algorithm which generates an integrable surface gradient vector field is described by Horn (1989). It simultaneously yields the surface gradients p_{uv} and q_{uv} and the depth z_{uv} . Here, the assumption of a smooth surface according to Eq. (2.21) is replaced by the departure from integrability error expressed by the error term

$$e_{\text{int}} = \sum_{uv} \left[\left(\left\{ \frac{\partial z}{\partial x} \right\}_{uv} - p_{uv} \right)^2 + \left(\left\{ \frac{\partial z}{\partial y} \right\}_{uv} - q_{uv} \right)^2 \right]. \quad (2.26)$$

Accordingly, the shape from shading problem corresponds to a minimisation of the overall error term

$$f = e_i + \gamma e_{\text{int}}. \quad (2.27)$$

In the continuous formulation, satisfying the integrability constraint (2.26) corresponds to the variational problem of minimising the functional

$$\iint \left[\left(\frac{\partial z(x,y)}{\partial x} - p(x,y) \right)^2 + \left(\frac{\partial z(x,y)}{\partial y} - q(x,y) \right)^2 \right] dx dy \quad (2.28)$$

with respect to the surface gradients $p(x,y)$ and $q(x,y)$. The Euler equation of this problem is given by

$$\nabla^2 z = \frac{\partial p}{\partial x} + \frac{\partial q}{\partial y}, \quad (2.29)$$

where $\nabla^2 z$ denotes the Laplacian of z (Horn, 1986). Eq. (2.29) must hold at each position in the image. On a discrete pixel grid, the Laplacian can be approximated by the expression

$$\{\nabla^2 z\}_{uv} \approx \frac{\kappa}{\varepsilon^2} (\bar{z}_{uv} - z_{uv}), \quad (2.30)$$

where ε is the spacing between the pixels (it is convenient to set $\varepsilon = 1$) and $\kappa = 4$ when the local average \bar{z}_{uv} is computed using the four edge-adjacent neighbours (Horn, 1986). Setting the derivative of the error term f with respect to p_{uv} and q_{uv} to zero and combining Eqs. (2.29) and (2.30) then yields the following iteration scheme:

$$\begin{aligned} p_{uv}^{(n+1)} &= \left\{ \frac{\partial z}{\partial x} \right\}_{uv}^{(n)} + \frac{1}{\gamma} (I - R_I) \frac{\partial R_I}{\partial p} \\ q_{uv}^{(n+1)} &= \left\{ \frac{\partial z}{\partial y} \right\}_{uv}^{(n)} + \frac{1}{\gamma} (I - R_I) \frac{\partial R_I}{\partial q} \\ z_{uv}^{(n+1)} &= \bar{z}_{uv} - \frac{\varepsilon^2}{\kappa} \left(\left\{ \frac{\partial p}{\partial x} \right\}_{uv}^{(n+1)} + \left\{ \frac{\partial q}{\partial y} \right\}_{uv}^{(n+1)} \right). \end{aligned} \quad (2.31)$$

After each update of the surface gradients p_{uv} and q_{uv} , the corresponding height map z_{uv} is computed by means of a discrete approximation to the solution of the Euler-Lagrange differential equations of the corresponding variational problem. For a single light source and oblique illumination, this algorithm gives a good estimate even of the surface gradients perpendicular to the direction of incident light, as it adjusts them to the integrability constraint without affecting the intensity error e_i given by (2.20).

Even if the iterative update rule (2.25) is initialised with a solution that perfectly fits with the observed pixel intensities, i.e. $e_i = 0$, the algorithm will nevertheless yield a different, smoother surface, since the constraint of small partial derivatives of the surface is a strong restriction and not always physically reasonable. In contrast, the integrability constraint is a much weaker restriction and always physically correct. If the corresponding iterative update rule (2.31) is initialised with a surface for which the associated intensity error e_i is zero, the algorithm will retain this “perfect” solution. However, the algorithm based on the integrability constraint has the drawback of a small convergence radius, such that it needs to be initialised with a surface profile which is already close to the final solution (Horn, 1989).

In the single-image shape from shading scenario, the surface albedo ρ_{uv} has always been regarded as known. If this is not the case, in many applications the assumption of a uniform albedo ρ is made. Different albedo values yield different solutions of the shape from shading problem, since e.g. increasing ρ will result in a

surface inclined away from the light source and vice versa. Hence, it is often necessary to make additional assumptions about the surface, e.g. that the average surface slope is zero or obtains a predefined value. In Section 2.3 methods to cope with unknown and non-uniform surface albedos are described.

2.2.3 Reconstruction of Height from Gradients

Local techniques for computation of height from gradient rely on curve integrals and are based on specifying an integration path and a local neighbourhood. According to a technique described by Jiang and Bunke (1997), reconstruction of height is started at a given point (u_0, v_0) of the image, e.g. the centre, for which $z_{u_0, v_0} = 0$ is assumed, and the initial paths are forming a cross along image column u_0 and image row v_0 (cf. also Klette and Schlüns (1996)). The image origin is in the upper left corner. For the upper right quadrant, the height value z_{uv} is obtained according to

$$z_{uv} = \frac{1}{2} \left[z_{u-1,v} + \frac{1}{2} (p_{u-1,v} + p_{uv}) + z_{u,v+1} + \frac{1}{2} (q_{u,v+1} + q_{uv}) \right] \quad (2.32)$$

Analogous relations for the remaining three quadrants are readily obtained. In Eq. (2.32) deviations of the surface gradient field from integrability are accounted for by averaging over the surface gradients in horizontal and vertical direction. A drawback of this method is that the resulting height map z_{uv} depends on the initial location (u_0, v_0) .

In a more systematic way than in the rather elementary approach by Jiang and Bunke (1997), the three-dimensional reconstruction scheme based on the integrability error outlined in Section 2.2.2.2 can be used for adapting a surface to the generally non-integrable surface gradient field obtained by shape from shading. It is desired to obtain a surface z_{uv} with partial derivatives $\{\partial z / \partial x\}_{uv}$ and $\{\partial z / \partial y\}_{uv}$ which come as close as possible to the values p_{uv} and q_{uv} previously obtained by shape from shading, which are assumed to be known and fixed. The height map z_{uv} is chosen such that the integrability error (2.26) is minimised. Hence, Eqs. (2.29) and (2.30) directly yield an iterative scheme to determine the height map z_{uv} (cf. also Eq. (2.31)):

$$z_{uv}^{(n+1)} = \bar{z}_{uv}^{(n)} - \frac{\varepsilon^2}{\kappa} \left(\left\{ \frac{\partial p}{\partial x} \right\}_{uv} + \left\{ \frac{\partial q}{\partial y} \right\}_{uv} \right). \quad (2.33)$$

The main drawback of this variational approach is the large number of iterations necessary until convergence is achieved. Hence, Frankot and Chellappa (1988) propose a method to enforce integrability of the surface using the theory of projection onto convex sets. The given, generally non-integrable gradient field is projected onto the integrable gradient field which is nearest in the least-mean-squares sense. This concept is extended by Simchony et al. (1990) to solving the Poisson equation $\nabla^2 z = f$ with the Dirichlet boundary condition that z is known on the boundary of

the surface. The basic step consists of computing the Fourier transform of Eq. (2.29), which yields

$$Z_{\omega_u \omega_v} = -\frac{i\omega_u P_{\omega_u \omega_v} + i\omega_v Q_{\omega_u \omega_v}}{\omega_u^2 + \omega_v^2}, \quad (2.34)$$

where $i = \sqrt{-1}$, and $P_{\omega_u \omega_v}$, $Q_{\omega_u \omega_v}$, and $Z_{\omega_u \omega_v}$ are the Fourier transforms of p_{uv} , q_{uv} , and z_{uv} . The corresponding height map z_{uv} is readily obtained by computing the inverse Fourier transform of $Z_{\omega_u \omega_v}$.

The Fourier-based approach is extended by Wei and Klette (2004) towards a “strong integrability” error term that additionally takes into account the differences between the second-order derivatives of z_{uv} and the first-order derivatives of p_{uv} and q_{uv} according to

$$e_{\text{int}}^{\text{strong}} = e_{\text{int}} + \delta \sum_{u,v} \left[\left(\left\{ \frac{\partial^2 z}{\partial x^2} \right\}_{uv} - \left\{ \frac{\partial p}{\partial x} \right\}_{uv} \right)^2 + \left(\left\{ \frac{\partial^2 z}{\partial y^2} \right\}_{uv} - \left\{ \frac{\partial q}{\partial y} \right\}_{uv} \right)^2 \right], \quad (2.35)$$

where δ is a weight factor and the “weak integrability” error term e_{int} is given by Eq. (2.26). According to Wei and Klette (2004), minimisation of the error term (2.35) is performed based on a Fourier transform in the continuous image domain, which yields the expression

$$Z_{\omega_u \omega_v} = -\frac{i(\omega_u + \delta \omega_u^3)P_{\omega_u \omega_v} + i(\omega_v + \delta \omega_v^3)Q_{\omega_u \omega_v}}{\omega_u^2 + \omega_v^2 + \delta(\omega_u^4 + \omega_v^4)} \quad (2.36)$$

for the Fourier-transformed depth map $Z_{\omega_u \omega_v}$. Again, the height map z_{uv} is obtained by computing the inverse Fourier transform of $Z_{\omega_u \omega_v}$.

Agrawal et al. (2005) propose an algebraic approach to the reconstruction of height from gradients which exploits the information contained in the curl of the given non-integrable vector field, denoted by $\partial p/\partial y - \partial q/\partial x$. The curl of a vector field denotes the deviation from integrability. Integrability is enforced by finding a residual gradient field, which is added to the computed gradients p_{uv} and q_{uv} such that an integrable gradient field results. Agrawal et al. (2005) show that the method by Simchony et al. (1990) enforces integrability by finding a zero curl gradient field which has the same divergence $\partial p/\partial x - \partial q/\partial y$ as the given non-integrable gradient field. In principle, the residual gradient field can be obtained by solving a set of linear equations, which is, however, underdetermined in shape from shading applications since the curl tends to be nonzero almost everywhere. Hence, the number of unknown residual gradient values is about twice the number of pixels. Instead of solving for a minimum norm least squares solution, as suggested by Simchony et al. (1990), it is proposed by Agrawal et al. (2005) to reduce the number of unknowns by considering the surface gradients at image locations with curl values lower than a given threshold as error-free. A graph-based optimisation scheme is presented for recovering the maximum number of erroneous gradients.

In the applications of three-dimensional surface reconstruction techniques to real-world problems described in this work (cf. Chapters 5 and 7) we utilise a computationally efficient implementation of the approach by Simchony et al. (1990)

based on enforcing the “weak integrability” constraint (2.26). Any physically reasonable surface z_{uv} must satisfy this constraint. In our application scenarios, computational efficiency is relevant since the employed methods require a reconstruction of the height map z_{uv} from the gradients p_{uv} and q_{uv} at many different intermediate stages of the optimisation algorithm.

2.2.4 Surface Reconstruction Based on Eikonal Equations

The solution of the single-image shape from shading problem is generally not unique. For a point in the image plane, the surface gradient is generally constrained by the image irradiance equation (2.17) to a one-parameter manifold (Bruss, 1989) if the surface albedo is assumed to be constant. In the mathematical framework discussed in this section, the observed image intensities are assumed to be represented by a continuous function $I(x,y)$ instead of the discrete pixel grey values I_{uv} . For shape from shading problems involving a reflectance function of the special form $R(p,q) = p^2 + q^2$, the image irradiance equation (2.17) becomes

$$p^2 + q^2 = I(x,y) \quad (2.37)$$

which is referred to as an eikonal equation. Under certain conditions, which are derived in detail by Bruss (1989), unique solutions for the reconstructed surface can be obtained from a single image based on Eq. (2.37).

The eikonal equation (2.37) is defined to be constrained if $I(x,y)$ is a C^3 function (i.e. it is three times continuously differentiable) which satisfies the following conditions in a neighbourhood of the point (x_0, y_0) :

1. The point (x_0, y_0) is a stationary point of $I(x,y)$ with $\partial I / \partial x = \partial I / \partial y = 0$.
2. $I(x_0, y_0) = 0$
3. $I(x,y) > 0$ for $(x,y) \neq (x_0, y_0)$
4. $I(x,y)$ vanishes precisely to second order at (x_0, y_0) .

The point $P = (x, y, p, q) = (x_0, y_0, 0, 0)$ is a critical point of a constrained eikonal equation, i.e. (x_0, y_0) is a stationary point of $I(x,y)$, $(0,0)$ is a stationary point of $R(p,q) = p^2 + q^2$, and the values $(x_0, y_0, 0, 0)$ satisfy Eq. (2.37). It is demonstrated by Bruss (1989) that for a constrained eikonal equation a unique locally convex solution for the surface gradients $p(x,y)$ and $q(x,y)$ and thus the surface $z(x,y)$ exists in a neighbourhood of (x_0, y_0) . Hence, the point (x_0, y_0) is a singular point of the constrained eikonal equation. If $I(x,y)$ is a C^∞ function, there exists a unique, locally convex power series solution to the constrained eikonal equation in a neighbourhood of the singular point.

An image irradiance equation according to Eq. (2.17) is defined to be singular if there exist finite values for x and y denoted by x_0 and y_0 for which

$$\lim_{(x,y) \rightarrow (x_0, y_0)} I(x,y) = \pm\infty. \quad (2.38)$$

The set of points (x_0, y_0) is termed b-silhouette. According to Bruss (1989), it is assumed that $R(p, q) = p^2 + q^2$, the b-silhouette is a closed, smooth curve in the xy plane, the points (x, y) at which the image irradiance equation is defined lie inside the b-silhouette, and the function $I(x, y)$ has exactly one stationary point (x_0, y_0) and additionally satisfies conditions 2–4 stated above. If a C^2 surface $z(x, y)$ exists which satisfies the image irradiance equation, the only other solution to this equation is $-z(x, y)$.

Reflectance functions of the form $R(p, q) = f(p^2 + q^2)$ with f as a bijection yield image irradiance equations that can be transformed into a form equivalent to Eq. (2.37). A practically relevant scenario, in which all conditions are fulfilled under which a unique solution of the image irradiance equation is obtained, is a Lambertian surface illuminated by a light source situated at the same position as the camera (Bruss, 1989; Kimmel and Sethian, 2001). In this special case, $I(x, y)$ is given by $I(x, y) = 1/\sqrt{1 + p^2 + q^2}$, leading to the eikonal equation

$$\|\nabla z(x, y)\| = \sqrt{\frac{1}{I(x, y)^2} - 1}. \quad (2.39)$$

For orthographic projection, i.e. infinite distance between the surface and the camera, and parallel incident light this situation corresponds to the zero phase angle ($\alpha = 0$) case.

Numerical solutions for this formulation of the shape from shading problem have e.g. been proposed by Rouy and Tourin (1992) and Kimmel and Bruckstein (1995). More recently, the shape from shading approach based on the eikonal equation has been extended to the more general scenario of a Lambertian surface observed under oblique illumination by Kimmel and Sethian (2001). They introduce a numerical approximation of an equation of the form $\|\nabla z(x, y)\| = f(x, y)$ according to

$$[\max(D_{uv}^{-x}z, -D_{uv}^{+x}z, 0)]^2 + [\max(D_{uv}^{-y}z, -D_{uv}^{+y}z, 0)]^2 = f_{uv}^2, \quad (2.40)$$

where $z_{uv} = z(u\Delta x, v\Delta y)$ is the discretised version of $z(x, y)$ defined on the pixel raster, $D_{uv}^{-x}z = (z_{uv} - z_{u-1,v})/\Delta x$ the backwards approximation of the derivative in x direction, and $D_{uv}^{+x}z = (z_{u+1,v} - z_{uv})/\Delta x$ the corresponding forward approximation (and analogous for the y direction). Rouy and Tourin (1992) show that this approximation selects the correct solution for the shape from shading problem. Information always flows from small to large values of the solution z . Hence, Kimmel and Sethian (2001) suggest to initialise all z values to infinity except at the local minimum points, where they are initialised with the correct height value. If only a single known minimum point exists, the surface is initialised with $z_{uv} = \infty$ except for the minimum point, which is set to zero. The update step for z_{uv} is then written as follows:

1. Let $z_1 = \min(z_{u-1,v}, z_{u+1,v})$ and $z_2 = \min(z_{u,v-1}, z_{u,v+1})$.
2. If $|z_1 - z_2| < f_{uv}$ then $z_{uv} = \frac{1}{2} (z_1 + z_2 + \sqrt{2f_{uv}^2 - (z_1 - z_2)^2})$
else $z_{uv} = \min(z_1, z_2) + f_{uv}$.

Kimmel and Sethian (2001) show that the computational complexity of this scheme is between $\mathcal{O}(N)$ and $\mathcal{O}(N^2)$ depending on the surface, where N is the number of pixels. Furthermore, they introduce a fast marching technique relying on the systematic causality relationship based on upwinding, combined with a heap structure for efficiently ordering the updated points (Sethian, 1999). This approach yields a worst case complexity of $\mathcal{O}(N \log N)$. The proposed scheme can be used directly to solve the shape from shading problem in the zero phase angle case.

In the more general scenario of a Lambertian surface illuminated by a light source located in a direction which is not identical to that of the camera, the shading image is given by the scalar product $I(x, y) = \mathbf{s} \cdot \mathbf{n}$, where it is assumed that the albedo corresponds to unity and $\|\mathbf{s}\| = \|\mathbf{n}\| = 1$. Without loss of generality, the value of s_y is set to zero. However, the described numerical solution scheme for the eikonal equation cannot be used directly to reconstruct the surface. Hence, the observed image is transformed into the coordinate system of the light source, leading to the expression

$$\tilde{p}^2 + \tilde{q}^2 = \frac{1}{\tilde{I}(\tilde{x}, y)^2} - 1 \quad (2.41)$$

which formally corresponds to the eikonal equation (2.39). In Eq. (2.41), \tilde{p} , \tilde{q} , \tilde{x} , and \tilde{I} are defined in the light source coordinate system. Due to the fact that the observed brightness of a Lambertian surface is independent of the viewing direction, the relation between the original and the transformed image brightness is fairly simple and amounts to

$$\tilde{I}(\tilde{x}, y) = I(s_z \tilde{x} + s_x \tilde{z}, y). \quad (2.42)$$

Inserting Eq. (2.42) into Eq. (2.41) allows to determine the surface height values z_{uv} . Alternatively, Kimmel and Sethian (2001) propose an extended fast marching scheme that directly yields the solution for z_{uv} .

The main advantage of the shape from shading approach outlined in this section, compared to the variational method described in Section 2.2.2, is the fact that it does not require a smoothness or integrability constraint but directly yields a reconstruction of the surface exclusively based on the observed pixel grey values. It is, however, restricted to Lambertian surfaces, which is a major drawback in real-world scenarios (cf. Chapters 5 and 7) that generally involve more complex reflectance functions. Furthermore, a-priori knowledge about the location of a local minimum of the surface, which is a precondition for the determination of the unique solution, is not necessarily available.

2.3 Photometric Stereo

The solution of shape from shading based on a single image is ambiguous as long as the surface albedo is unknown and no assumption about the surface can be made. Furthermore, for oblique illumination and reflectance maps similar to the Lambert law (2.18), the surface gradients perpendicular to the direction of incident light are

much less well-defined than those in the direction of incident light. These drawbacks can be overcome by the analysis of several images, a procedure termed photometric stereo in analogy to standard geometric stereo image analysis techniques.

2.3.1 Classical Photometric Stereo Approaches

The most straightforward way to extend the shape from shading method outlined in Section 2.2.2.2 is to acquire a set of L pixel-synchronous images of the surface under known illumination conditions described by the vectors \mathbf{s}_l . The intensity error is then defined as a sum over all L images according to

$$e_i = \sum_{l=1}^L \sum_{u,v} \left[I_{uv}^{(l)} - R_I(\mathbf{s}_l, p_{uv}, q_{uv}) \right]^2 \quad (2.43)$$

while the smoothness constraint (2.21) and the integrability constraint (2.26) remain unchanged. The corresponding iterative update schemes for the surface gradients are analogous to Eq. (2.25) or (2.31). In this setting we obtain reliable information about the surface gradients in all directions already with $L = 2$ light sources, but still a known uniform albedo ρ has to be assumed. In principle, once the value of ρ is determined, it is possible to obtain without further constraints the surface gradients p_{uv} and q_{uv} from the $L = 2$ intensities available for each pixel. In many applications, however, it is advantageous to keep on assuming a smooth surface, as this prevents the surface from being strongly influenced by noise or small-scale specular reflections.

Even recent photometric stereo approaches such as multi-image shape from shading (Lohse and Heipke, 2003, 2004; Lohse et al., 2006) assume a uniform surface albedo. This technique is based on the direct reconstruction of the surface in an object-centred coordinate system. It is used to obtain digital elevation models of the lunar polar regions based on images acquired by the Clementine spacecraft. Due to the fact that nearly all images taken by this spacecraft were acquired at local lunar noon, the method makes use of image pairs acquired under virtually identical incidence angles but different emission angles, exploiting the comparably weak influence of the emission angle θ_e on the image brightness under oblique illumination rather than the strong influence of incidence angle. This approach is feasible due to the non-Lambertian reflectance behaviour of the lunar surface, but it requires accurate knowledge about the reflectance map and a precise absolute radiometric calibration of the image data. What is more, the algorithm requires a very good initialisation, which has to be obtained by stereophotogrammetry, and is thus only suitable for strongly textured surface regions.

The classical approach to photometric stereo for surfaces of non-uniform albedo is the method introduced by Woodham (1980) which relies on three pixel-synchronous images of the surface acquired under different illumination conditions (cf. also Horn (1986) and Klette et al. (1996)), defined by the illumination vectors \mathbf{s}_l

with $\|\mathbf{s}_l\| = 1$ and $l \in \{1, 2, 3\}$. The surface is assumed to display a Lambertian reflectance behaviour according to Eq. (2.18) and a non-uniform albedo ρ_{uv} . The light sources are situated at infinite distance and their irradiances are identical. For each pixel (u, v) , three intensity values are measured and expressed as the three components of the vector \mathbf{I}_{uv} . If we define the surface normal \mathbf{n}_{uv} as a unit vector, we find that the pixel intensities are given by

$$\mathbf{I}_{uv} = \rho_{uv} S \cdot \mathbf{n}_{uv}, \quad (2.44)$$

where the rows of the 3×3 matrix S contain the illumination vectors \mathbf{s}_l . Due to the unit length of \mathbf{n}_{uv} , Eq. (2.44) immediately yields

$$\rho_{uv} = \|S^{-1}\mathbf{I}_{uv}\|, \quad (2.45)$$

and the surface normal amounts to

$$\mathbf{n}_{uv} = \frac{1}{\rho_{uv}} S^{-1}\mathbf{I}_{uv}. \quad (2.46)$$

This algorithm has the considerable advantage that it copes with surfaces of arbitrary non-uniform surface albedo. However, a drawback is that the inverse matrix S^{-1} exists only if the three illumination vectors \mathbf{s}_l are not coplanar, and that it is restricted to surfaces of purely Lambertian reflectance.

2.3.2 Photometric Stereo Approaches Based on Ratio Images

This section regards photometric stereo approaches which are suited for non-uniform surface albedos, coplanar illumination vectors (which may be of high practical relevance especially in the field of remote sensing applications, as pointed out by Wöhler and Hafezi (2005) and Lena et al. (2006)), and a much more general class of reflectance characteristics.

2.3.2.1 Ratio-based Photoclinometry of Surfaces with Non-uniform Albedo

Similar to the scenario of single-image photoclinometry (cf. Section 2.2.2.1) we assume that the scene is illuminated along the image rows ($q_s = 0$). For oblique illumination, the dependence of the reflectance map on p is much more pronounced than the dependence on q , provided that it has no strongly specular component. Hence, we again approximate the surface gradient q perpendicular to the direction of incident light with zero values.

Two images of the surface acquired under different illumination angles are required to separate intensity variations due to topographic relief from those due to albedo variations. The images have to be pixel-synchronous, which is achieved by

an image registration step (Gottesfeld Brown, 1992). We do not have to restrict ourselves to Lambertian reflectance, instead we assume to have a reflectance map of the form

$$R_I(\rho_{uv}, \mathbf{s}, p_{uv}, q_{uv}) = \rho_{uv} \tilde{R}_I(\mathbf{s}, p_{uv}, q_{uv}). \quad (2.47)$$

Photoclinometry is then performed along image rows by extending Eq. (2.17) as described by McEwen (1985) and determining p_{uv} such that

$$\frac{I_{uv}^{(1)}}{I_{uv}^{(2)}} = \frac{\tilde{R}_I(\mathbf{s}_1, p_{uv}, q_{uv})}{\tilde{R}_I(\mathbf{s}_2, p_{uv}, q_{uv})}. \quad (2.48)$$

The surface gradients q_{uv} are still kept zero, and the albedo ρ_{uv} cancels out, such that Eq. (2.48) directly yields the surface gradient p_{uv} individually for each pixel. It is generally not possible to obtain an analytical solution for p_{uv} , such that numerical techniques like the Newton method have to be used.

When the images are not absolutely calibrated radiometrically, their average pixel intensities have to be adjusted according to the different illumination angles, relying on the assumption that on the average the surface is flat. As long as $\tilde{R}_I(p, q)$ is not strongly nonlinear in p and q , it is sufficient to normalise $I_{uv}^{(1)}$ by multiplying its pixel intensities with the factor

$$\frac{\tilde{R}_I(\mathbf{s}_1, 0, 0) \langle I_{uv}^{(2)} \rangle_{u,v}}{\tilde{R}_I(\mathbf{s}_2, 0, 0) \langle I_{uv}^{(1)} \rangle_{u,v}}.$$

The non-uniform surface albedo ρ_{uv} is then recovered by

$$\rho_{uv} = \frac{1}{L} \sum_{l=1}^L \frac{I_{uv}^{(l)}}{\tilde{R}_I(\mathbf{s}_l, p_{uv}, q_{uv})}. \quad (2.49)$$

In the next step, the albedo map ρ_{uv} is inserted into one of the single-image shape from shading schemes described in Section 2.2.2, preferably relying on the image acquired under the more oblique illumination conditions in order to extract the relief as accurately as possible. The surface gradients p_{uv} determined based on Eq. (2.48) are used as initial values for the corresponding iterative update rule (2.25) or (2.31). Accordingly, p_{uv} hardly changes in the course of the iteration process while q_{uv} obtains values consistent with the smoothness constraint (2.21) or the integrability constraint (2.26).

For strongly non-Lambertian surfaces it may be necessary to repeat the ratio-based photoclinometry step, now setting q_{uv} in Eq. (2.48) to the values obtained by the shape from shading scheme. New values are then obtained for p_{uv} and subsequently for ρ_{uv} (cf. Eq. (2.49)), with which the single-image shape from shading scheme is initialised, and so on. Eventually this iterative procedure yields a self-consistent solution for the surface gradients p_{uv} and q_{uv} and the non-uniform surface albedo ρ_{uv} .

2.3.2.2 Ratio-based Variational Photometric Stereo Approach

Another approach to cope with a non-uniform surface albedo ρ_{uv} is to return to the single-image shape from shading schemes in Section 2.2.2 and to replace the single-image intensity error term (2.20) by the modified ratio-based error term

$$\tilde{e}_i = \sum_{u,v} \left[\frac{I_{uv}^{(1)} \tilde{R}_I(\mathbf{s}_2, p_{uv}, q_{uv})}{I_{uv}^{(2)} \tilde{R}_I(\mathbf{s}_1, p_{uv}, q_{uv})} - 1 \right]^2 \quad (2.50)$$

(Wöhler and Hafezi, 2005). This leads to the ratio-based iterative update rule

$$\begin{aligned} p_{uv}^{(n+1)} &= \bar{p}_{uv}^{(n)} + \lambda \left(\frac{I_{uv}^{(1)} \tilde{R}_I(\mathbf{s}_2, \bar{p}_{uv}^{(n)}, \bar{q}_{uv}^{(n)})}{I_{uv}^{(2)} \tilde{R}_I(\mathbf{s}_1, \bar{p}_{uv}^{(n)}, \bar{q}_{uv}^{(n)})} - 1 \right) \frac{I_{uv}^{(1)}}{I_{uv}^{(2)}} \frac{\partial}{\partial p} \frac{\tilde{R}_I(\mathbf{s}_2, p_{uv}, q_{uv})}{\tilde{R}_I(\mathbf{s}_1, p_{uv}, q_{uv})} \Big|_{\bar{p}_{uv}^{(n)}, \bar{q}_{uv}^{(n)}} \\ q_{uv}^{(n+1)} &= \bar{q}_{uv}^{(n)} + \lambda \left(\frac{I_{uv}^{(1)} \tilde{R}_I(\mathbf{s}_2, \bar{p}_{uv}^{(n)}, \bar{q}_{uv}^{(n)})}{I_{uv}^{(2)} \tilde{R}_I(\mathbf{s}_1, \bar{p}_{uv}^{(n)}, \bar{q}_{uv}^{(n)})} - 1 \right) \frac{I_{uv}^{(1)}}{I_{uv}^{(2)}} \frac{\partial}{\partial q} \frac{\tilde{R}_I(\mathbf{s}_2, p_{uv}, q_{uv})}{\tilde{R}_I(\mathbf{s}_1, p_{uv}, q_{uv})} \Big|_{\bar{p}_{uv}^{(n)}, \bar{q}_{uv}^{(n)}} \end{aligned} \quad (2.51)$$

The non-uniform albedo ρ_{uv} cancels out and can be recovered by Eq. (2.49) after determination of the surface gradients p_{uv} and q_{uv} . The properties of the results obtained with the ratio-based iterative scheme (2.51) are comparable to those obtained by single-image shape from shading analysis, except that one is not restricted to surfaces of uniform albedo. For more than two light sources and images ($L > 2$), the error term (2.50) can be extended to a sum over all $L(L-1)/2$ possible pairs of images, which reveals surface gradients in all directions if the light sources are appropriately distributed.

A drawback of the presented method is the fact that it has to be initialised with a surface which is already close to the final solution, since otherwise the algorithm diverges or gets stuck in local minima. Hence, as long as the albedo variations are not strong, it is advantageous to combine the two shape from shading approaches described in this section as follows: First, the surface profile is reconstructed using the multi-image intensity error (2.43), resulting in values for both p_{uv} and q_{uv} and a uniform albedo ρ . As a second step, the iterative update rule (2.51) is initialised with the results of the first step and then started. This procedure changes the surface profile only at the locations of albedo variations. The albedo map ρ_{uv} is then obtained according to Eq. (2.49). As a third step, p_{uv} and q_{uv} are recomputed according to Eq. (2.43). These three steps are repeated until p_{uv} , q_{uv} , and ρ_{uv} converge towards a self-consistent solution. Experimental results obtained with this approach are described in detail in Section 5.3.

2.4 Shape from Polarisation

An important drawback of the single-image shape from shading approach is the fact that an ill-posed problem has to be solved, since it is desired to determine two surface gradients (and sometimes also an albedo value) for each pixel based on a single intensity measurement. This drawback is overcome either by introducing constraints on the reconstructed surface, such as knowledge about the surface at the boundaries, smoothness, or uniform albedo, or by regarding several images of the surface acquired under different illumination conditions.

A further possibility to transform the ill-posed problem into a well-posed one is offered by extending the measurement of intensities towards the additional measurement of polarisation properties of the light reflected from the surface. This approach yields up to three measurements per pixel, i.e. intensity, degree of polarisation, and direction of polarisation, which may allow an unambiguous reconstruction of the surface gradient field under the conditions typically encountered in real-world scenarios.

2.4.1 Surface Orientation from Dielectric Polarisation Models

In many shape from shading algorithms a Lambertian reflectance map is explicitly assumed, as it is e.g. the case for the photometric stereo method described in Section 2.3. The computed surface shape is inaccurate when specular reflections are present in the image. An early approach to incorporate polarisation information into the shape from shading framework consists of the detection and subsequent removal of specular reflections in the image based on the analysis of colour and polarisation data (Nayar et al., 1993). They use the degree of polarisation to select pixels characterised by specular reflection and derive a technique to locally determine the colour of the specular component. This allows to constrain the colour of the diffuse component to a one-dimensional linear colour subspace, which is used to find neighbouring pixels with consistent colours. The result is an image of the surface that displays only the diffuse (Lambertian) component of the reflected light.

The existing methods that aim for a reconstruction of surface shape from polarisation measurements deal with dielectrical materials. The term polarisation image has been introduced by Wolff (1987), who presents a method to recover surface orientation by polarisation. Here, the illumination and viewing direction as well as the polarisation properties of the surface have to be known. This approach is extended towards the estimation of surface orientation by polarisation analysis using a stereo camera setup by Wolff (1989).

In the approach by Rahmann (1999), a polarisation image is generated by acquiring images of the surface with a camera equipped with a linear polarising filter. The intensity of a pixel depends on the rotation angle ω of the polarising filter according to

$$I(\omega) = I_c + I_v \cos [2(\omega - \Phi)]. \quad (2.52)$$

In Eq. (2.52), I_c is equal to half the overall intensity made up by the polarised and the unpolarised component. The polarisation degree D_p is defined as

$$D_p = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} = \frac{I_v}{I_c}. \quad (2.53)$$

The rotation angle of the polarising filter for which maximum intensity I_{\max} is observed corresponds to the polarisation angle Φ . To determine the polarisation image, at least three pixel-synchronous images of the surface are acquired. For each pixel, Eq. (2.52) is fitted to the measured pixel intensities, which yields the parameters I_c , I_v , and Φ .

Atkinson and Hancock (2005a) reconstruct the orientation of dielectric surfaces based on Fresnel theory. In this scenario, the refraction index n_i of the external medium, e.g. air, can be approximated as $n_i = 1$, while the refraction index of the dielectric material is denoted by n_t . The angle θ_i is the incidence angle as defined in Section 2.2.2.1 and θ_t the angle between the surface normal and the direction of the light inside the material. The angles θ_i and θ_t are interrelated by Snellius' refraction law (Hecht, 2001) according to

$$n_i \sin \theta_i = n_t \sin \theta_t. \quad (2.54)$$

The Fresnel reflection coefficients are defined as

$$F_{\perp} = \frac{n_i \cos \theta_i - n_t \cos \theta_t}{n_i \cos \theta_i + n_t \cos \theta_t} \quad (2.55)$$

$$F_{\parallel} = \frac{n_t \cos \theta_i - n_i \cos \theta_t}{n_t \cos \theta_i + n_i \cos \theta_t}, \quad (2.56)$$

where Eq. (2.55) yields the reflection ratio for light polarised perpendicular to the plane of incidence and Eq. (2.56) denotes the corresponding ratio for the light polarised parallel to the plane of incidence. Since the Fresnel coefficients denote amplitudes, the corresponding intensity ratios are given by F_{\perp}^2 and F_{\parallel}^2 . For specular reflection there exists a unique relation between the incidence angle θ_i and the polarisation degree D_p^{spec} , which yields for $n_i = 1$:

$$D_p^{\text{spec}}(\theta_i) = \frac{F_{\perp}^2 - F_{\parallel}^2}{F_{\perp}^2 + F_{\parallel}^2} = \frac{2 \sin^2 \theta_i \cos \theta_i \sqrt{n_t^2 - \sin^2 \theta_i}}{n_t^2 - \sin^2 \theta_i - n_t^2 \sin^2 \theta_i + 2 \sin^4 \theta_i} \quad (2.57)$$

(Atkinson and Hancock, 2005b). Eq. (2.57), however, is not valid for diffuse polarisation, which is due to internal scattering of incident light penetrating the surface. The light becomes depolarised as a consequence of the random nature of internal scattering. A fraction of the light is refracted back into the air and therefore partially polarised. Analogous to the case of specular reflection, Snellius' law and the Fresnel equations yield an equation for the degree of polarisation D_p^{diff} of light diffusely scattered by the surface in terms of the incidence angle (again, $n_i = 1$ is assumed):

$$D_p^{\text{diff}} = \frac{\left(n_t - \frac{1}{n_t}\right)^2 \sin^2 \theta_i}{2 + 2n_t^2 - \left(n + \frac{1}{n_t}\right)^2 \sin^2 \theta_i + 4 \cos \theta_i \sqrt{n_t^2 - \sin^2 \theta_i}}. \quad (2.58)$$

Rahmann (1999) exploits the fact that according to the Fresnel equations, light which is specularly reflected by a surface splits up into a parallel and an orthogonal component relative to the plane of reflection, i.e. the plane spanned by the illumination vector \mathbf{s} and the viewing direction \mathbf{v} . As long as the incident light is unpolarised, the unpolarised component of the reflected light corresponds to the amount of light reflected parallel to the plane of reflection. Perpendicular to the plane of reflection, the sum of the unpolarised and the polarised component is observed. Hence, the polarisation angle Φ , associated with the orientation of the polarising filter for which maximum intensity is observed, represents the direction normal to the plane of reflection (Rahmann, 1999). However, it must be possible to distinguish specular from diffuse reflection. While this distinction is usually possible for dielectric surfaces, it will be shown in Section 2.4.2 that the reflectance behaviour of metallic surfaces is more complex and that they display several specular reflection components which strongly overlap with the diffuse reflection component.

It is favourable to employ polarisation imaging for the reconstruction of specularly reflecting surfaces based on polarisation imaging (Rahmann and Canterakis, 2001), utilising for shape recovery the projection of the surface normals directly provided by the polarisation angle. Based on the measured polarisation angle information, iso-depth curves corresponding to surface profiles parallel to the image plane can be obtained, the absolute depth of which, however, is unknown. These level curves provide correspondences between the different views, which are in turn used for triangulation, yielding absolute depth values. In principle, three polarisation images are sufficient for reconstruction of the surface. By formulating the surface reconstruction problem as a minimisation of an error functional (cf. Section 2.2.2.2) describing the mean squared difference between the observed and the modelled polarisation angle, Rahmann and Canterakis (2001) show that the surface of an asymmetrical object can be reconstructed already based on two views, while three views are necessary for simple symmetrical, e.g. spherical, surfaces.

Atkinson and Hancock (2005a) derive the polarisation degree and the polarisation angle from a set of 36 images acquired under different orientations of the linear polarisation filter mounted in front of the lens. They systematically examine the accuracy of the value of θ_i obtained based on Eqs. (2.53) and (2.58). Furthermore, they utilise the result to estimate the BRDF at zero phase angle (parallel illumination vector \mathbf{s} and viewing direction \mathbf{v}) of the surface material. This method to estimate the surface normals based on Fresnel theory is extended by Atkinson and Hancock (2005b) towards a multi-view framework in which the polarisation information is exploited to establish potential correspondences between different views in the absence of surface texture. Potential correspondences are located according to similar values of $|\theta_D| = \arctan(\sin \Phi \tan \theta_i)$ and then refined by locally reconstructing the depth of surface parts near the selected points and comparing the results across the different views. The resulting surface gradient field yields the depth map

of the surface according to the algorithm by Frankot and Chellappa (1988). This polarisation-based stereo approach is applicable to surfaces which are inaccessible to classical stereo vision approaches.

A polarisation-based method to estimate three-dimensional shape, texture, surface roughness, and illumination distribution for an object from a single viewpoint is described by Miyazaki et al. (2003). The proposed method relies on an estimation of three-dimensional shape by computing the surface normals from polarisation data, an extraction of the surface texture from the diffuse reflection component, a determination of the directions to the light sources from the maxima of the specular reflection component, and a computation of the surface roughness based on the previously estimated illumination distribution. A specularity-free image is constructed by elementary pixel-wise transformations of the original colour image in hue-saturation-intensity space in order to separate the specular from the diffuse reflection component, where a surface reflectance behaviour according to the Torrance-Sparrow model (cf. Section 2.2.1) is assumed. The shape from polarisation approach is extended to transparent objects by Miyazaki et al. (2004). They determine the surface orientation based on the polarisation angle Φ and the specular polarisation degree D_p^{spec} according to Eq. (2.57). The two-fold ambiguity of the value of θ_i determined from D_p^{spec} is resolved by rotating the object and establishing correspondences between the views, which yields the surface normal unambiguously. This approach is extended by Miyazaki and Ikeuchi (2005), who introduce a method for polarisation raytracing to compute both the path of the light and the polarisation state, considering the reflection, refraction, interreflection, and transmission of the light occurring at the surface of and inside a transparent object. An iterative optimisation approach estimates the three-dimensional surface shape by minimising the difference between the measured polarisation data and the modelled polarisation data obtained based on the raytracing approach.

2.4.2 Determination of Polarimetric Properties of Rough Metallic Surfaces for Three-dimensional Reconstruction Purposes

The specular polarisation degree D_p^{spec} defined according to Eq. (2.57) cannot be directly applied to metallic surfaces since the refraction index of metals is complex. It is written in the form $\hat{n} = n(1 + ik)$ with the imaginary part k as the attenuation index. Morel et al. (2005) apply the approximation $|\hat{n}|^2 = n_r^2(1 + k^2) \gg 1$, which is generally true for visible wavelengths, and accordingly obtain the relation

$$D_p^{\text{met}} = \frac{2n \tan \theta_i \sin \theta_i}{\tan^2 \theta_i \sin^2 \theta_i + |\hat{n}|^2} \quad (2.59)$$

for the polarisation degree in terms of the incidence angle. Eq. (2.59) is valid for smooth, specularly reflecting metallic surfaces. Provided that the complex refraction index of the material is known, Morel et al. (2005) compute the surface gradients

and determine the height map of the surface by integration of the gradient field. The surface is illuminated diffusely by a hemispherical dome. The described method is used in the context of quality inspection of very smooth polished mirror surfaces.

According to polarised light scattering measurements performed by Germer et al. (2000), it is questionable, however, if a simple polarisation model like Eq. (2.59) can be applied to metallic surfaces. The steel surface samples examined by Germer et al. (2000) were polished with various polishing emulsions, and some were etched in a sulphuric acid solution. In the course of their ellipsometric measurements, they determine the BRDF of the surface, the degree of circular polarisation, the total degree of polarisation, and the polarisation angle of the light reflected from the surface for a constant incidence and emission angle of $\theta_i = \theta_e = 60^\circ$ over a range of incident polarisation states and azimuthal scattering angles. The surface roughness is determined by means of atomic force microscopy. The polarimetric measurements are compared to theoretical predictions for scattering from surface topography (micro-roughness) and subsurface permittivity variations. For most values of the azimuthal scattering angle, the measurements are neither consistent with the microroughness model nor with the subsurface scattering model but have to be explained by a combination of the two different scattering mechanisms. The relative amounts of the two scattering sources is also dependent on sample preparation.

The experimental results by Germer et al. (2000) give an impression of the difficulties encountered when attempting to apply polarisation models to metallic surfaces. This is especially true for rough metallic surfaces. As a consequence, for the raw forged iron surfaces regarded in the industrial quality inspection scenarios described in Section 5.3, we found that it is favourable to determine empirically the BRDF and the polarisation properties instead of relying on physical models.

The measurement procedure employed for determining the polarisation properties of the surface is similar to the method described by Atkinson and Hancock (2005a). A flat sample part is attached to a goniometer, which allows a rotation of the sample around two orthogonal axes. The corresponding goniometer angles γ_1 and γ_2 can be adjusted at an accuracy of a few arcseconds. As illustrated in Fig. 2.5, adjusting γ_1 is equivalent to rotating the surface normal \mathbf{n} around an axis perpendicular to the plane spanned by the vectors \mathbf{s} and \mathbf{v} , while adjusting γ_2 causes a rotation of \mathbf{n} around an axis lying in that plane. The phase angle α between \mathbf{s} and \mathbf{v} is independent of γ_1 and γ_2 , since the centre of rotation lies on the sample surface, and is assumed to be constant over the image. It is straightforward to determine the surface normal \mathbf{n} , the incidence angle θ_i , and the emission angle θ_e from the goniometer angles γ_1 and γ_2 and the vectors \mathbf{s} and \mathbf{v} . For each configuration of goniometer angles, five images are acquired through a linear polarisation filter at orientation angles ω of 0° , 45° , 90° , 135° , and 180° . Due to the encountered wide range of reflected intensities, a high dynamic range image is synthesised from four low dynamic range images acquired with different exposure times. For each filter orientation ω , an average pixel intensity over an image area containing a flat part of the sample surface is computed. A sinusoidal function of the form (2.52) is then fitted to the measured pixel intensities. The filter orientation Φ for which maximum intensity is observed corresponds to the polarisation angle, and the polarisation degree is readily obtained

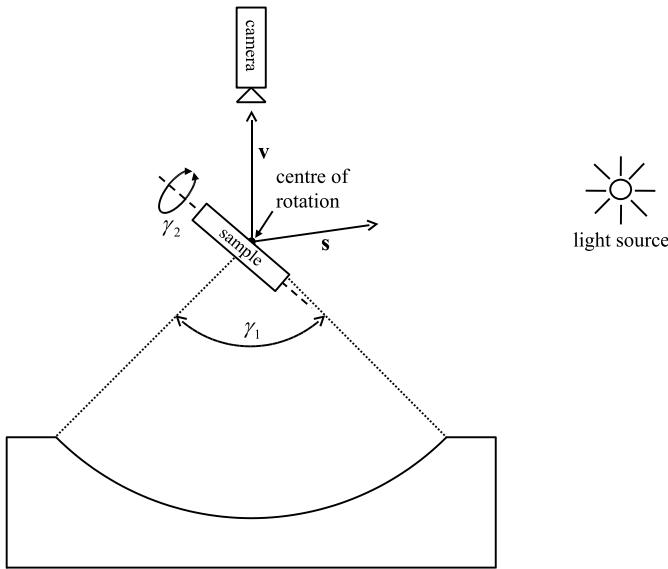


Fig. 2.5 Definition of the goniometer angles γ_1 and γ_2 .

from the sinusoidal fit according to $D_p = I_v/I_c$ (cf. Section 2.4.1). In principle, three measurements would be sufficient to determine the three parameters I_c , I_v , and Φ , but the fit becomes less noise-sensitive when more measurements are used.

As apparent from the previous discussion, no accurate physically motivated model for the polarisation properties of rough metallic surfaces is available. Hence, we fit a phenomenological model, here a polynomial in terms of the goniometer angles γ_1 and γ_2 , to the measured values of the polarisation angle and degree. The polarisation angle is represented by an incomplete third-degree polynomial of the form

$$R_\Phi(\gamma_1, \gamma_2) = a_\Phi + b_\Phi \gamma_2 + c_\Phi \gamma_1 \gamma_2 + d_\Phi \gamma_1^2 \gamma_2 + e_\Phi \gamma_2^3, \quad (2.60)$$

which is antisymmetric in γ_2 , and $R_\Phi(\gamma_1, 0) = a_\Phi = \text{const}$, corresponding to coplanar vectors \mathbf{n} , \mathbf{s} , and \mathbf{v} . In an analogous manner, the polarisation degree is represented by an incomplete polynomial of the form

$$R_D(\gamma_1, \gamma_2) = a_D + b_D \gamma_1 + c_D \gamma_1^2 + d_D \gamma_2^2 + e_D \gamma_1^2 \gamma_2^2 + f_D \gamma_2^4 + g_D \gamma_1 \gamma_2^4 + h_D \gamma_1^2 \gamma_2^4, \quad (2.61)$$

which is symmetric in γ_2 . The symmetry properties are required for geometrical reasons as long as an isotropic interaction between incident light and surface material can be assumed. The polarisation properties of a raw forged iron surface measured at a phase angle of $\alpha = 79^\circ$ are illustrated in Fig. 2.6, along with the polynomial fits according to Eqs. (2.60) and (2.61). Interestingly, while in the framework based on Fresnel theory outlined in Section 2.4.1 the polarisation angle corresponds to the projection of the surface normal into the image plane (up to a 90° phase shift for specular polarisation), a close inspection of our polarisation angle measurements re-

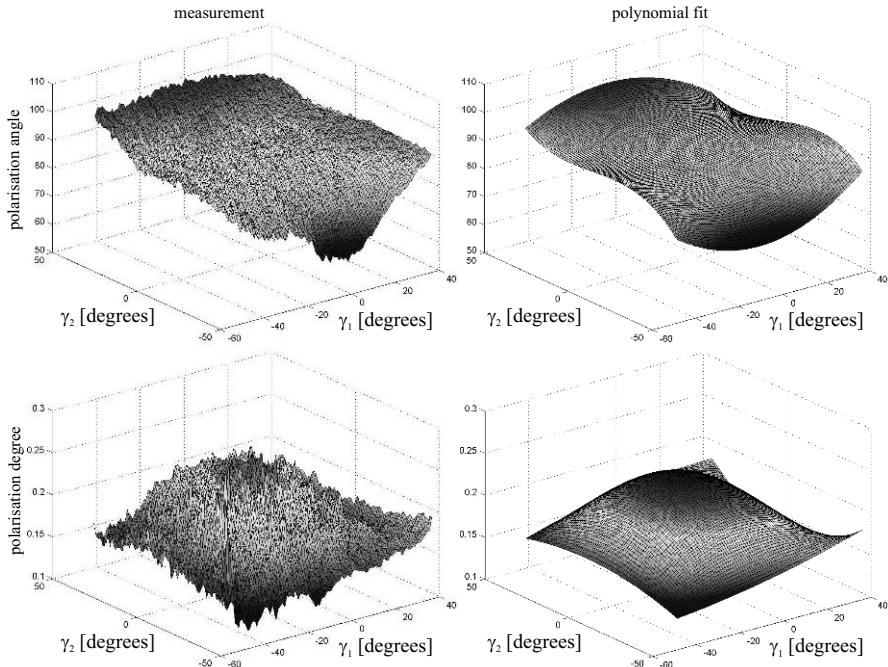


Fig. 2.6 Measured and modelled polarisation properties of a raw forged iron surface at a phase angle of $\alpha = 79^\circ$. Top: polarisation angle. Bottom: polarisation degree.

veals that this simple relation does not hold for the rough metallic surfaces regarded in this section.

At this point it is straightforward to determine the surface gradients p and q associated with the goniometer angles γ_1 and γ_2 based on the mechanical setup of the goniometer. Based on the fitted phenomenological laws for the polarisation angle Φ and degree D_p , the functions $R_\Phi(p, q)$ and $R_D(p, q)$ yielding the polarisation properties in terms of the surface gradients p and q are obtained. In analogy to the intensity reflectance map $R_I(p, q)$, we termed $R_\Phi(p, q)$ polarisation angle reflectance map and $R_D(p, q)$ polarisation degree reflectance map. Results concerning three-dimensional reconstruction of rough metallic surfaces based on the evaluation of polarisation features are described in the context of the shape from photopolarimetric reflectance framework in Section 4.3 and the application scenario of quality inspection outlined in Section 5.3.

Chapter 3

Real-aperture Approaches to Three-dimensional Scene Reconstruction

All methods for three-dimensional scene reconstruction described in the previous chapters rely on a pinhole camera model, i.e. the images are well focused regardless of object distance. Physically, this corresponds to the unrealistic assumption of a pinhole of zero radius. Real optical systems, of course, have apertures with finite diameters. The shape of the aperture is usually given by the shape of the iris, which is generally approximately circular but may also be hexagonal or triangular.

As a result of the finite aperture of the optical system, the observed image I_{uv} can be modelled by the convolution of the “ideal” image $I_{uv}^{(0)}$ with the spatially varying point spread function (PSF) G_{uv} according to

$$I_{uv} = G_{uv} * I_{uv}^{(0)}. \quad (3.1)$$

It should be noted that the convolution operation according to Eq. (3.1) in principle assumes an input image and a PSF of infinite size. In practice, image and PSF are of finite size, and we perform this convolution operation in frequency space using the fast Fourier transform (FFT) algorithm (Press et al., 1992). This method assumes a periodically continued version of the finite-sized input image and the PSF. The problem of discontinuities at the image borders that may be introduced by the periodic continuation is regarded in Section 3.2.1.

For monochromatic light, an exact description of the PSF due to diffraction of light at a circular aperture is given by the radially symmetric Airy pattern $A(r) \propto [J_1(r)/r]^2$, where $J_1(r)$ is a Bessel function of the first kind of first order (Hecht, 2001). Consequently, the image of a point light source is radially symmetric and displays an intensity maximum at its centre and concentric rings surrounding the maximum with brightnesses that decrease for increasing ring radius (Fig. 3.1). The central maximum is also called Airy disk (Hecht, 2001). For non-monochromatic light, the rings produced by the diffraction vary with wavelength in amplitude, width, and position. For the range of visual wavelengths, the phases of the rings change by as much as 90° , leading to a superposition of minima and maxima corresponding to different wavelengths. If additional imaging system artifacts like chromatic aberration and digital sampling are considered, a radially symmetric Gaussian

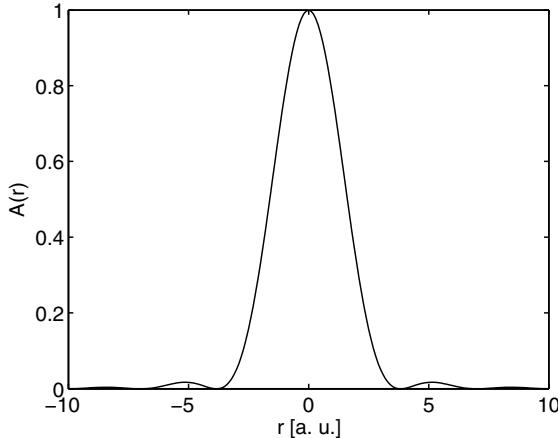


Fig. 3.1 One-dimensional section of the Airy pattern $A(r)$.

function is a reasonable approximation to the true PSF (Pentland, 1987; Chaudhuri and Rajagopalan, 1999):

$$G(r, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{r^2}{2\sigma^2}}. \quad (3.2)$$

Determining depth information by actively controlling the camera parameters and thus exploiting the effect of the PSF on the image is an alternative to the previously described geometric and photometric techniques. Important advantages of such focus-based methods are that they do not have a correspondence problem since the images of the scene can be acquired in a pixel-synchronous manner, and no knowledge about viewing direction, illumination direction or the reflectance properties of the object surfaces is required. Basically, there are two distinct approaches to utilising focus information for depth recovery. The depth from focus method outlined in Section 3.1 aims for determining depth by acquiring a large number of images at different known focus settings. The configuration for which the sharpest image is observed is used to compute the distance to the scene point based on the known intrinsic camera parameters. The depth from defocus approach described in Section 3.2 consists of acquiring a small number of images at different focus setting, where the differences of the PSF across the set of images are exploited to recover the depth of scene points.

3.1 Depth from Focus

The PSF of a defocused optical system acts like a low-pass filter on the original image (Nayar, 1989). This low-pass filter influences most strongly the high spatial frequencies of the original image. Hence, if it is desired to measure the effect of the optical system the original image must contain high spatial frequencies, i.e. the surfaces of the objects present have to be rough or textured. Otherwise it is still possible to illuminate the scene with a light pattern containing high spatial frequencies. The image is focused if the lens law

$$\frac{1}{b} + \frac{1}{z_0} = \frac{1}{f} \quad (3.3)$$

is fulfilled, where b denotes the principal distance, i.e. the distance between the optical centre and the image plane, z_0 the depth of the perfectly focused scene point, and f the focal length of the optical system. The values of b and f are determined based on a calibration procedure, such that the value of z_0 is directly given by Eq. (3.3). If the depth deviates from the value z_0 in Eq. (3.3) for which the scene point appears best focused while b and f are kept constant, the image of the scene point becomes blurred. Nayar (1989) proposes to attach the object under study to a reference plane, which is moved towards the camera in small depth intervals Δz known at high precision. Hence, for image n the depth of the reference plane amounts to $z_n = z_0 + n\Delta z$, where z_0 is the initial depth. If a point on the surface appears best focused in image n , its distance to the camera must correspond to z_0 , and the index n directly determines the distance between the scene point and the reference plane.

An automatic detection of best focus involves the definition of a focus measure. For this purpose, Nayar (1989) utilises the value of a sum-modified Laplacian (SML) according to

$$F_{uv} = \sum_{x=u-N}^{u+N} \sum_{y=v-N}^{v+N} \left[\left| \frac{\partial^2 I}{\partial x^2} \right| + \left| \frac{\partial^2 I}{\partial y^2} \right| \right] \quad (3.4)$$

with I as the pixel grey value. The sum is computed over an image window of size $(2N+1) \times (2N+1)$ pixels centred around the pixel at position (u, v) . Nayar (1989) compares the SML focus measure to the grey value variance, the sum-Laplacian, and the Tenengrad measure (Krotkov, 1987), and it is shown that the SML measure for most regarded material samples yields the steepest and most accurately located maximum. The focus measurements are interpolated by fitting a Gaussian function to the maximum SML value and the two neighbouring measurements in order to obtain depth measurements which are more accurate than the interval Δz .

If the profile of the focus measure is assumed to be unimodal in the interval $[x, y]$, and if $x < x_1 < x_2 < y$ and $F(x_1) < F(x_2)$ with $F(x)$ as the focus measure, then the maximum of $F(x)$ cannot be located in the interval $[x, x_1]$, since otherwise the unimodality assumption would not be fulfilled. Hence, if it is possible to select values for x_1 and x_2 accordingly e.g. by bisection, the maximum can be found

optimally. The search method based on iteratively narrowing the search interval is termed Fibonacci search and corresponds to the optimal search under the assumption of unimodality (Xiong and Shafer, 1993). In the presence of noise, however, the function $F(x)$ is not strictly unimodal but may display a large number of local maxima. Hence, Xiong and Shafer (1993) propose to terminate the Fibonacci search for an interval width below a given threshold to fit a Gaussian interpolation function to this search interval.

The depth from focus method has the advantage that it yields an accurate depth map. A disadvantage is that a large number of images have to be acquired, resulting in a considerable computational effort. Hence, only static scenes can be reconstructed with this technique. Furthermore, the intrinsic camera parameters, especially the focal length and the principal distance, as well as the displacement of the object or the camera during image acquisition have to be known at high accuracy. Applications of the depth from focus method include automated quality control of micro-technology components (Schaper, 2002) and the detection of obstacles in the context of autonomous mobile robot guidance (Nourbakhsh et al., 1997).

3.2 Depth from Defocus

The amount of generally spatially varying blur in a defocused image contains information about a scene point. We will see in this section that in principle already two images of different focus are sufficient to recover blur difference between the images and in turn the depth of the scene. A-priori information about the image intensity distribution, e.g. the presence of sharp discontinuities (edges), allows the computation of the distribution of blur based on a single defocused image. If no a-priori information is available, the ideally sharp image $I_{uv}^{(0)}$ in Eq. (3.1) can be approximated by an image acquired with a very small (pinhole) aperture. This approach to depth estimation based on exploiting the PSF difference between a pair of images is termed depth from defocus. Pentland (1987) states that defocus information is also used by biological visual systems to infer the depth of the scene based on essentially the same principle, since e.g. the focal length of the human eye is varying in a sinusoidal manner at a frequency of about 2 Hz. Hence, two differently focused views of the same scene are obtained within a time interval of 250 ms.

3.2.1 Basic Principles

Pentland (1987) introduces the concept of the blur circle (also known as circle of confusion) into the depth from defocus framework. The lens law (3.3) yields for the distance z_0 between the lens and a scene point in perfect focus the relation $z_0 = fb_0/(b_0 - f)$, where b_0 is the distance between the optical centre and the image plane in perfect focus and f the focal length of the lens (cf. Fig. 3.2). If the distance

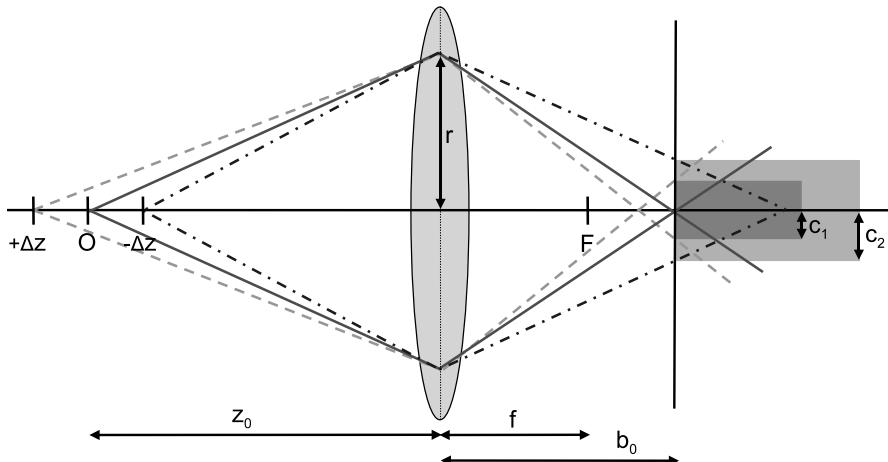


Fig. 3.2 Relation between a depth offset Δz and the resulting circles of confusion with radii c_1 and c_2 .

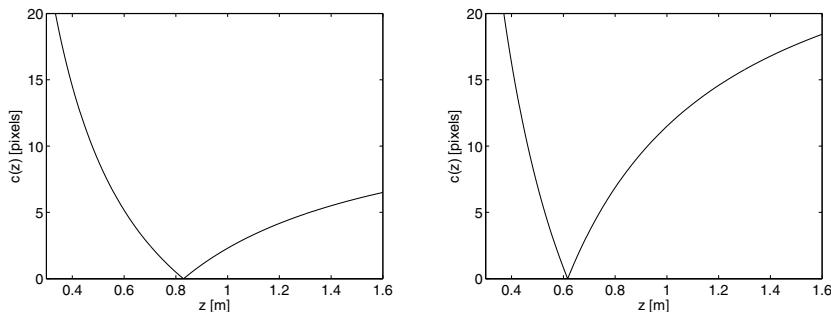


Fig. 3.3 Dependence of the radius c of the circle of confusion on the depth z for a lens of $f = 12$ mm at $\kappa = 1.4$ (left) and for a lens of $f = 20$ mm at $\kappa = 2.4$ (right). The value of c is given in pixels, where the size of the square pixels corresponds to $4.65 \mu\text{m}$. For each configuration, the principal distance b is assumed to be fixed. The image is best focused at a depth of about 0.8 m and 0.6 m, respectively.

between the optical centre and the image plane is set to a value $b = b_0 - \Delta b$, points at distance $z = z_0 + \Delta z = fb/(b - f)$ are in perfect focus. If the image plane remains at a distance b_0 to the optical centre while the distance of the scene point is increased from z_0 to $z_0 + \Delta z$, the image of this point appears focused in front of the image plane. Fig. 3.2 reveals that $r/b = c/(b - b_0)$ with r as the lens radius and c as the radius of the circle into which the scene point is spread in the image plane. Hence, we have $z = frb/[rb - f(r + c)]$, corresponding to

$$z = \frac{fb}{b - f - 2c\kappa} \quad (3.5)$$

with $\kappa = f/(2r)$ as the f-stop number of the lens. For illustration, the dependence of the radius c on the depth z for fixed focal length f and principal distance b according to Eq. (3.5) is shown for a lens of $f = 12$ mm at $\kappa = 1.4$ and for a lens of $f = 20$ mm at $\kappa = 2.4$ in Fig. 3.3. To allow a comparison with the experimental results described later on, the value of c is computed for a Baumer CCD camera with square pixels of size 4.65 μm .

The relation between the depth z of the scene point and the corresponding radius c of the circle of confusion expressed by Eq. (3.5) is the result of elementary geometric optics considerations. At this point, an important conclusion is drawn by Pentland (1987), namely that the width σ of the PSF $G(\sigma)$, which is assumed to be of Gaussian shape, is taken to be proportional to the radius c of the circle of confusion, such that

$$\sigma = \gamma c \quad (3.6)$$

with the camera-specific proportionality constant γ . As a consequence, the depth from defocus problem reduces to the determination of the PSF difference between a focused and a defocused image.

The classical depth from defocus approach introduced by Pentland (1987) and by Subbarao (1988) uses two possibly defocused images I_1 and I_2 of the same scene taken at two different focal settings. Let $\mathcal{I}_1(\omega_u, \omega_v)$ and $\mathcal{I}_2(\omega_u, \omega_v)$ be the amplitude spectra of I_1 and I_2 . In frequency space, the convolution of the perfectly focused image $I_{uv}^{(0)}$ with the PSF G according to Eq. (3.1) becomes an element-wise multiplication. Hence, by dividing $\mathcal{I}_1(\omega_u, \omega_v)$ by $\mathcal{I}_2(\omega_u, \omega_v)$, the unknown ideally focused image $I_{uv}^{(0)}$ can be eliminated, leading to

$$\frac{\mathcal{I}_1(\omega_u, \omega_v)}{\mathcal{I}_2(\omega_u, \omega_v)} = \exp \left[-\frac{1}{2} (\omega_u^2 + \omega_v^2) (\sigma_1^2 - \sigma_2^2) \right]. \quad (3.7)$$

The term $(\sigma_1^2 - \sigma_2^2)$ is then given by

$$\sigma_1^2 - \sigma_2^2 = -2 \left\langle \frac{1}{\omega_u^2 + \omega_v^2} \ln \frac{\mathcal{I}_1(\omega_u, \omega_v)}{\mathcal{I}_2(\omega_u, \omega_v)} \right\rangle_{\omega_u, \omega_v}, \quad (3.8)$$

where $\langle \dots \rangle_{\omega_u, \omega_v}$ denotes an average over ω_u and ω_v . In principle, only one pair of amplitudes $\mathcal{I}_1(\omega_u, \omega_v)$ and $\mathcal{I}_2(\omega_u, \omega_v)$ is required to compute $(\sigma_1^2 - \sigma_2^2)$, but averaging over a larger domain of the amplitude spectrum as described in Section 3.2.2 is favourable.

If one of the images is perfectly focused, i.e. $\sigma_1 = 0$, Eqs. (3.5), (3.6), and (3.8) allow a direct computation of the depth z . If the two images are blurred, i.e. $\sigma_1 > 0$ and $\sigma_2 > 0$, due to different focus settings and thus different principal distances b_1 and b_2 , it follows from Eq. (3.5) that $z = b_1 / (b_1 - f - 2c_1\kappa) = b_2 / (b_2 - f - 2c_2\kappa)$, which in combination with Eq. (3.6) leads to a linear relation between σ_1 and σ_2 according to

$$\sigma_1 = \alpha\sigma_2 + \beta \quad \text{with} \quad \alpha = \frac{b_1}{b_2} \quad \text{and} \quad \beta = \frac{fb_1\gamma}{2\kappa} \left(\frac{1}{b_2} - \frac{1}{b_1} \right). \quad (3.9)$$

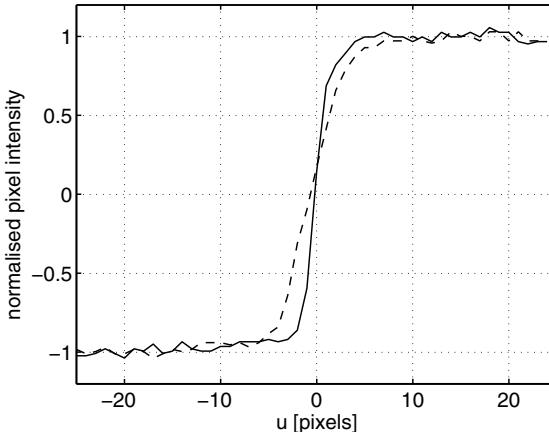


Fig. 3.4 Two intensity profiles, extracted orthogonal to a slightly defocused (solid curve) and a strongly defocused (dashed curve) object boundary, respectively.

Eq. (3.9) directly yields the relation

$$\sigma_1^2 - \sigma_2^2 = (\alpha^2 - 1)\sigma_2^2 + 2\alpha\beta\sigma_2 + \beta^2 \quad (3.10)$$

(Chaudhuri and Rajagopalan, 1999), which is readily solved for σ_2 . Determining $(\sigma_1^2 - \sigma_2^2)$ according to Eq. (3.8) then yields the value of σ_2 . By applying this technique to local image windows, a PSF radius and in turn a depth value can be computed for each pixel of an image. Subbarao and Surya (1994) extend this approach to PSFs of arbitrary, non-Gaussian shape by introducing a generalisation of the parameter σ of the Gaussian PSF $G(r, \sigma)$ as the square root of the second central moment of the arbitrarily shaped PSF.

An alternative approach, relying on a single image of the scene, is to determine the amount of defocus in the image by assuming that intensity transitions along object boundaries are ideal edges in the real world, which are blurred by the optical system of the camera (Pentland, 1987). As an example, two intensity profiles extracted orthogonal to an object boundary, displaying different amounts of defocus, are shown in Fig. 3.4. The intensity profile $I(u)$ generated by blurring an ideal edge with a Gaussian PSF can be represented by a function of the form

$$I(u) = a \operatorname{erf}\left(\frac{u - u_0}{\sqrt{2}\sigma}\right) + b \quad (3.11)$$

with u as the pixel coordinate orthogonal to the boundary, a as the amplitude of the edge, b as an offset parameter, u_0 as the position of the steepest brightness gradient, and σ as the PSF radius in pixel units (cf. Section 1.4.8). The error function $\operatorname{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-s^2} ds$ is the step response of the Gaussian PSF. The edge is well focused for $\sigma \rightarrow 0$ while the amount of defocus increases for increasing σ . This technique is

further regarded in Section 4.5.2 in the context of three-dimensional pose estimation based on the integration of depth from defocus information into the CCD algorithm described in Section 1.6.2.3.

The recovery of depth from defocused images is a space-variant blur identification problem. In this context, Subbarao and Wei (1992) introduce a computationally efficient method to determine the width of the PSF based on the evaluation of one-dimensional brightness profiles rather than two-dimensional image regions. The one-dimensional profiles are obtained by summing the pixel rows. The FFT algorithm (Press et al., 1992) used to compute the amplitude spectrum assumes that the image continues periodically at its borders, which may introduce spurious contributions from high spatial frequencies due to discontinuities between the left and the right as well as the upper and the lower image border, respectively. To circumvent this problem, Subbarao and Wei (1992) perform a pixel-wise multiplication of the grey values with a two-dimensional Gaussian function centred at the image centre, having a width of about $2/3$ of the image size.

As a note of interest, an alternative method to avoid such discontinuities is to pad the image, which is assumed to be of size $N \times N$ pixels, into the lower left quadrant of a new image matrix of size $2N \times 2N$ pixels. Then the image is mirrored at its right and its upper border, respectively, and the new pixel grey values are padded into the lower right and the upper left quadrant of the $2N \times 2N$ matrix, respectively. Finally a rotation by 180° of the image is performed around its upper right corner, and the resulting pixel grey values are padded into the upper right quadrant. No discontinuities occur at the borders of the new image of size $2N \times 2N$ pixels if it is continued periodically in horizontal and vertical direction (Donner, 1995).

Another computationally efficient scheme to determine the PSF difference between two images based on the inverse S-transform is introduced by Subbarao and Surya (1994), involving only local spatial filtering operations rather than a two-dimensional Fourier transform.

Windowing for local PSF analysis assumes a constant depth of all scene points captured in the window, such that tilted surfaces or depth discontinuities will result in increased measurement errors. Additionally, an interaction occurs between neighbouring windows due to the “spreading” of scene points from one window into neighbouring windows. This effect is particularly relevant for small image windows. To take into account such cases, Chaudhuri and Rajagopalan (1999) introduce the block shift-variant blur algorithm that takes into account the interaction of blur among neighbouring subimages. Instead of a window-wise Fourier transform, the blurring is regarded in a space-frequency representation framework. Furthermore, a variational approach incorporating a smoothness constraint on the PSF parameter is introduced. Based on a maximum likelihood approach, a criterion for the optimal relative blurring of the two regarded images is inferred. These frameworks are extended towards a depth estimation based on multiple blurred images of the scene. Based on a simulated annealing approach, maximum a-posteriori estimates of the depth and the focused image of the scene are derived. This mathematical framework established by Chaudhuri and Rajagopalan (1999) is especially relevant in the

case of a PSF which is highly variable across the image and in the presence of strong depth discontinuities in the scene.

Depth from defocus is more sensitive to inaccurate camera and blur models than depth from focus. Some commonly used lenses show non-Gaussian PSFs, sometimes depending on whether the image is focused in front or behind the image plane. Similar to depth from focus, the image content should display a significant contribution from high spatial frequencies, as it is the case e.g. for textured or rough surfaces. A favourable property of the depth from defocus approach is that the computed depth map is dense, i.e. for each pixel a depth value is determined, but these depth values tend to display a considerable scatter. In Section 3.2.2 we describe to what extent relatively small depth differences across surfaces can be recovered with the depth from defocus method under realistic circumstances. Section 3.2.3 introduces a framework for the determination of absolute depth values across broad ranges, showing the necessity to put aside the simplifying assumption of a proportionality between the PSF radius σ and the radius c of the circle of confusion and introducing a more appropriate empirical calibration approach.

3.2.2 Determination of Small Depth Differences

For clarity, we utilise in this section the radius Σ of the Gaussian PSF in frequency space as a measure for the image blur, where $\Sigma \propto 1/\sigma$. The observed image blur decreases with increasing value of Σ , such that a perfectly sharp image is characterised by $\Sigma \rightarrow \infty$, corresponding to $\sigma = 0$. Following the approach by d'Angelo and Wöhler (2005c), the calibration procedure for estimating depth from defocus then involves the determination of the lens-specific characteristic curve $\Sigma(z - z_0)$. For this purpose we acquire two pixel-synchronous images of a rough, uniformly textured plane surface consisting of forged iron, inclined by 45° with respect to the optical axis. The image part in which the intensity displays maximum standard deviation (i.e. most pronounced high spatial frequencies) is sharp and thus situated at distance z_0 . A given difference in pixel coordinates with respect to that image location directly yields the corresponding depth offset $(z - z_0)$. The first image I_1 is taken with small aperture, i.e. $f/8$, resulting in virtually absent image blur, while the second image I_2 is taken with a large aperture, i.e. $f/2$, resulting in a perceivable image blur that depends on the depth offset $(z - z_0)$.

The images are partitioned into quadratic windows of size 32×32 pixels, for each of which the average depth offset $(z - z_0)$ is known. This window size is at least one order of magnitude larger than the spatial PSF radii encountered during our measurements. The PSF radius Σ in frequency space is computed based on the ratio $\mathcal{I}_2(\omega_u, \omega_v)/\mathcal{I}_1(\omega_u, \omega_v)$ of the amplitude spectra of the corresponding windows of the second and the first image, respectively:

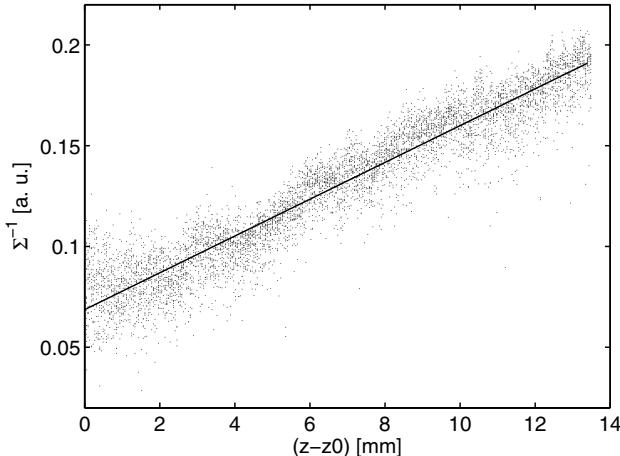


Fig. 3.5 Calibration of the depth from defocus algorithm: measurements $(\Sigma^{-1}, (z - z_0))$ and fitted characteristic curve.

$$\Sigma^2 = - \left\langle \frac{\omega_u^2 + \omega_v^2}{2 \ln \frac{\mathcal{I}_2(\omega_u, \omega_v)}{\mathcal{I}_1(\omega_u, \omega_v)}} \right\rangle_{\omega_u, \omega_v} \quad (3.12)$$

(Subbarao, 1988), where $\langle \dots \rangle_{\omega_u, \omega_v}$ denotes the average over the coordinates ω_u and ω_v in frequency space. Only the range of intermediate spatial frequencies is regarded in order to reduce the influence of noise on the resulting value of Σ . If the amplitude spectrum of the examined image window displays a very low value at (ω_u, ω_v) , the corresponding amplitude ratio tends to be inaccurate, which may result in a substantial error of Σ . Hence, we first compute Σ according to Eq. (3.12), identify all spatial frequencies (ω_u, ω_v) for which the term in brackets in Eq. (3.12) deviates by more than one standard deviation from Σ , and recompute Σ after neglecting these outliers.

For a given value of Σ , the corresponding value of $(z - z_0)$ is ambiguous since two depth values $z_1 < z_0$ and $z_2 > z_0$ may correspond to the same value of Σ . In our experiments we avoided this two-fold ambiguity by placing the complete surface to be reconstructed behind the plane at distance z_0 , implying $z > z_0$. One would expect $\Sigma \rightarrow \infty$ for $z \rightarrow z_0$, since ideally the small-aperture image and the large-aperture image are identical for $z = z_0$. We found empirically, however, that due to the imperfections of the optical system, even for $z = z_0$ an image window acquired with larger aperture is slightly more blurred than the corresponding image window acquired with smaller aperture. This remains true as long as the small aperture is sufficiently large for diffraction effects to be small. As a consequence, Σ obtains a finite maximum value at $z = z_0$ and decreases continuously for increasing z .

The geometric optics based approach by Pentland (1987) implies that the radius c of the circle of confusion is proportional to the value of Δb . In turn, the PSF radius in image space (being proportional to Σ^{-1}) is assumed to be proportional to

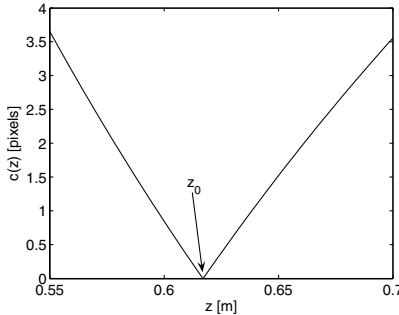


Fig. 3.6 Enlarged section of the right diagram of Fig. 3.3 ($f = 20$ mm and $\kappa = 2.4$), showing that c is approximately proportional to $|z - z_0|$ for small $|z - z_0|$.

c , implying $\Sigma^{-1} \rightarrow 0$ for $z \rightarrow z_0$. For the lenses, CCD sensors, and object distances regarded in our experiments (cf. Section 5), it follows from the models by Pentland (1987) and Subbarao (1988) that c and thus Σ^{-1} are proportional to $|z - z_0|$ for small values of $|z - z_0|$ of some millimetres and for small radii of the circle of confusion of a few pixels (cf. Fig. 3.6). It turns out that the measured $(\Sigma^{-1}, (z - z_0))$ data points can be represented fairly well by a linear function (cf. Fig. 3.5), displaying a non-zero offset due to the aforementioned imperfections of the optical system. The root-mean-square deviation of the depth from defocus measurements from the fitted line corresponds to 1.2 mm at a distance of the reference object of 480 mm, which corresponds to a relative depth error of 0.25 percent. This simple relationship holds due to the regarded small depth range. For larger depth ranges, more appropriate representations of the relationship between PSF radius and depth are introduced in Section 3.2.3.

For the raw forged iron surface of a connection rod we acquired images from a distance of about 480 mm with a lens of $f = 25$ mm focal length at a resolution of 86 μm per pixel at $\kappa = 8$ and $\kappa = 2$, respectively (cf. Fig. 3.7a). The raw depth from defocus measurements are shown in Fig. 3.7b. Smoothing this result by a median filter of 30×30 pixels size yields the surface profile shown in Fig. 3.7c, which provides a reasonable impression of the overall surface shape. However, the profile shows many spurious structures especially on small scales, such that shape details are not reliably recovered. According to the calibration curve shown in Fig. 3.5, the standard error of the depth values obtained by depth from defocus corresponds to 1.2 mm, with an average object distance of 480 mm. Although this corresponds to an absolute depth error of only 0.3 percent, due to the relatively small depth extension of the scene of about 4 mm the extracted set of three-dimensional points contains reasonably reliable information only about the average surface gradient but not about higher-order surface properties. It is shown in Sections 4.3 and 4.5 that in combination with geometric and photometric methods, such depth from defocus measurements may nevertheless contribute useful information to the three-

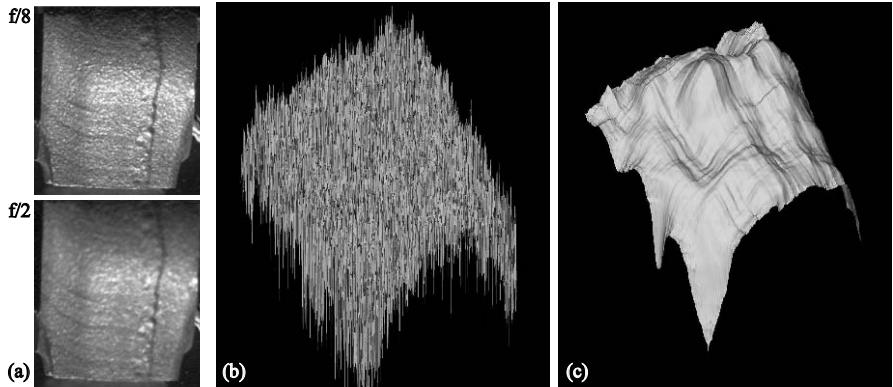


Fig. 3.7 Depth from defocus analysis of a raw forged iron surface. (a) Images acquired with different apertures. (b) Raw measurements. (c) Raw measurements smoothed with a median filter of 30×30 pixels size.

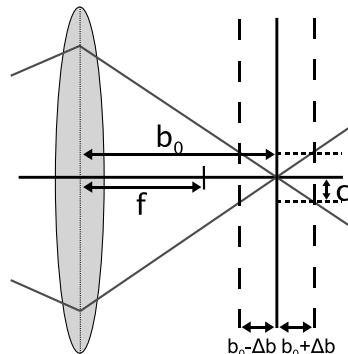


Fig. 3.8 Relation between the offset $\Delta b = |b - b_0|$ of the image plane and the radius c of the circle of confusion.

dimensional reconstruction of a surface or an object, as pointed out by d'Angelo and Wöhler (2005c).

3.2.3 Determination of Absolute Depth Across Broad Ranges

In the geometric optics approximation, a point in the scene is transformed into a circle of confusion of radius

$$c = \Delta b / (2\kappa) \quad (3.13)$$

in the image plane, where $\Delta b = |b - b_0|$ and κ is the f-stop number (cf. Fig. 3.8). Assuming that the PSF radius σ is proportional to the value of c directly leads to Eq. (3.5). Empirically, however, we found that for typical video lenses σ is not

proportional to $|b - b_0|$, as it should be according to the assumption of a proportionality between σ and c . Kuhl et al. (2006) demonstrate that the behaviour of σ is much better represented by a function which is symmetric in $|b - b_0|$, displays a minimum with a continuous first derivative at $b = b_0$, and saturates for larger values of $|b - b_0|$. A good representation is obtained with a Gaussian of the form

$$\sigma(\Delta b) = \frac{1}{\phi_1} \exp\left(-\frac{(\Delta b)^2}{\phi_2}\right) + \phi_3, \quad (3.14)$$

where ϕ_1 , ϕ_2 , and ϕ_3 are empirically determined parameters. We can safely assume that the radius c of the circle of confusion and the PSF radius σ are related to each other in that σ is a monotonically increasing nonlinear function of c . Hence, the symmetric behaviour of $c(\Delta b)$ apparent from Fig. 3.8 implies a symmetric behaviour of $\sigma(\Delta b)$.

Depending on the constructional properties of the lens, analytic forms different from Eq. (3.14) but also symmetric in Δb may better represent the observed behaviour of the PSF. For example, in the implementation by Krauß (2006) the PSF radius σ is determined based on the adaptation of a function of sigmoidal shape of the form $I(u) = a \tanh[\zeta(u - u_0)] + b$, where $\zeta = 1/(2\tilde{\sigma})$ according to Eq. (1.97), to an intensity profile extracted orthogonal to an object boundary. Setting $\sigma = \tilde{\sigma}/\sqrt{\pi/8}$ yields an identical slope at $u = u_0$ of the fitted sigmoidal profile and the profile according to Eq. (3.11) (cf. Section 1.4.8 for different representations of an ideal edge blurred by a Gaussian PSF). For the lens utilised by Krauß (2006), the dependence of σ on Δb can be represented by a Lorentz function of the form

$$\frac{1}{\sigma(\Delta b)} = \frac{\psi_1}{(\Delta b)^2 + \psi_2^2} + \psi_3 \quad (3.15)$$

with ψ_1 , ψ_2 , and ψ_3 as empirical parameters.

3.2.3.1 Definition of the Depth–Defocus Function

To obtain a relation between the depth of an object and the PSF radius σ , the image plane is assumed to be fixed at a distance b_0 from the lens while the distance z of the object varies by the amount Δz , such that $\Delta z = 0$ refers to an object in best focus. But since neither z nor Δz are known, the functional relation needs to be modelled with respect to Δb according to

$$\frac{1}{b_0 + \Delta b} + \frac{1}{z} = \frac{1}{f}. \quad (3.16)$$

A value of $\Delta b \neq 0$ refers to a defocused object point. We assume that the relation between σ and Δb is represented by Eq. (3.14) (i.e. we abandon the assumption of a proportionality between the radius c of the circle of confusion and the PSF radius σ). Hence, solving Eq. (3.16) for Δb and inserting Δb in Eq. (3.14) yields the

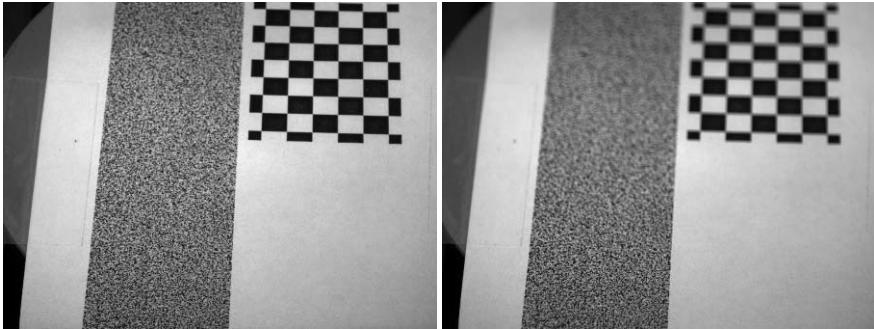


Fig. 3.9 Pixel-synchronous image pair (left: $\kappa = 8$, right: $\kappa = 2$) used for calibration of the depth-defocus function with a stationary camera.

relation

$$\mathcal{S}(z) = \frac{1}{\phi_1} \exp \left(-\frac{1}{\phi_2 \left(\frac{zf}{z-f} - b_0 \right)^2} \right) + \phi_3 \quad (3.17)$$

between the depth z and the corresponding expected PSF radius $\mathcal{S}(z)$, for which the term depth-defocus function is introduced by Kuhl et al. (2006). Calibrating the depth-defocus function $\mathcal{S}(z)$ for a given lens corresponds to determining the parameters ϕ_1 , ϕ_2 , ϕ_3 , and f in Eq. (3.17) by performing a least mean squares fit to a large set of measured (σ, z) data points.

3.2.3.2 Calibration of the Depth-Defocus Function

Stationary Camera

Barrois and Wöhler (2007) propose a depth from defocus calibration approach that requires a stationary camera calibrated geometrically e.g. with one of the methods described in Section 1.4—for all experiments described here, the semi-automatic approach by Krüger et al. (2004) described in Section 1.4.7 is employed. Two pixel-synchronous images of the calibration pattern shown in Fig. 3.9 are acquired at a small and a large aperture, respectively, as described in Section 3.2.2. The calibration pattern consists of a random noise pattern on the left, which is especially suitable for estimating the PSF radius Σ in frequency space based on Fourier-transformed image windows according to Eq. (3.12), and of a chequerboard pattern of known size on the right. The pose of the chequerboard pattern is obtained at high accuracy by extracting the corner points and applying bundle adjustment (cf. Section 1.6.1). Hence, the coordinates of each corner point are known in the camera coordinate system. The position of the random dot pattern with respect to the chequerboard pattern and thus the depth z of each pixel on the random dot pattern are also known. A window of size 32×32 pixels is extracted around each pixel on

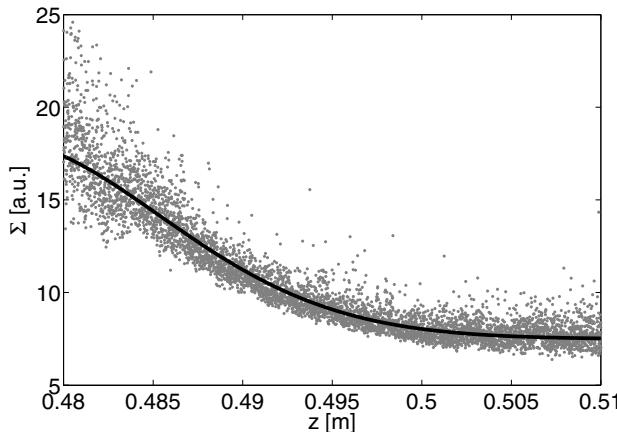


Fig. 3.10 Depth–defocus function obtained with a stationary camera based on the calibration pattern shown in Fig. 3.9.

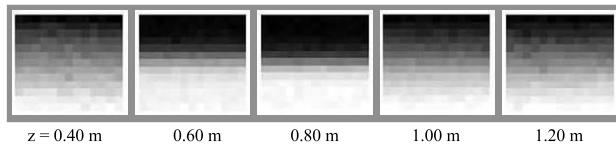


Fig. 3.11 Appearance of a ROI in different images for increasing depth z .

the random dot pattern. The window size is much larger than the PSF radius, such that the influence of neighbouring pixels outside a window on the pixels inside the window is negligible. Fitting Eq. (3.17) with $\mathcal{S}(z) = 1/\Sigma(z)$ to the measured (Σ, z) data points yields a depth–defocus function like the one shown in Fig. 3.10.

Moving Camera

The second calibration approach analyses an image sequence acquired by a camera moving towards a chequerboard pattern of known size. The corner points are tracked across the sequence using the KLT technique (Shi and Tomasi, 1994). Again, the coordinates of the corner points in the camera coordinate system are determined by bundle adjustment. Following the approach by Kuhl et al. (2006), small ROIs denoted by I_{ij} are extracted from each image j at the points on the chequerboard situated in the middle between each pair i of neighbouring corner points, such that they display edges of different sharpness (cf. Fig. 3.11). The amount of defocus is estimated for each ROI I_{ij} based on its grey value variance χ_{ij} , where the value of χ_{ij} increases with increasing sharpness. For each pair i of corner points, a parabola is fitted to the values of χ_{ij} around the maximum, which yields the index f_i of the image in which the ROI is best focused. This ROI is denoted by I_{if_i} . The fitting

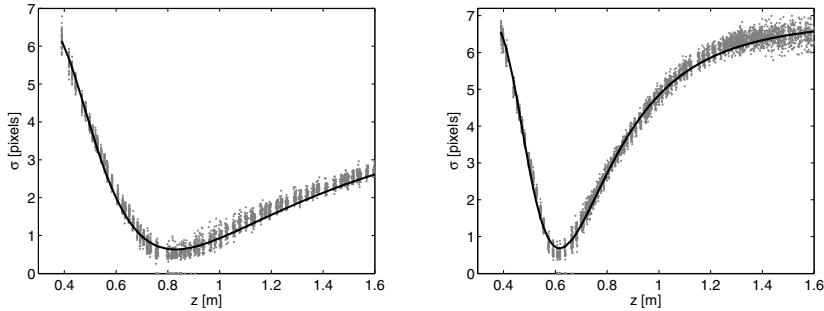


Fig. 3.12 Depth-defocus functions for a lens with $f = 12$ mm and $\kappa = 1.4$ (left) and a lens with $f = 20$ mm and $\kappa = 2.4$ (right), fitted to the measured data points according to Eq. (3.17), respectively.

procedure is applied to introduce robustness with respect to pixel noise. An alternative technique is to directly select the ROI with the maximum grey value variance, i.e. $f_i = \arg \max_j \chi_{ij}$. Using the variance as a defocus measure is convenient for the chequerboard pattern, but depending on the texture of the object surfaces, the high-frequency integral as defined in the next paragraph (cf. Eq. (3.20)) may be more appropriate.

The ROI I_{ij} extracted between corner pair i in image j is given by a convolution of the best-focused version I_{if_i} with the appropriate PSF of radius σ_{ij} according to

$$I_{ij} = G(\sigma_{ij}) * I_{if_i}. \quad (3.18)$$

Ideally, I_{if_i} , I_{ij} , and $G(\sigma_{ij})$ are defined on an infinite domain, but in practice they are represented by image windows of finite size. For each ROI I_{ij} the value σ_{ij} relative to the previously determined best focused ROI I_{if_i} needs to be determined. It would of course be possible to estimate the PSF radius Σ in frequency space according to Eq. (3.12), which is in turn proportional to $1/\sigma$. However, it is illustrative at this point to demonstrate that different techniques can be applied to estimate the PSF. Hence, based on the bisection method, we directly determine the value of σ_{ij} for which the root mean square deviation

$$E_{\text{ROI}} = \sqrt{\langle [G(\sigma_{ij}) * I_{if_i} - I_{ij}]^2 \rangle} \quad (3.19)$$

between the best-focused ROI convolved with a Gaussian PSF of radius σ_{ij} and the currently observed ROI becomes minimal. In Eq. (3.19), the expression $\langle \dots \rangle$ denotes the average over the pixels of the ROI. The PSF does not vary rapidly across the chequerboard pattern, such that it is not required to employ more sophisticated techniques like the block shift-variant blur model (Chaudhuri and Rajagopalan, 1999). As the depth z_{ij} is known for each ROI I_{ij} , the depth-defocus function is then obtained by fitting Eq. (3.17) to all determined (σ_{ij}, z_{ij}) data points. Two examples are

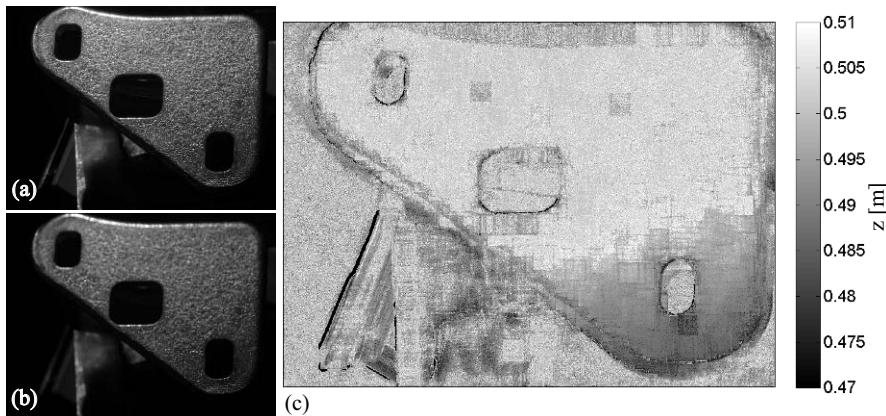


Fig. 3.13 Example of a depth map obtained with the depth from defocus approach based on two images acquired with a stationary camera. (a) Sharp input image, acquired at $\kappa = 8$. (b) Unsharp input image, acquired at $\kappa = 2$. (c) Resulting depth map. For the black pixels no depth value could be computed. The pixel grey values are absolutely scaled in metres in the camera coordinate system.

shown in Fig. 3.12 for lenses with focal lengths of 12 mm and 20 mm and f-stop numbers of 1.4 and 2.4, respectively. Objects at a distance of about 0.8 m and 0.6 m are in focus, corresponding to the minimum of the curve, respectively. At the minimum of the depth–defocus function one would expect a PSF radius of zero, but since E_{ROI} according to Eq. (3.19) is not very sensitive to σ_{ij} for $\sigma_{ij} < 1$ pixel, and we always have $\sigma_{ij} \geq 0$, the influence of pixel noise may yield small nonzero positive values of σ_{ij} near the minimum, leading to the behaviour observed in Fig. 3.12.

3.2.3.3 Determination of the Depth Map

Stationary Camera

An example result of the depth from defocus method based on a pixel-synchronous pair of images acquired at different f-stop numbers with a stationary camera is shown in Fig. 3.13 for a door hinge (Barrois and Wöhler, 2007). This procedure may be automated using a lens equipped with a motorised iris. The raw cast iron surface of this automotive part displays a sufficient amount of texture to allow a reasonable estimation of the position-dependent PSF radius. The depth from defocus method was calibrated according to the method illustrated in Fig. 3.9. The depth–defocus function shown in Fig. 3.10 was used to determine the depth map in Fig. 3.13c, clearly illustrating that the plane surface is tilted such that its lower part is closer to the camera than its upper part.

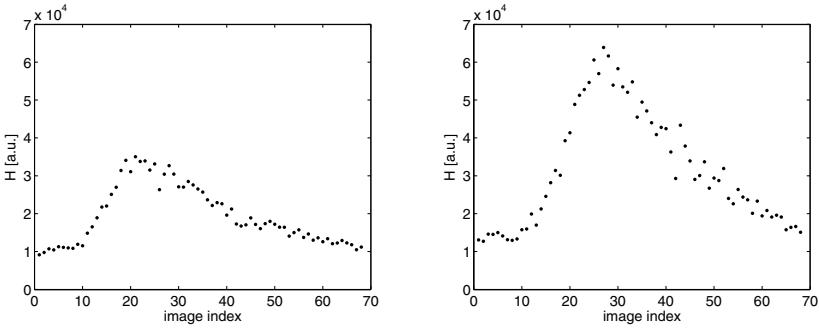


Fig. 3.14 Image index vs. defocus measure H according to Eq. (3.20) for two different tracked image features.

Moving Camera

Processing the image sequence acquired with a moving camera involves the extraction of salient features from the image sequence. These features are tracked using the KLT technique (Shi and Tomasi, 1994), which is based on the Harris corner detector (Harris and Stephens, 1988) and takes into account affine transformations between subsequent views. A ROI of constant size is extracted around each feature point at each time step. For each tracked feature, the best focused image has to be identified in order to obtain the increase of defocus for the other images. We found that the grey value variance as a measure for defocus does not perform well on features other than black-and-white corners. Instead we make use of the amplitude spectrum $|\mathcal{J}(\omega_u, \omega_v)|$ of the ROI extracted around the feature position (Kuhl et al., 2006). High-frequency components of the amplitude spectrum denote sharp details, whereas low-frequency components refer to large-scale features. Hence, the integral over the high-frequency components can be used as a measure for the sharpness of a certain tracked feature. However, since the highest-frequency components are considerably affected by pixel noise and defocus has no perceivable effect on the low-frequency components, a frequency band between ω_0 and ω_1 is taken into account according to

$$H = \iint_{\omega_0}^{\omega_1} |\mathcal{J}(\omega_u, \omega_v)| d\omega_u d\omega_v \quad (3.20)$$

with $\omega_0 = \frac{1}{4}\omega_{\max}$ and $\omega_1 = \frac{3}{4}\omega_{\max}$, where ω_{\max} is the maximum spatial frequency. The amount of defocus increases with decreasing value of H . The defocus measure H is used to determine the index of the best focused ROI for each tracked feature in the same manner as the grey value variance χ in Section 3.2.3.2. Fig. 3.14 shows the behaviour of H for two different example features tracked across a sequence. It is obvious that the value of H cannot be used for comparing the amount of defocus among different feature points since the maximum value of H depends on the image content. The same is true for the grey value variance. Hence, both the integral

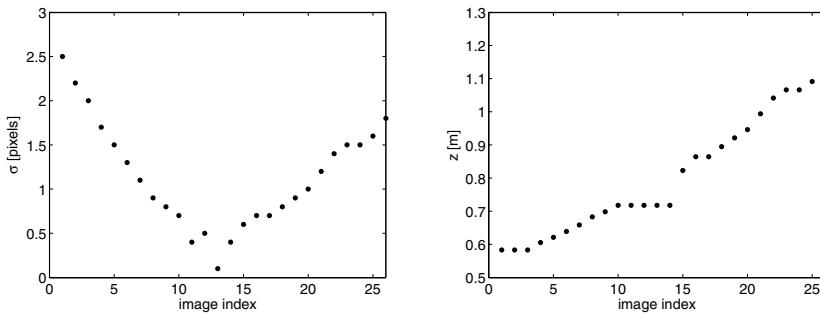


Fig. 3.15 Typical behaviour of the PSF radius σ (left) and depth z (right) across an image sequence for an example feature.

H of the amplitude spectrum as well as the grey value variance are merely used for determining the index of the image in which a certain feature is best focused. Other defocus measures may be used, as e.g. those regarded by Nayar (1989), but it depends on the individual sequence which defocus measure is most useful to determine the sharpest ROI of the sequence. A parabola is fitted to the measurements around the maximum in order to increase robustness with respect to noise.

The PSF radius is computed relative to the best focused ROI according to Eq. (3.19). The depth z is obtained by inverting the depth-defocus function $\mathcal{S}(z)$ according to Eq. (3.17). The encountered two-fold ambiguity is resolved by using information about the direction of camera motion, which is obtained either based on a-priori knowledge or by performing a structure from motion analysis e.g. according to Section 1.2, yielding information about the path of the camera. If the estimated PSF radius is smaller than the minimum of $\mathcal{S}(z)$, the depth is set to the value of z at which the function $\mathcal{S}(z)$ obtains its minimum. For an example feature, the computed PSF radii and the inferred depth values are shown in Fig. 3.15, illustrating that the camera moves away from the object at an approximately constant velocity.

An integration of the techniques of depth from defocus based on an image sequence acquired with a moving camera and structure from motion in order to obtain an accurate and absolutely scaled three-dimensional scene reconstruction is described in detail in Section 4.1.

3.2.3.4 Estimation of the Useful Depth Range

At this point it is useful to examine which focal length and lens aperture are required to obtain depth values of a given accuracy with the depth from defocus method. Assume that for a lens of focal length f_1 , an object is well focused at depth z_0 , and a certain amount of defocus is observed at depth hz_0 , where the factor h is assumed to be close to 1 with $|h - 1| \ll 1$. The depth offset $\Delta z = (h - 1)z_0$ implies a circle of confusion of radius c_1 with

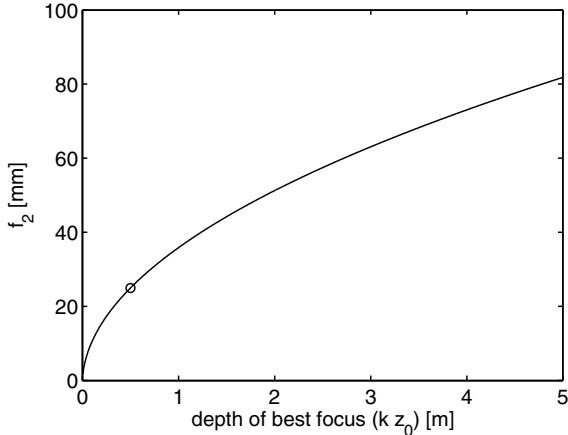


Fig. 3.16 Focal length f_2 according to Eq. (3.24) required to obtain the same relative depth resolution $|h - 1|$ as a lens of $f_1 = 25$ mm at a distance of $z_0 = 0.5$ m (open circle).

$$c_1 = \frac{1}{2\kappa} \left(\frac{1}{f_1^{-1} - z_0^{-1}} - \frac{1}{f_1^{-1} - (hz_0)^{-1}} \right). \quad (3.21)$$

Now let the focal length be changed to f_2 and the object depth be set to a larger distance kz_0 with $k > 1$. The radius c_2 of the corresponding circle of confusion is readily obtained by

$$c_2 = \frac{1}{2\kappa} \left(\frac{1}{f_2^{-1} - (kz_0)^{-1}} - \frac{1}{f_2^{-1} - (hkz_0)^{-1}} \right). \quad (3.22)$$

The f-stop number κ and the pixel size remain unchanged. Since the radius of the circle of confusion is a monotonically increasing function of the PSF radius σ , we assume that observing the same amount of defocus in both scenarios implies an identical radius of the corresponding circle of confusion. With the abbreviations

$$\begin{aligned} K &= \frac{1}{f_1^{-1} - z_0^{-1}} - \frac{1}{f_1^{-1} - (hz_0)^{-1}} \\ L_1 &= \frac{1 - h^{-1}}{kz_0} \\ L_2 &= \frac{1 + h^{-1}}{kz_0} \\ M &= \frac{1}{hk^2 z_0^2}, \end{aligned} \quad (3.23)$$

setting $c_1 = c_2$ yields the focal length f_2 according to

$$f_2 = \left(\frac{L_2}{2} \pm \sqrt{\frac{L_1}{K} - M + \frac{L_2^2}{4}} \right)^{-1}. \quad (3.24)$$

Only the solution with the plus sign before the square root yields positive values for f_2 . Close inspection of Eq. (3.24) reveals that the value of f_2 is approximately proportional to \sqrt{k} independent of the chosen value of h as long as $|h - 1| \ll 1$. Hence, for constant f-stop number κ , constant relative variation $|h - 1|$ of the object depth z , and constant pixel size, the required focal length and thus also the aperture of the lens are largely proportional to $\sqrt{z_0}$. Fig. 3.16 depicts the focal length f_2 according to Eq. (3.24) required to obtain the same relative depth resolution $|h - 1|$ as a lens of $f_1 = 25$ mm at a distance of $z_0 = 0.5$ m. Hence, for standard video cameras and lenses like those used for the experiments in this work, the depth from defocus approach is favourably applied in the close-range domain where z_0 is not larger than a few metres, since otherwise the field of view would become too small and the required lens aperture too large.

Chapter 4

Integrated Frameworks for Three-dimensional Scene Reconstruction

It has been shown in Chapters 1–3 that the problem of three-dimensional scene reconstruction can be addressed with a variety of approaches. Geometric approaches (cf. Chapter 1) rely on correspondences of points or higher-order features between several images of a scene acquired either with a moving camera or with several cameras from different viewpoints. These methods are accurate and do not require a-priori knowledge about the scene or the cameras used. On the contrary, as long as the scene points are suitably distributed they do not only yield the scene structure but also the intrinsic and extrinsic camera parameters, i.e. they perform a camera calibration simultaneously with the scene reconstruction. Geometric approaches, however, are restricted to parts of the scene with a sufficient amount of texture to decide which part of a certain image belongs to which part of another image. Occlusions may occur, such that corresponding points or features are hidden in some images, the appearance of the objects may change from image to image due to perspective distortions, and in the presence of objects with non-Lambertian surface properties the apparent intensity distribution across the scene may vary strongly from image to image, such that establishing correspondences between images becomes inaccurate or impossible at all.

Photometric approaches to three-dimensional scene reconstruction (cf. Chapter 2) exploit the intensity distribution across the image by determining the surface normal for each image pixel. They are especially suited for textureless parts of the scene, but if several images of the scene are available, it is also possible to separate texture from shading effects. Drawbacks are that the reflectance properties of the regarded surfaces need to be known, the reconstructed scene structure may be ambiguous especially with respect to its large-scale properties, and small systematic errors of the estimated surface gradients may cumulate into large depth errors on large scales.

Real-aperture approaches (cf. Chapter 3) directly estimate the depth of scene points based on several images acquired at different focus settings. While the depth from focus method determines depth values based on the configuration of maximum sharpness, the problem of depth estimation reduces to an estimation of the PSF difference between images for the depth from defocus method. Depth from focus is

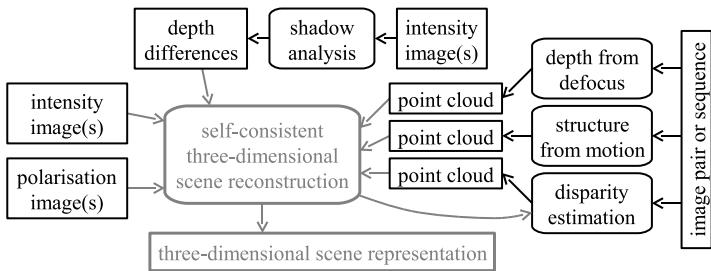


Fig. 4.1 Summary of geometric, photometric, and real-aperture methods described in Chapters 1–3. This chapter regards different integrated frameworks to obtain a three-dimensional scene representation.

very accurate but also time-consuming due to the large number of images required, such that it should be preferentially applied in measurement applications rather than computer vision scenarios. Depth from defocus can be easily applied and no a-priori knowledge about the scene needs to be available, but a sufficient amount of surface texture is required. Due to the fact that estimation of the PSF is sensitive with respect to pixel noise, the resulting depth values tend to be rather inaccurate.

These considerations illustrate that each of the approaches described in Chapters 1–3 has its specific advantages and drawbacks. Some of the techniques are complementary; as an example, geometric methods yield three-dimensional point clouds describing textured parts of the scene while photometric methods may be able to reconstruct textureless regions between the points. Hence, just like the human visual system achieves a dense three-dimensional scene reconstruction based on combinations of different cues, it appears to be favourable for computer vision systems to integrate different three-dimensional scene reconstruction methods into a unifying framework (cf. Fig. 4.1). This chapter describes several approaches of this kind and discusses their specific preconditions, advantages, limitations, and preferential application domains.

4.1 Monocular Three-dimensional Scene Reconstruction at Absolute Scale

This section describes a method for combining geometric and real-aperture methods for monocular three-dimensional reconstruction of static scenes at absolute scale introduced by Kuhl et al. (2006). The algorithm (cf. Fig. 4.2) relies on a sequence of images of the object acquired by a monocular camera of fixed focal setting from different viewpoints. Object features are tracked over a range of distances from the camera with a small depth of field, leading to a varying degree of defocus for each feature. Information on absolute depth is obtained based on a depth from defocus approach. The parameters of the point spread functions estimated by depth from defocus are used as a regularisation term for a structure from motion algorithm

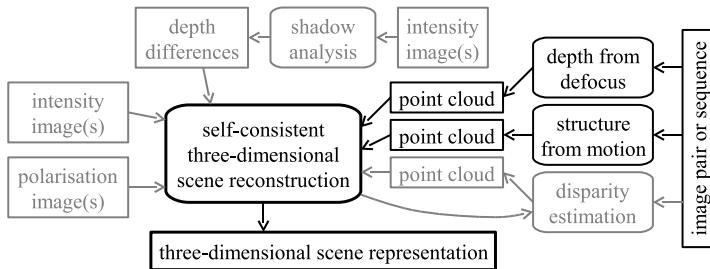


Fig. 4.2 Overview of the integrated approach to monocular three-dimensional scene reconstruction at absolute scale.

based on bundle adjustment. The reprojection error obtained from bundle adjustment and the absolute depth error obtained from depth from defocus are simultaneously minimised for all tracked object features. The proposed method yields absolutely scaled three-dimensional coordinates of the scene points without any prior knowledge about scene structure and the camera motion. The implementation of the proposed method is described both as an offline and as an online algorithm. Evaluating the algorithm on real-world data we demonstrate that it yields typical relative scale errors of a few percent. We examine the influence of random effects, i.e. the noise of the pixel grey values, and systematic effects, caused by thermal expansion of the optical system or by inclusion of strongly blurred images, on the accuracy of the three-dimensional reconstruction result.

To our knowledge, prior to the work by Kuhl et al. (2006) no attempt has been made to combine the precise relative scene reconstruction of structure from motion with the absolute depth data of depth from defocus. A work related to the presented approach has been published by Myles and da Vitoria Lobo (1998), where a method to recover affine motion and defocus simultaneously is proposed. However, the spatial extent of the scene is not reconstructed by their method, since it requires planar objects. In contrast, the method described in this section yields a three-dimensional scene reconstruction at absolute scale based on an image sequence acquired with a monocular camera.

4.1.1 Combining Motion, Structure, and Defocus

The structure from motion analysis employed in this section involves the extraction of salient features from the image sequence which are tracked using the KLT technique (Shi and Tomasi, 1994). A depth from defocus analysis is performed for these features according to the method introduced by Kuhl et al. (2006) as described in detail in Section 3.2.3. We found experimentally that the random scatter of the feature positions extracted by the KLT tracker, which is due to the noise of the pixel grey values, is largely independent of the image blur for PSF radii smaller than 5 pixels and is always of the order 0.1 pixels. However, more features are detected

and less features are lost by the tracker when the tracking procedure is started on a well-focused image. Hence, the tracking procedure is repeated, starting from the sharpest image located near the middle of the sequence which displays the largest value of H according to Eq. (3.20) averaged over all previously detected features, proceeding towards either end of the sequence and using the ROIs extracted from this image as reference patterns. The three-dimensional coordinates ${}^W\mathbf{x}_k$ of the scene points are then computed by extending the classical bundle adjustment error term (1.11) with an additional error term that takes into account the depth from defocus measurements, leading to the combined error term

$$\begin{aligned} E_{\text{comb}} = & \sum_{i=1}^N \sum_{k=1}^K \left[\left\| {}^{S_i}_{I_i} T^{-1} \left(\mathcal{P} \left({}^{C_i}_W T, \{c_j\}_i, {}^W \mathbf{x}_k \right) - {}^{S_i} \mathbf{x}_k \right) \right\|^2 \right. \\ & \left. + \alpha \left(\mathcal{S} \left(\left[{}^{C_i}_W T {}^W \mathbf{x}_k \right]_z \right) - \sigma_{ik} \right)^2 \right]. \end{aligned} \quad (4.1)$$

The error term E_{comb} is minimised with respect to the N camera transforms ${}^{C_i}_W T$ and the K scene points ${}^W \mathbf{x}_k$. The value of σ_{ik} corresponds to the estimated PSF radius for feature k in image i , α is a weighting factor, \mathcal{S} the depth-defocus function that yields the expected PSF radius of feature k in image i , and $\left[{}^{C_i}_W T {}^W \mathbf{x}_k \right]_z$ the z coordinate (depth) of a scene point transformed to camera coordinate system i . The estimated radii σ_{ik} of the Gaussian PSFs define a regularisation term in Eq. (4.1), such that absolutely scaled three-dimensional coordinates ${}^W \mathbf{x}_k$ of the scene points are obtained, where for convenience the world coordinate system is set equal to camera coordinate system 1. The values of ${}^W \mathbf{x}_k$ are initialised according to the depth values estimated based on the depth from defocus approach. To minimise the error term E_{comb} the Levenberg-Marquardt algorithm (Press et al., 1992) is used.

4.1.2 Online Version of the Algorithm

The integrated three-dimensional reconstruction method described so far is an offline algorithm. The error term (4.1) is minimised once for the complete image sequence after acquisition. This section describes an online version of the proposed combination of structure from motion and depth from defocus, processing the acquired images instantaneously and thus generating a refined reconstruction result as soon as a new image has been acquired. This is a desired property for systems e.g. in the context of mobile robot navigation or in-situ exploration.

The online version of the algorithm starts by acquiring the current image. Features already present in the previous image are searched by the KLT tracker, and lost features may be replaced with new ones. The amount of defocus is obtained for each feature within the current frame based on the high-frequency integral H of the ROI around the feature position (cf. Eq. (3.20)). For each tracked feature, the

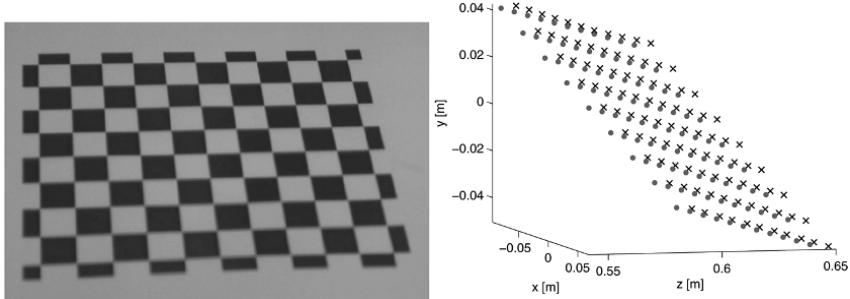


Fig. 4.3 True (dots) and reconstructed (crosses) three-dimensional pose of the chequerboard, obtained with a weight factor of $\alpha = 0.42$.

best focused ROI is determined by fitting a second-degree polynomial to the values of H . Possible candidates that may already have passed their point of maximum sharpness are identified based on a threshold rating. The initial depth value is then computed for each tracked feature by estimating the PSF radius σ as outlined in Section 3.2.3.2.

The Levenberg-Marquardt scheme which minimises the error function (4.1) and thus determines the camera transforms and three-dimensional feature points is initialised with the depth from defocus estimates for each feature point. To reduce the effect of outliers, the M-estimator technique with the “fair” weighting function $w(x) = 1/(1 + |x|/c)$ is employed, where $c = 1.3998$ is recommended as a favourable choice by Rey (1983). The optimisation result for the current time step is used as an initialisation to the subsequent time step upon acquisition of the next image.

4.1.3 Experimental Evaluation Based on Tabletop Scenes

In the tabletop experiments a Baumer industrial CCD camera with an image size of 1032×776 pixels, equipped with a Cosmicar-Pentax video lens of 12 mm focal length, was used. The pixels are skewless and of size $4.65 \times 4.65 \mu\text{m}^2$. In order to validate our approach we first reconstructed a planar object with reference points of precisely known mutual distance. A chequerboard as shown in Fig. 4.3 with 10×8 squares of size $15 \times 15 \text{ mm}^2$, respectively, was used. The 99 corners serve as features and were extracted in every image using the method described by Krüger et al. (2004) to assure sub-pixel accuracy. The true pose of the chequerboard was obtained according to Bouguet (2007) based on the given size of the squares (cf. Section 1.6.1). Note that Bouguet (2007) determines the true pose of the chequerboard by applying a least mean squares fit on a single image, whereas the proposed algorithm estimates the three-dimensional structure of a scene by means of a least mean

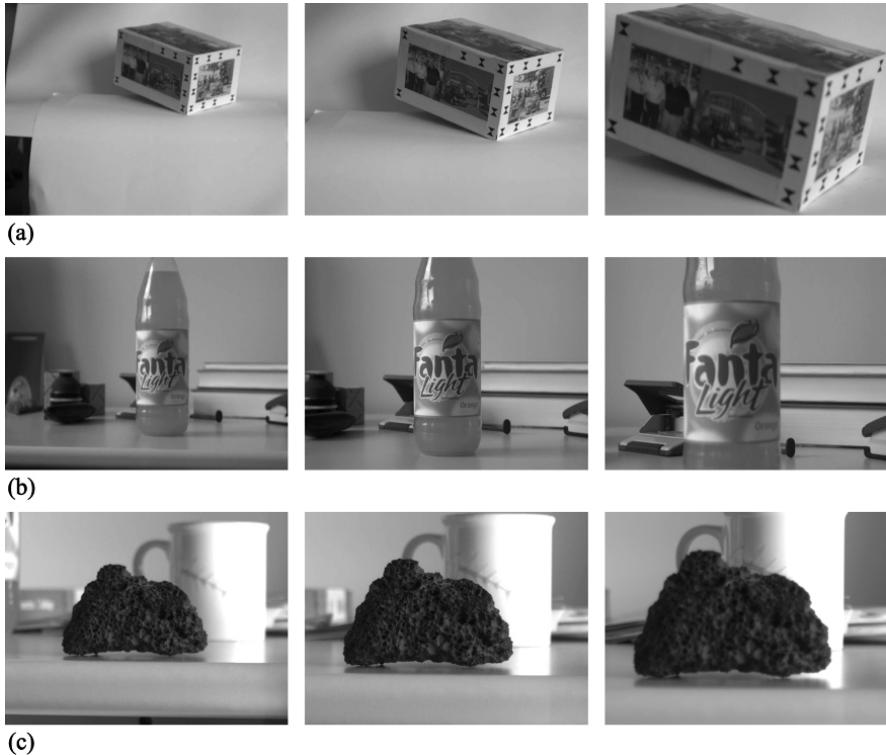


Fig. 4.4 Images from the beginning, the middle, and the end of (a) the cuboid sequence, (b) the bottle sequence, and (c) the lava stone sequence.

squares fit applied to the whole image sequence. Comparing the obtained results with the determined true pose of the object is actually a comparison between two methods conducting different least mean squares fits.

Experiments involving real-world objects were conducted based on image sequences that display a cuboid with markings at known positions, a bottle of known diameter, and a lava stone with a pronounced surface texture. Images from the beginning, the middle, and the end of each sequence are shown in Fig. 4.4 (Wöhler et al., 2009).

4.1.3.1 Evaluation of the Offline Algorithm

To analyse the three-dimensional reconstruction results of the combined structure from motion and depth from defocus approach, we define several error measures. The deviation E_{rec} of the reconstructed three-dimensional scene point coordinates ${}^W\mathbf{x}_k$ from the ground truth values ${}^W\mathbf{x}_k^{\text{true}}$ is given by

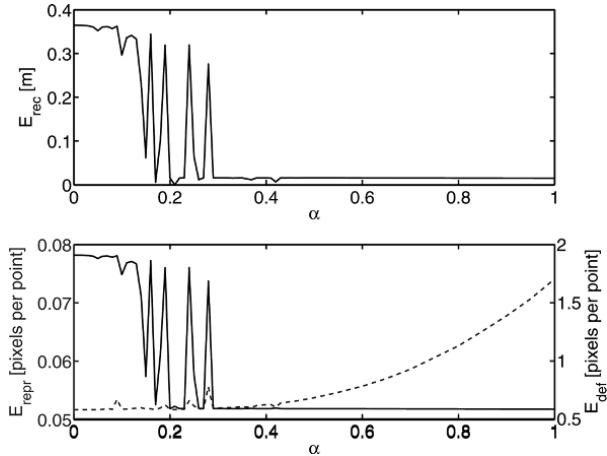


Fig. 4.5 Dependence of E_{rec} (upper diagram), E_{repr} (lower diagram, dashed curve, left axis), and E_{def} (lower diagram, solid curve, right axis) on the weight parameter α .

$$E_{\text{rec}} = \sqrt{\frac{1}{K} \sum_{k=1}^K \|{}^W \mathbf{x}_k - {}^W \mathbf{x}_k^{\text{true}}\|^2}, \quad (4.2)$$

where K denotes the number of scene points. To determine an appropriate weight parameter α in Eq. (4.1) we computed E_{rec} for different α values in the range between 0 and 1. For $\alpha = 0$ the global minimisation is equivalent to structure from motion initialised with the calculated depth from defocus values. One must keep in mind, however, that the absolute scaling factor is then part of the gauge freedom of the bundle adjustment method, resulting in a corresponding “flatness” of the error function. Small α values lead to an unstable convergence. The value of E_{rec} levels off to 16 mm for $\alpha \approx 0.3$ and obtains its minimum value of 7 mm for $\alpha = 0.42$. The root mean square deviation of the reconstructed size of the squares from the true value of 15 mm then amounts to 0.2 mm or 1.3 percent. The most accurate scene reconstruction results are obtained with α between 0.3 and 0.5. The reconstructed three-dimensional scene points ${}^W \mathbf{x}_k$ for $\alpha = 0.42$ are illustrated in Fig. 4.3, the dependence of E_{rec} on α in Fig. 4.5 (upper diagram).

In addition to the reconstruction error E_{rec} , a further important error measure is the reprojection error

$$E_{\text{repr}} = \sqrt{\frac{1}{KN} \sum_{i=1}^N \sum_{k=1}^K \left\| {}^{S_i} T^{-1} \left(\mathcal{P} \left({}^C_i T, \{c_j\}_i, {}^W \mathbf{x}_k \right) - {}^{S_i} \mathbf{x}_k \right) \right\|^2}, \quad (4.3)$$

denoting the root-mean-square deviation between the measured two-dimensional feature positions ${}^{S_i} \mathbf{x}_k$ in the sensor plane after transformation by ${}^{S_i} T^{-1}$ into the image plane, and the reconstructed three-dimensional scene points ${}^W \mathbf{x}_k$ reprojected

into the image plane using the reconstructed camera transforms ${}^{C_i}_W T$. The number of images is denoted by N .

The defocus error corresponds to the root-mean-square deviation between measured and expected radii σ_{ik} of the Gaussian PSFs according to

$$E_{\text{def}} = \sqrt{\frac{1}{KN} \sum_{i=1}^N \sum_{k=1}^K \left(\mathcal{S} \left(\begin{bmatrix} {}^{C_i}_W T & {}^W \mathbf{x}_k \end{bmatrix}_z \right) - \sigma_{ik} \right)^2}. \quad (4.4)$$

Fig. 4.5 (bottom) shows the relation between the weight parameter α , the reprojection error E_{repr} , and the defocus error E_{def} . For $\alpha > 0.3$ the defocus error stabilises to 0.58 pixels per feature. Larger α values lead to a stronger influence of the depth from defocus values on the optimisation result, leading to an increasing reprojection error E_{repr} due to the inaccuracy of the estimated σ_{ik} values. Although the depth values derived by depth from defocus are noisy, they are sufficient to establish a reasonably accurate absolute scale. Hence, this first evaluation shows that the combined approach is able to reconstruct scenes at absolute scale without prior knowledge. As shown in Section 3.2.3.4, the described approach is favourably used in the close-range domain at object distances of not more than a few metres as long as standard video cameras and lenses (focal length below about 20 mm, pixel size around 10 μm , image size approximately 10^6 pixels) are used.

Further experiments regarding several real-world objects are described in the following paragraphs. Here we report results obtained based on the image sequences examined by Kuhl et al. (2006) (cf. Fig. 4.4), which were refined and somewhat extended in later experiments (d'Angelo, 2007; Wöhler et al., 2009). In order to distinguish random fluctuations from systematic deviations, the error measures for 100 runs were computed for each example, respectively. For the utilised camera, the noise of the pixel grey values is proportional to the square root of the grey values themselves, where the proportionality constant has been determined empirically. For each of the 100 runs, a corresponding amount of Gaussian noise was added to the images of the sequence. The noise leads to a standard deviation of the feature positions ${}^{S_i} \mathbf{x}_k$ obtained by the KLT tracker of 0.1 pixels.

Cuboid Sequence

To demonstrate the performance of our approach on a non-planar test object of known geometry, the method was applied to the cuboid-shaped object shown in Fig. 4.4a. This object displays a sufficient amount of texture to generate “good features to track” according to Shi and Tomasi (1994). In addition, black markers on white background with known mutual distances are placed near the edges of the cuboid. As described in Section 4.1.1, feature points are extracted and tracked using the KLT algorithm, and the three-dimensional coordinates of the scene points are obtained by minimising the error term E_{comb} according to Eq. (4.1) with $\alpha = 0.5$ as the weight parameter. This value of α will be used in all subsequent experiments. Tracking outliers are removed by determining the features with associated very large

Table 4.1 Summary of the evaluation results for the offline version of the combined structure from motion and depth from defocus algorithm. The indicated error intervals correspond to the standard deviations.

Sequence	Length (images)	E_{repr} (pixels)	E_{def} (pixels)	Reference length (mm) Ground truth	Reconstruction
Cuboid	46	0.642	0.636	32.0	34.1 ± 1.6
Bottle	26	0.747	0.387	80.0	82.8 ± 1.4
Lava stone	15	0.357	0.174	60.0	58.3 ± 0.8

reprojection errors of more than $3E_{\text{repr}}$ and neglecting them in a subsequent second bundle adjustment step.

The three-dimensional reconstruction result for the cuboid sequence is shown in Fig. 4.6. We obtain for the reprojection error $E_{\text{repr}} = 0.642$ pixels and for the defocus error $E_{\text{def}} = 0.636$ pixels. In order to verify the absolute scale, we compared the reconstructed pairwise distances between the black markers on the object (as seen e.g. in the top right corner of the front side) to the corresponding true distances. For this comparison we utilised a set of six pairs of markers with an average true distance of 32.0 mm. The corresponding reconstructed average distance amounts to 34.1 mm (cf. Table 4.1).

Bottle Sequence

In a further experiment, our combined structure from motion and depth from defocus approach was applied to a bottle, the lower part of which is of cylindrical shape with a diameter of 80.0 mm, as shown in Fig. 4.4b. No background features are selected by the algorithm since for none of these features maximum sharpness could be observed in the acquired sequence. The three-dimensional reconstruction result is shown in Fig. 4.7. We obtained $E_{\text{repr}} = 0.747$ pixels for the reprojection error and $E_{\text{def}} = 0.387$ pixels for the defocus error. To quantify the accuracy of the determined absolute scale, we compared the diameter of the reconstructed object with that of the real bottle. For this purpose the reconstructed points were projected on the xz plane of the camera coordinate system, in which the x axis is parallel to the image rows, the y axis is parallel to the image columns (and thus to the central axis of the cylinder), and the z axis is parallel to the optical axis. A circle fit was applied to the three-dimensional point cloud projected on the xz plane, yielding a diameter of the bottle of 82.8 mm, which is in good correspondence with the known diameter of 80.0 mm (cf. Fig. 4.7 and Table 4.1).

Lava Stone Sequence

As a further real-world object we examined the lava stone shown in Fig. 4.4c. The three-dimensional reconstruction result is shown in Fig. 4.8. The shaded view of the object was obtained by Delaunay triangulation of the reconstructed set of three-

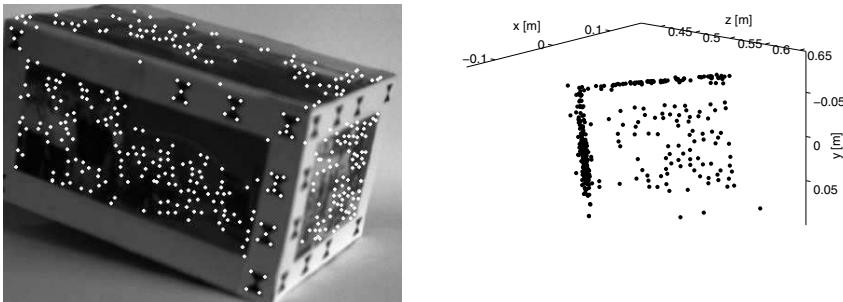


Fig. 4.6 Three-dimensional reconstruction of the cuboid.

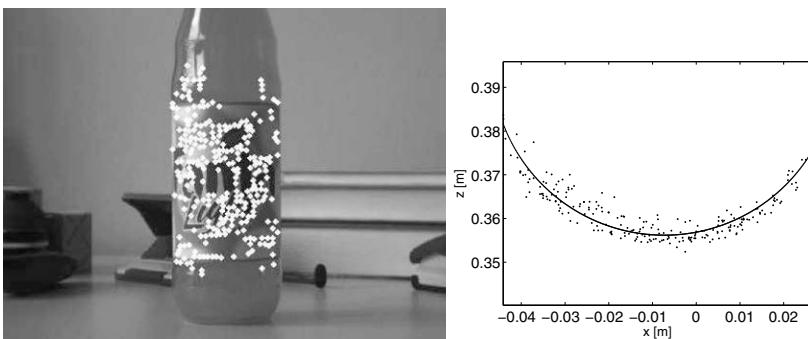


Fig. 4.7 Three-dimensional reconstruction of the cylindrical surface of the bottle.

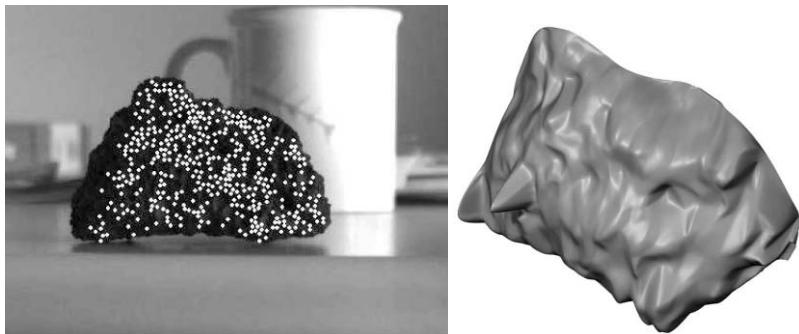


Fig. 4.8 Three-dimensional reconstruction of the lava stone. The cusp visible in the left part of the reconstructed surface is produced by three outlier points.

dimensional points. The cusp visible in the left part of the reconstructed surface is due to three outlier points presumably generated by inaccurately determined feature positions. For this sequence we obtain reconstruction errors of $E_{\text{repr}} = 0.357$ pixels and $E_{\text{def}} = 0.174$ pixels. The very low standard deviations indicate a high consistency of the depth from defocus measurements due to the strong texture of the object surface. Two well-defined points on the object with a true mutual distance of

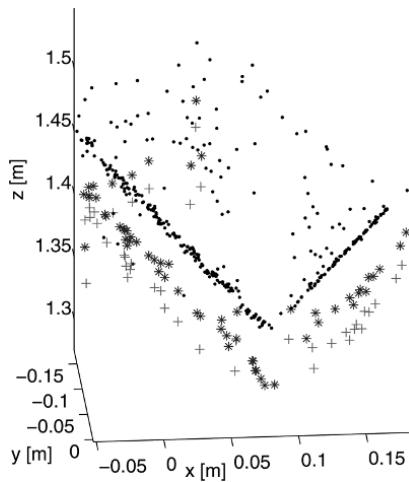


Fig. 4.9 Three-dimensional reconstruction of the cuboid, obtained with the online algorithm. The result is displayed after 5 (crosses), 9 (stars), and 25 (dots) processed images. The first three-dimensional reconstruction result is obtained after acquisition of the 21st image.

60.0 mm were chosen as a reference for estimating the accuracy of the inferred absolute scale. The reconstructed distance of the reference points amounts to 58.3 mm, which is consistent with the known distance of 60.0 mm (cf. Table 4.1).

4.1.3.2 Evaluation of the Online Algorithm

A systematic evaluation of the online algorithm was performed for the cuboid, bottle, and lava stone sequences (Wöhler et al., 2009), again based on the image sequences regarded by Kuhl (2005) and Kuhl et al. (2006). The online algorithm generally starts with a very noisy set of three-dimensional points due to the small number of features already having reached their maximum sharpness at the beginning of the image sequence. After processing an increasing number of images, the three-dimensional reconstruction result more and more resembles the result of the offline algorithm. The results are not identical because generally a similar but not identical index f_i (cf. Eq. (3.18)) is determined for the best focused ROI of each tracked feature by the offline and the online algorithm, respectively.

In the cuboid example shown in Fig. 4.9, the first three-dimensional reconstruction result can be obtained after processing 21 images. After 5 further processed images, still only 40 features are available which have passed their point of maximum sharpness, respectively. The same six pairs of reference points as those regarded in the offline experiments were used for determining the accuracy of the inferred absolute scale. The standard deviations across the 100 online runs are indicated by

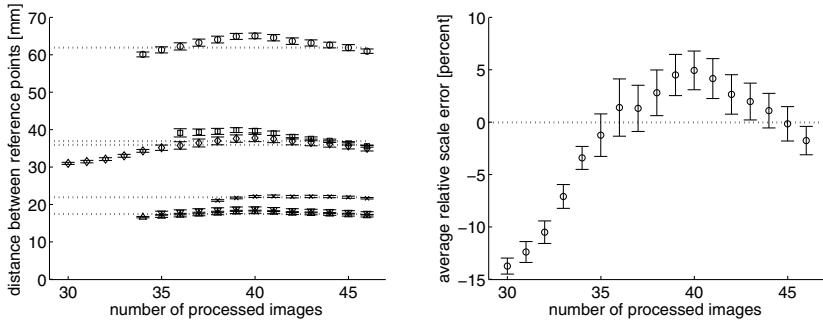


Fig. 4.10 Left: Behaviour of the distances between six pairs of reference points on the cuboid surface, obtained with the online algorithm, for increasing number of iterations, compared to the corresponding true values. The standard deviations across the 100 online runs are indicated by error bars. Right: Relative accuracy of the inferred absolute scale, given by the average relative deviation between the measured and true absolute distances between those pairs of reference points already having passed their point of maximum sharpness, for increasing number of processed images.

error bars in Figs. 4.10–4.12. The behaviour of the reconstruction accuracy with increasing number of processed images is illustrated in Fig. 4.10, where Gaussian noise was added to the images before starting a processing run in the same manner as in the offline experiments described in the previous paragraph. For less than 38 processed images, not all reference points have passed their point of maximum sharpness, such that their three-dimensional positions are not yet available. The average relative scale error shown in Fig. 4.10 is thus derived from those pairs of reference points that have already passed their point of maximum sharpness.

Fig. 4.9 suggests that with increasing number of available features the reconstructed size and shape of the cuboid become more accurate. Fig. 4.10 shows that the best value for the inferred absolute scale is already obtained after 45 processed images. The reason for this observation is the fact that in the last image of the sequence (i.e. image 46) all tracked features appear strongly blurred, leading to higher inaccuracies of the depth values derived from the estimated PSF radii σ_{ik} due to the low slope of the depth–defocus function (3.17) far away from its minimum.

Analogous experimental evaluations were conducted for the bottle sequence and the lava stone sequence. The results are shown in Fig. 4.11. For the bottle sequence, the average accuracy of the determined absolute scale (represented by the inferred diameter of the bottle) is better than 3.0 percent already after 12 processed images. However, at the beginning of the sequence the random scatter across the 100 runs is about two times larger than at the end (more than 23 processed images). The final difference between measured and true absolute scale corresponds to 1.9 standard deviations. For the lava stone sequence, the determined absolute scale is about 1.2 percent too small after 12 processed images, corresponding to one standard deviation. The deviation corresponds to several standard deviations and is most likely of systematic nature when 13 and more images are processed. The fact that the last three images of the lava stone sequence are strongly blurred presumably is the

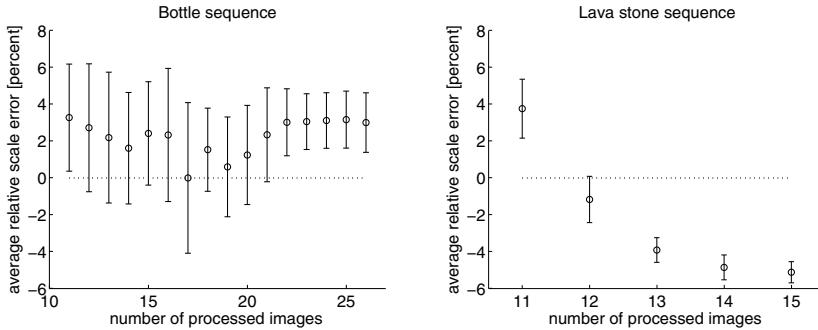


Fig. 4.11 Relative deviation between measured and true absolute scale for increasing number of processed images for the bottle sequence (left) and the lava stone sequence (right).

main reason for this behaviour. Hence, it is favourable for this sequence to adopt the three-dimensional reconstruction result obtained after processing 12 images and to avoid utilising the last three, strongly blurred, images. It is straightforward to take into account this criterion in the implementation of the online algorithm as the PSF radii estimated for all tracked features are available to the system.

4.1.3.3 Random Errors vs. Systematic Deviations

An important source of random errors of the combined structure from motion and depth from defocus method is the pixel noise of the CCD sensor, which influences the estimation of the PSF radius according to the depth–defocus function (3.17) and furthermore leads to a random scatter of the extracted feature positions of the order 0.1 pixels. According to Table 4.1, the reprojection error is always significantly larger than the random scatter of the KLT tracker and amounts to several tenths of a pixel. Presumably, systematic deviations are introduced by the changing appearance of tracked features across the sequence which cannot be fully described by affine deformation and thus cannot be fully compensated by the KLT tracker. The three-dimensional reconstruction results obtained with the offline algorithm for 100 runs over the cuboid, bottle, and lava stone sequences, respectively, show that the relative differences between the ground truth and the reconstructed absolute scale of the scene amount to a few percent and always correspond to between 1 and 2 standard deviations (cf. Table 4.1). The observed deviations are presumably due to a combination of random fluctuations and small systematic errors. Both types of error appear to be of the same order of magnitude.

Systematic errors may become important at the end of the sequence, where the images tend to be strongly blurred (Wöhler et al., 2009). For PSF radii smaller than 5 pixels we found that the random scatter of the feature positions extracted by the KLT tracker is of the order 0.1 pixels and independent of the PSF radius, such that the extracted feature positions do not introduce systematic errors. However, the ob-

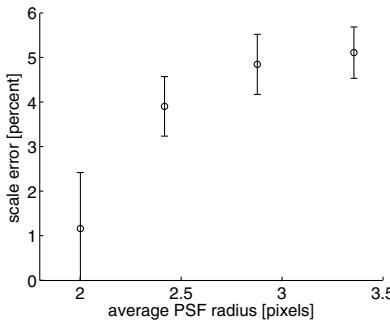


Fig. 4.12 Correlation between scale error and average PSF radius for the last four images of the lava stone sequence.

served relation $\sigma(z)$ between PSF radius and depth is accurately represented only for small and intermediate values of σ by the depth-defocus function $\mathcal{S}(z)$ according to Eq. (3.17). Fig. 3.12 illustrates that the 12 mm lens shows this effect for values of σ between 2 and 3 pixels on both sides of the minimum of $\mathcal{S}(z)$. Systematic errors might also be introduced by the nonlinearity of the depth-defocus function $\mathcal{S}(z)$. Effectively, the estimation of the depth z is based on an inversion of Eq. (3.17). Even if we assume that the measurement errors of $\sigma(z)$ for a certain depth z can be described by a Gaussian distribution of zero mean (which is a good approximation to the observed behaviour), the statistical properties of the inverse relation $z(\sigma)$ generally cannot be described in terms of a zero-mean Gaussian distribution. Due to the nonlinear nature of $\mathcal{S}(z)$, the average deviation between the measured depth value z and its value predicted by the depth-defocus function deviates from zero. For small PSF radii, i.e. close to the inflexion points of the depth-defocus function $\mathcal{S}(z)$, where its curvature is close to zero and its shape is largely linear, this effect is only minor, but its importance increases for large PSF radii, where $\mathcal{S}(z)$ may display a strong curvature as apparent in Fig. 3.12. For the lava stone sequence processed with the online algorithm, Fig. 4.12 illustrates the correlation between scale error and average PSF radius of the last processed image. The systematic effect is especially pronounced for this sequence since it comprises only 15 images (cf. Table 4.1). Measurement errors obtained while processing the last three images, which are strongly blurred with $\sigma > 2$ pixels, thus have a substantial effect on the three-dimensional reconstruction result. These findings suggest that features with large associated PSF radii should be excluded from the three-dimensional reconstruction process, where the range of favourable PSF radii depends on the depth-defocus function $\mathcal{S}(z)$.

A further important source of systematic errors is the thermal expansion of the optical system (Wöhler et al., 2009). The body of the lens used for our experiments consists of aluminium with a relative thermal expansion coefficient of $\nu = 2.3 \times 10^{-5} \text{ K}^{-1}$. With a lens of focal length f at calibration temperature T_0 an image of maximum sharpness is observed at depth z_0 . The corresponding principal

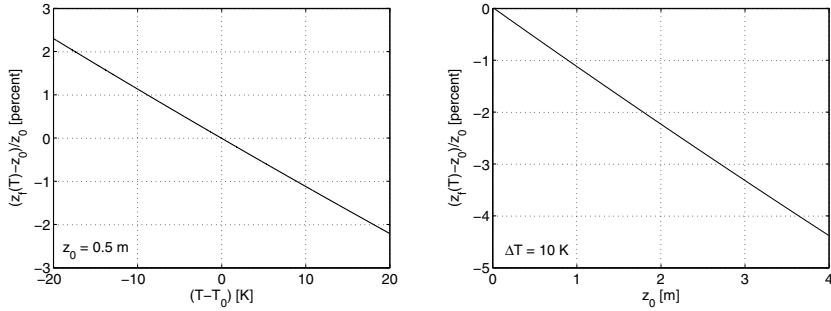


Fig. 4.13 Temperature-induced relative systematic error of the depth from defocus measurements, corresponding to $(z_f(T) - z_0)/z_0$. The focal length is set to $f = 20 \text{ mm}$.

distance v_0 is obtained according to the lens law (3.3). Assuming that image acquisition is performed at temperature T , the thermal expansion of the aluminium lens body yields a principal distance $b(T) = [1 + v(T - T_0)]b_0$, and the corresponding depth $z_f(T)$ for which an image of maximum sharpness is observed at temperature T is computed according to the lens law (3.3). Since the thermal expansion coefficient of the aluminium body is much larger than that of the glass components, we assume that the glass components remain unchanged across the regarded temperature range, such that the only effect of temperature variations is a changing value of the principal distance b . The focal length f is assumed to be constant. As a result, the depth–defocus function (3.17) is shifted by the amount $z_f(T) - z_0$ along the z axis, which introduces a relative systematic error of $[z_f(T) - z_0]/z_0$. We find that for a given temperature difference $|T - T_0| \ll T_0$, the relative systematic deviation of the depth from defocus measurement is largely proportional to z_0 , and for a given value of $z_0 \gg f$, it is largely proportional to $(T - T_0)$ (cf. Fig. 4.13). For a lens with $f = 20 \text{ mm}$, a temperature difference of about 10 K is thus able to explain the systematic deviations of the inferred absolute scale from the ground truth observed in the experiments, which are of the order of a few percent.

Further possible sources of systematic deviations are vibrations and shocks occurring after calibration (which were avoided during the experiments) and systematic variations of the appearance of the extracted ROIs across the image sequence, especially when the assumption of affine deformation does not hold or for specular surfaces. We show in Section 4.4 how to use the effects of specular reflection in a favourable way to improve the three-dimensional reconstruction results.

4.1.4 Discussion

In this section we have described a method for combining geometric and real-aperture methods for monocular three-dimensional reconstruction of static scenes

at absolute scale. The proposed algorithm is based on a sequence of images of the object acquired by a monocular camera of fixed focal setting from different viewpoints. Feature points are tracked over a range of distances from the camera, resulting in a varying degree of defocus for each tracked feature point. The determination of absolute depth based on the depth from defocus approach is described in Section 3.2.3. The inferred PSF radii for the corresponding scene points are used to compute a regularisation term for an extended bundle adjustment algorithm that simultaneously optimises the reprojection error and the absolute depth error for all feature points tracked across the image sequence. The proposed method yields absolutely scaled three-dimensional coordinates of the object feature points without any prior knowledge about scene structure and camera motion. The proposed method has been implemented as an offline and as an online algorithm.

Based on tabletop experiments with real-world objects, we have demonstrated that the offline version of the proposed algorithm yields absolutely scaled three-dimensional coordinates of the feature points with typical relative errors of a few percent. For the online algorithm, the reconstruction accuracy increases with increasing number of processed images as long as the images do not become strongly blurred. At the end of the sequence, the reconstruction results of the offline and the online versions of the proposed algorithm are of comparable accuracy. The deviations between the obtained reconstruction results and the ground truth can be explained by a combination of the random scatter of the extracted feature positions and the estimated PSF radii, which are both due to the noise of the pixel grey values, systematic errors introduced by a changing appearance of the tracked features over time when the assumption of affine deformation does not hold, and systematic deviations of the order 1 percent due to thermal expansion of the optical system. Further systematic errors may be introduced if the image sequence contains strongly blurred images with an average PSF radius larger than about two pixels, due to deviations of the observed depth dependence of the PSF radius from the analytic model used for the depth defocus function. Since the PSF radius is continuously computed in the course of the three-dimensional reconstruction process, it is possible and favourable to reject such strongly blurred images.

4.2 Self-consistent Combination of Shadow and Shading Features

This section describes a framework for three-dimensional surface reconstruction by self-consistent fusion of shading and shadow features introduced by Wöhler and Hafezi (2005). Based on the analysis of at least two pixel-synchronous images of the scene under different illumination conditions, this framework combines a shape from shading approach (cf. Section 2.2) for estimating surface gradients and depth variations on small scales with a shadow analysis method (cf. Section 2.1) that allows the determination of the large-scale properties of the surface (cf. Fig. 4.14). As a first step, the result of shadow analysis is used for selecting a consistent so-

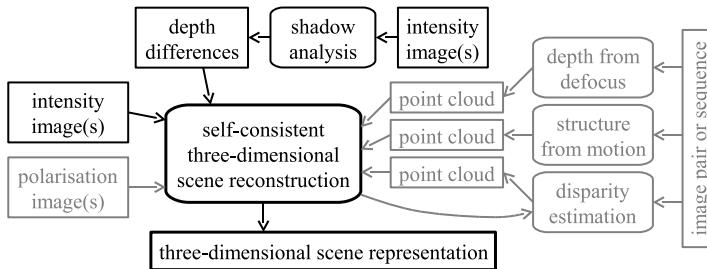


Fig. 4.14 Overview of the integrated approach to self-consistent combination of shadow and shading features.

lution of the shape from shading reconstruction algorithm. As a second step, an additional error term derived from the fine-structure of the shadow is incorporated into the reconstruction algorithm. This approach is extended to the analysis of two or more shadows under different illumination conditions leading to an appropriate initialisation of the shape from shading algorithm.

4.2.1 Selection of a Shape from Shading Solution Based on Shadow Analysis

This section introduces a self-consistent fusion scheme that combines the shape from shading approach based on a single light source (cf. Section 2.2) with the result of shadow analysis. Two images of the scene are registered at subpixel accuracy such that a pixel at position (u, v) corresponds to the same surface point in both images (Gottesfeld Brown, 1992). One of the images, the so-called shading image, is essentially shadow-free, while the other one shows one or several shadow areas and is therefore termed shadow image. The shadow pixels are segmented either by binarisation of the shadow image itself with a suitable threshold or by binarisation of the ratio image. The latter approach is insensitive to albedo variations on the surface and is therefore used throughout this section and in the applications regarded in Chapters 5 and 7.

According to Fig. 4.15a, the shadow is regarded as being composed of a number of S shadow lines. Shadow line s begins at pixel position $(u_i^{(s)}, v^{(s)})$ and ends at position $(u_e^{(s)}, v^{(s)})$ with $u_i^{(s)} \leq u_e^{(s)}$. The information about depth differences derived from the shadow is quite accurate while it is available only for a small fraction of pixels. The depth information derived from shape from shading is dense, but it strongly depends on the reflectance function, which is not necessarily exactly known, and may be affected by imperfections of the CCD sensor used for image acquisition e.g. due to saturation effects, by scattering or by specular reflection of incoming light at the surface. Furthermore, even if the slopes p_{uv} and q_{uv} are reason-

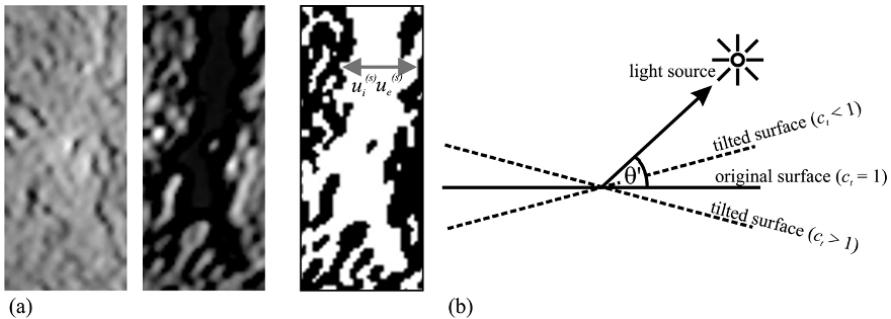


Fig. 4.15 (a) Definition of a shadow line. Left: Image used for shape from shading analysis. Middle: Image used for shadow analysis. Right: Binarised ratio image (shadow shown in white), the arrow indicating shadow line s ranging from $u_i^{(s)}$ to $u_e^{(s)}$. (b) Illustration of the adjustment factor c_t defined in Eq. (4.7).

ably well known, small but correlated errors of p_{uv} and q_{uv} will cumulate and finally result in large errors of z_{uv} over large scales. Therefore, depth differences estimated based on shape from shading are usually significantly more reliable on small scales than on large scales.

The algorithm presented in this section copes with any reflectance function $R(p, q)$ that is varying in a smooth and monotonical manner with $\cos \theta_i$. For the sake of simplicity we will exemplify the reconstruction method using a Lambertian reflectance function, keeping in mind that the proposed framework is suitable for arbitrary reflectance functions with physically reasonable properties (cf. Chapter 7), in order to illustrate how to incorporate a shadow-related error term into the shape from shading formalism. For the application scenario of three-dimensional reconstruction of small sections of the lunar surface as described in Chapter 7, the Lambertian reflectance values may be inaccurate by a few percent, but this error is usually smaller than the errors caused by nonlinearities and other imperfections of the CCD sensor. Anyway, the shadow-related error term permits significant deviations of the reconstruction result from the solution obtained from reflectance only.

For reflectance functions accurately modelling the photometric surface properties of planetary bodies one should refer to the detailed descriptions by Hapke (1993) or by McEwen (1991) (cf. Section 7.1.2). For the metallic surfaces of the industrial parts examined in Section 5.3, we performed measurements which indicate that for moderate surface gradients, viewing directions roughly perpendicular to the surface, and oblique illumination the assumption of a Lambertian reflectance function is a good approximation. For shallow slopes ($p, q \ll 1$) and oblique illumination many realistic reflectance functions can be approximated over a wide range in terms of a linear expansion in p and q at a good accuracy. Sophisticated shape from shading algorithms have been developed for this important class of reflectance functions by Horn (1989), which should all be applicable within the presented framework for surfaces with shallow slopes, perpendicular view, and oblique illumination.

With a Lambertian reflectance function, the result of the iterative shape from shading scheme according to Eq. (2.25) depends on the initial values $p_{uv}^{(0)}$ and $q_{uv}^{(0)}$ of the surface gradients and on the value chosen for the surface albedo ρ . A different initialisation usually yields a different result of the algorithm, as although the regularisation (e.g. smoothness or integrability) constraint strongly reduces the range of possible solutions, it still allows for an infinite number of them. Without loss of generality, it is assumed that the scene is illuminated exactly from the right or the left hand side.

To obtain a solution for z_{uv} which both minimizes the error term (2.22) and is at the same time consistent with the average depth difference derived from shadow analysis, we propose an iterative procedure:

1. As the regarded applications deal with relatively smooth and flat surfaces, the initial values $p_{uv}^{(0)}$ and $q_{uv}^{(0)}$ in Eq. (2.25) are set to zero. The iteration index m is set to $m = 0$. The surface gradients are then computed according to Eq. (2.25). Subsequently the initial surface profile $z_{uv}^{(m)}$ is reconstructed by numerical integration of the obtained values of p_{uv} and q_{uv} according to Jiang and Bunke (1997) as described in Section 2.2.3.
2. The average depth difference $(\Delta z)_{\text{shadow}}^{\text{ave}}$ based on shadow analysis is given by

$$(\Delta z)_{\text{shadow}}^{\text{ave}} = \frac{1}{S} \sum_{s=1}^S \left(|u_e^{(s)} - u_i^{(s)}| + 1 \right) \tan \mu_{\text{shadow}}. \quad (4.5)$$

In Eq. (4.5) the effective shadow length is given by $(|u_e^{(s)} - u_i^{(s)}| + 1)$, such that the shadow length is 1 if $u_i^{(s)} = u_e^{(s)}$, i.e. a single pixel along the direction of incident light lies in the shadow. The corresponding depth difference $(\Delta z)_{\text{sfs}}^{\text{ave}}$ given by shape from shading analysis is obtained by

$$(\Delta z)_{\text{sfs}}^{\text{ave}} = \frac{1}{S} \sum_{s=1}^S \left[z_m \left(u_e^{(s)}, v^{(s)} \right) - z_m \left(u_i^{(s)}, v^{(s)} \right) \right]. \quad (4.6)$$

At the pixels marking the ridge that casts the shadow, denoted by $(u_i^{(s)}, v^{(s)})$ or $(u_e^{(s)}, v^{(s)})$ depending on the direction of illumination, the surface gradient p in horizontal image direction corresponds to $\tan \mu_{\text{shadow}}$. These values are kept constant throughout the following steps.

3. Assuming that the scene is illuminated exactly in horizontal image direction, shape from shading analysis cannot yield reliable information about the surface gradient q_{uv} in the vertical image direction. Especially for small illumination angles μ , changing q_{uv} for a certain surface element does not significantly change the angle θ_i between the corresponding surface normal \mathbf{n} and the direction of illumination \mathbf{s} , which results in a small value of $\partial R / \partial q$ in Eq. (2.25) for Lambertian reflection. Once the initial values of q_{uv} are small, they remain small during the iteration process according to Eq. (2.25). The angle θ_i is mainly governed by the surface gradient p_{uv} in horizontal image direction. Hence, for all surface

elements the angle $\theta'_i = \pi/2 - \theta_i$ between the respective surface element (not its surface normal) and the illumination direction \mathbf{s} is multiplied with a constant factor c_t such that $(\Delta z)_{\text{shadow}}^{\text{ave}} = (\Delta z)_{\text{sfs}}^{\text{ave}}$. For small values of q_{uv} , the horizontal surface gradient p_{uv} is replaced by the new value \tilde{p}_{uv} according to

$$\tilde{p}_{uv} = -\cot\left(\mu + \frac{\pi}{2} - c_t \theta'_i\right) \quad \text{and} \quad \theta'_i = \mu \pm \frac{\pi}{2} + \arctan \frac{1}{p_{uv}}. \quad (4.7)$$

In Eq. (4.7) the plus sign is used for $p_{uv} < 0$ and the minus sign for $p_{uv} > 0$. The surface is tilted away from the light source if $c_t < 1$ and to the light source if $c_t > 1$ as illustrated in Fig. 4.15b. The value of c_t necessary to adjust the value of $(\Delta z)_{\text{sfs}}^{\text{ave}}$ to that of $(\Delta z)_{\text{shadow}}^{\text{ave}}$ is determined by means of the bisection method. This procedure has the strong advantage that for small illumination angles μ , small surface gradients p_{uv} and thus small angles θ'_i between the surface and the illumination direction \mathbf{s} , changing p_{uv} to \tilde{p}_{uv} according to (4.7) hardly changes the resulting relative pixel intensities. The corresponding Lambertian reflectance is given by $\tilde{\rho} \sin(c_t \theta'_i) \approx \rho \sin \theta'_i$ with $\tilde{\rho} = \rho/c_t$, resulting in a new value $\tilde{\rho}$ for the surface albedo but almost unchanged relative intensities throughout the image. The new gradient \tilde{p}_{uv} along with the new albedo $\tilde{\rho}$ is still a near-optimum configuration if $\theta_i \ll \pi/2$, which remains true also if $R(p, q)$ contains higher-order terms in $\cos \theta_i$. Hence, the iterative update rule (2.25) is expected to converge quickly.

4. The iterative update rule (2.25) is initialised with $p_{uv}^{(0)} = \tilde{p}_{uv}$, and the iteration procedure is started. After execution, the iteration index m is incremented ($m \leftarrow m + 1$) and the surface profile $z_{uv}^{(m)}$ is computed by numerical integration of p_{uv} and q_{uv} .
5. Steps 2, 3, and 4 are repeated until the average change of the surface profile falls below a user-defined threshold Θ_z , i.e. until $\left\langle \left(z_{uv}^{(m)} - z_{uv}^{(m-1)} \right)^2 \right\rangle_{u,v}^{1/2} < \Theta_z$. In all described experiments a value of $\Theta_z = 0.01$ pixels is used.

As long as the lateral pixel resolution in metric units is undefined, the depth profile z_{uv} is computed in pixel units. However, multiplying these values by the lateral pixel resolution on the reconstructed surface (e.g. measured in metres per pixel) readily yields the depth profile in metric units.

4.2.2 Accounting for the Detailed Shadow Structure in the Shape from Shading Formalism

This section describes how the detailed shadow structure rather than the average depth difference derived from shadow analysis is incorporated into the shape from shading formalism. The depth difference $(\Delta z)_{\text{shadow}}^{(s)}$ along shadow line s amounts to

$$(\Delta z)_{\text{shadow}}^{(s)} = \left(|u_e^{(s)} - u_i^{(s)}| + 1 \right) \tan \mu_{\text{shadow}}. \quad (4.8)$$

In the reconstructed profile, this depth difference is desired to match the depth difference

$$(\Delta z)_{\text{sfs}}^{(s)} = z(u_e^{(s)}, v^{(s)}) - z(u_i^{(s)}, v^{(s)}) \quad (4.9)$$

obtained by the shape from shading algorithm. Therefore it is useful to add a shadow-related error term to Eq. (2.22), leading to

$$e = e_s + \lambda e_i + \eta e_z \quad \text{with} \quad e_z = \sum_{s=1}^S \left[\frac{(\Delta z)_{\text{sfs}}^{(s)} - (\Delta z)_{\text{shadow}}^{(s)}}{|u_e^{(s)} - u_i^{(s)}| + 1} \right]^2. \quad (4.10)$$

The scene is supposed to be illuminated from either the right or the left hand side. The depth difference $(\Delta z)_{\text{sfs}}^{(s)}$ can be derived from the surface gradients by means of a discrete approximation of the total differential $dz = \frac{\partial z}{\partial x} dx + \frac{\partial z}{\partial y} dy$. Hence, as depth differences are evaluated along image rows only, the second term becomes zero, and we obtain

$$(\Delta z)_{\text{sfs}}^{(s)} = \sum_{u=u_i^{(s)}}^{u_e^{(s)}} p(u, v^{(s)}). \quad (4.11)$$

The derivative of e_z for pixel (u, v) with respect to the surface gradients p and q is then

$$\begin{aligned} \frac{\partial e_z}{\partial p} \Big|_{u,v} &= \begin{cases} 2 \frac{(\Delta z)_{\text{sfs}}^{(s)} - (\Delta z)_{\text{shadow}}^{(s)}}{(|u_e^{(s)} - u_i^{(s)}| + 1)^2} & \text{if } u_i^{(s)} \leq u \leq u_e^{(s)} \text{ and } v = v^{(s)} \\ 0 & \text{otherwise} \end{cases} \\ \frac{\partial e_z}{\partial q} \Big|_{u,v} &= 0. \end{aligned} \quad (4.12)$$

This leads to an extended update rule for the surface gradient p (cf. Eq. (2.25)):

$$p_{uv}^{(n+1)} = \bar{p}_{uv}^{(n)} + \lambda (I - R(\bar{p}_{uv}^{(n)}, \bar{q}_{uv}^{(n)})) \frac{\partial R}{\partial p} \Big|_{\bar{p}_{uv}^{(n)}, \bar{q}_{uv}^{(n)}} - \eta \frac{\partial e_z}{\partial p} \Big|_{u,v}. \quad (4.13)$$

The update rule for q in Eq. (2.25) remains unchanged. The iteration is initialised with the result of the algorithm described in Section 4.2.1. After each iteration step the depth profile z_{uv} needs to be reconstructed by numerical integration based on the current values of $p_{uv}^{(n)}$ and $q_{uv}^{(n)}$ (cf. Section 2.2.3) in order to determine the values of $(\Delta z)_{\text{sfs}}^{(s)}$ for the next iteration step.

A further condition is that if pixel (u, v) is outside the shadow, the angle between surface normal \mathbf{n} and vector \mathbf{s} of incident light must be less than $\pi/2$, and the values p_{uv} of the pixels outside the shadow have to be limited accordingly during the iteration process. This means that p_{uv} and q_{uv} must fulfill the condition $\mathbf{n} \cdot \mathbf{s}_{\text{shadow}} > 0$, which for incident light from the left or right hand side (zero component of $\mathbf{s}_{\text{shadow}}$

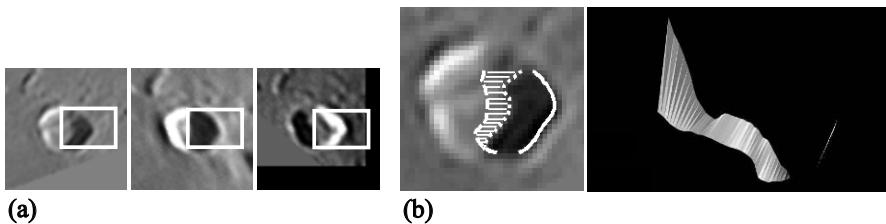


Fig. 4.16 Initialisation of the shape from shading algorithm by shadow analysis according to Section 4.2.3. (a) The images to the left in the middle are evaluated with respect to shadow, the image to the right with respect to shading. The reconstructed surface part is marked by a white rectangle, respectively. (b) Surface patch between the two shadow lines (hatched) along with the initial surface profile derived from shadow analysis.

in vertical image direction) becomes $p_{uv} < \tan \mu_{\text{shadow}}$ if $\mu_{\text{shadow}} \in [0, \dots, \pi/2[$ and $p_{uv} > \tan \mu_{\text{shadow}}$ if $\mu_{\text{shadow}} \in]\pi/2, \dots, \pi]$.

It is important to note that the shadow-related error term can be incorporated into any error function applied to a variational optimisation scheme, an overview of which is given in Section 2.2. Especially, the iterative update rule (2.31) based on the integrability error term (2.26) is readily extended to take into account depth differences indicated by shadow analysis in a manner analogous to Eq. (4.13). Three-dimensional reconstruction results of lunar tectonic faults and a wrinkle ridge obtained with this approach are presented in Section 7.3.

4.2.3 *Initialisation of the Shape from Shading Algorithm Based on Shadow Analysis*

In this section, the integration of shading and shadow features described in Section 4.2.1 is modified in that two or more shadows are used to initialize the shape from shading algorithm and compute a reliable value for the surface albedo ρ . The basic idea is developed based on the example of lunar surface reconstruction with a single light source used for shape from shading analysis, but the proposed method can also be used with shape from shading algorithms that rely on multiple light sources. Shape from shading analysis is performed by making use of error function (4.10) described in Section 4.2.2, which takes into account the fine structure of the shadow along with the shading information. The approach described in this section is applicable to arbitrary reflectance functions $R(p, q)$, but for the sake of simplicity we again assume Lambertian reflectance.

Fig. 4.16 shows a lunar crater (Theaetetus) at sunrise (illumination from the right) at solar elevation angles $\mu_{\text{shadow}}^{(1)} = 17.9^\circ$ and $\mu_{\text{shadow}}^{(2)} = 12.2^\circ$ and at sunset (illumination from the left) at a solar elevation angle of $\mu = 165.0^\circ$. The third image is used as the shading image, the eastern (right) half of the crater is reconstructed as

Table 4.2 Comparison of reconstruction results to ground truth data for the synthetic data examples given in Fig. 4.17.

example	μ_1	μ_2	$\mu_{\text{shadow}}^{(1)}$	$\mu_{\text{shadow}}^{(2)}$	z	RMSE (pixels)	ρ	RMSE (percent)
(a)	4°		2.5°		0.043		2.0	
(b)	6°	8°	2.5°		0.008		0.2	
(c)	6°	8°	2.5°		0.052		5.3	
(d)	5°	4°	5°		0.135		1.3	
					0.174		1.1	

indicated by the rectangular boxes in Fig. 4.16a. To achieve this, we again follow an iterative approach:

- Initially, it is assumed that the depths of the ridges casting the shadows are constant and identical. The iteration index m is set to $m = 0$. Based on the depth differences with respect to the ridges, the three-dimensional profile $\tilde{z}_m(u, v)$ of the small surface patch between the two shadow lines can be derived from the shadow lengths measured according to Section 4.2.1 (cf. Fig. 4.16b).
- The surface profile $\tilde{z}_m(u, v)$ directly yields the slopes $p_{uv}^{(0)}$ and $q_{uv}^{(0)}$ in horizontal and vertical image direction, respectively, for all pixels belonging to the surface patch between the shadow lines. The known values of $p_{uv}^{(0)}$ and $q_{uv}^{(0)}$ are used to compute the albedo ρ , and they serve as initial values inside the surface patch between the two shadow lines for the shape from shading algorithm. These values are kept constant throughout the following steps of the algorithm. Outside the region between the shadow lines, the initial values of $p_{uv}^{(0)}$ and $q_{uv}^{(0)}$ are set to zero.
- Using the shape from shading algorithm with the initialisation applied in the previous step, the complete surface profile $z_{uv}^{(m)}$ is reconstructed based on the shading image. In general, the resulting depths of the ridges casting the shadows are not identical any more—they are extracted from the reconstructed surface profile $z_{uv}^{(m)}$. This yields a new profile $\tilde{z}_{m+1}(u, v)$ for the surface patch between the shadow lines.
- The iteration index m is incremented: $m \leftarrow m + 1$.
- Steps 2, 3, and 4 are repeated until the criterion $\left\langle \left(z_{uv}^{(m)} - z_{uv}^{(m-1)} \right)^2 \right\rangle_{u,v}^{1/2} < \Theta_z$ is fulfilled. Again, the threshold value $\Theta_z = 0.01$ pixels is used.

This iterative algorithm mutually adjusts in a self-consistent manner the depth profiles of the floor and the ridges that cast the shadows. It allows to determine not only the surface gradients p_{uv} in the direction of the incident light, as it can be achieved by shape from shading alone, but to estimate the surface gradients q_{uv} in the perpendicular direction as well. It is especially suited for surface reconstruction under coplanar light sources. Furthermore, the algorithm does not require any special form of the shape from shading algorithm or the reflectance function used, such that it can be extended in a straightforward manner to more than two shadows and to shape from shading algorithms based on multiple light sources.

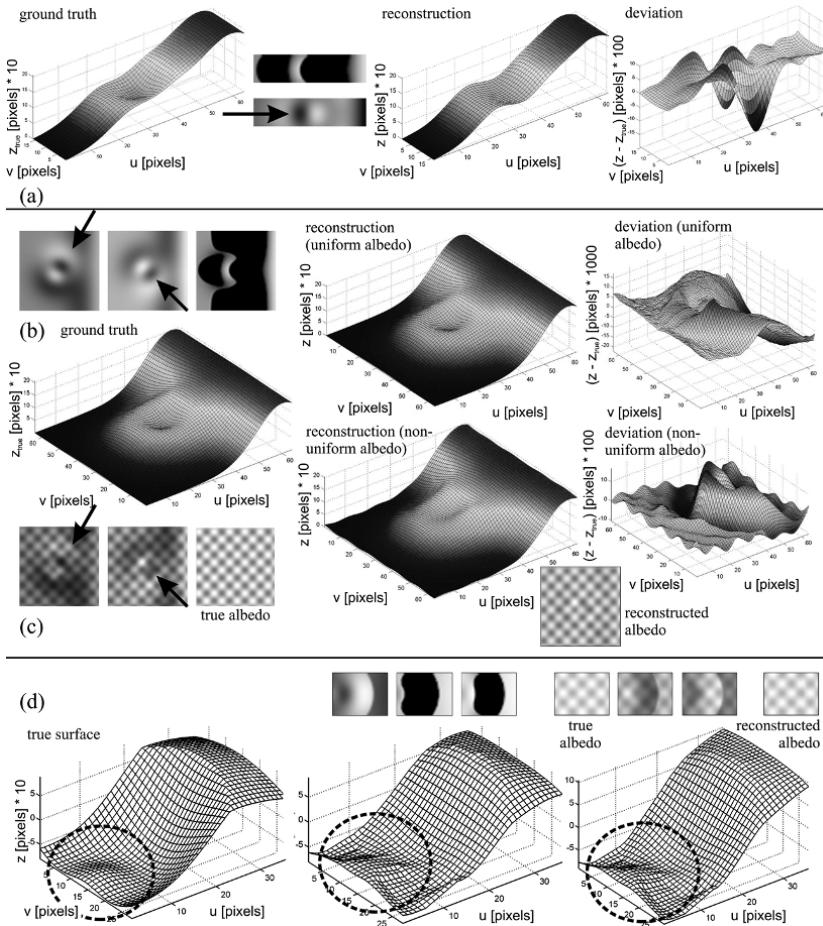


Fig. 4.17 Reconstruction results for synthetic data. (a) Surface reconstruction by a combined analysis of one shading and one shadow image according to Section 4.2.1, with uniform surface albedo. (b) Surface reconstruction by photometric stereo (two shading images) according to Section 2.3.1 and one shadow image, with uniform surface albedo. (c) Same as (b), but with non-uniform surface albedo. (d) Surface reconstruction by shadow-based initialisation of the surface profile according to Section 4.2.3, with uniform (middle) and non-uniform (right) surface albedo. The label of the z axis of the left plot is valid for all three plots. For detailed comments, cf. Section 4.2.4 and Table 4.2.

4.2.4 Experimental Evaluation Based on Synthetic Data

To demonstrate the accuracy of the surface reconstruction techniques presented in the previous section, they are applied to synthetic data for which the ground truth is known. A Lambertian reflectance function is assumed. In all examples of Fig. 4.17 we made use of the smooth surface constraint (2.21). In Fig. 4.17a–c, the true surface

profile is shown to the left along with the synthetically generated images used for reconstruction, the reconstructed profile in the middle, and the deviation ($z - z_{\text{true}}$) between reconstructed and true depth to the right, respectively. The directions of illumination for the shading images are indicated by arrows. Fig. 4.17a illustrates the result of the method described in Section 4.2.2 based on one shading and one shadow image. Consequently, the surface gradients in vertical image direction are slightly under-estimated. In Fig. 4.17b, two shading images and one shadow image are employed, and a uniform surface albedo is used. This yields a very accurate reconstruction result. The same illumination setting is assumed in Fig. 4.17c, but the albedo is strongly non-uniform. Here, the algorithm described in Section 2.3.2 based on the ratio of the images is employed, yielding a less accurate but still reasonable reconstruction result. The computed albedo map is shown next to the reconstructed surface profile. Fig. 4.17d illustrates the performance of the algorithm proposed in Section 4.2.3 on a synthetically generated object (left). In contrast to traditional shape from shading, the surface gradients perpendicular to the direction of incident light are revealed (middle). The single-image error term (2.20) was then replaced by the ratio-based error term (2.50) for a reconstruction of the same synthetic object but now with a non-uniform albedo (right). Consequently, two shading images are used in combination with the shadow information. As a result, a similar surface profile is obtained along with the surface albedo. Refer to Table 4.2 for a detailed comparison between ground truth and reconstruction results in terms of the root mean square error (RMSE).

4.2.5 Discussion

In this section we have described a self-consistent scheme for the integration of shading and shadow features based on at least two pixel-synchronous images of the same surface part under different illumination conditions. For a small number of surface points, the shadow analysis yields accurate depth differences, while dense but less accurate depth information is obtained with the shape from shading approach, which, however, permits an infinite number of solutions even with the regularisation constraint of a smooth or integrable surface. As a first step, a solution which is consistent with the average depth difference determined by shadow analysis is selected by means of the presented iterative scheme. As a second step, the error function to be minimised is extended by an error term that takes into account the detailed structure of the shadow.

The second described approach to combine shading and shadow features is especially suitable for surface reconstruction under coplanar light sources. It is based on the initialisation of the surface profile by analysis of at least two shadows observed at different illumination angles. In this setting, shadow analysis allows for deriving the surface gradients in both the horizontal and the vertical image direction along with the surface albedo for a small surface patch between the shadow lines. The surface profile is iteratively reconstructed based on the first proposed algorithm, re-

lying on one shading and one shadow image, which is initialised with the previously obtained result of the analysis of two or more shadows.

Based on synthetic data, we have shown that a high accuracy can be achieved with the proposed reconstruction techniques. This first evaluation is extended towards the three-dimensional reconstruction of metallic surfaces in Chapter 5 and lunar surface regions in Chapter 7.

4.3 Shape from Photopolarimetric Reflectance and Depth

This section introduces an image-based three-dimensional surface reconstruction method based on simultaneous evaluation of intensity and polarisation features (shape from photopolarimetric reflectance) and its combination with absolute depth data (shape from photopolarimetric reflectance and depth). A number of approaches to combine stereo and shape from shading have been proposed in the literature. Cryer et al. (1995) fuse low-pass filtered stereo depth data and high-pass filtered shape from shading depth data. Their approach, however, requires dense depth data and fuses the independently obtained results of two separate algorithms. Samaras et al. (2000) introduce a surface reconstruction algorithm that performs stereo analysis of a scene and uses a minimum description length metric to selectively apply shape from shading to regions with weak texture. A surface model described by finite elements is adjusted to minimise a combined depth, multi-image shape from shading, and smoothness error. The influence of a depth point on the surface, however, is restricted to a small local neighbourhood of the corresponding finite element, favouring the use of dense depth data. A related approach by Fassold et al. (2004) integrates stereo depth measurements into a variational shape from shading algorithm and estimates surface shape, light source direction, and diffuse reflectance map. In their approach, the influence of a depth point is restricted to a small local neighbourhood of the corresponding image pixel. Horovitz and Kiryati (2004) propose a method that enforces sparse depth points during the surface gradient integration step performed in many shape from shading algorithms, involving a heuristically chosen parameterised weight function governing the local influence of a depth point on the reconstructed surface. They propose a second approach, suggesting a subtraction of the large-scale deviation between the depth results independently obtained by stereo and shape from shading, respectively, from the shape from shading solution. For sparse stereo data, the large-scale deviation is obtained by fitting a sufficiently smooth parameterised surface model to the depth difference values. Both approaches fuse independently obtained results of two separate algorithms.

The technique proposed in this section is based on the analysis of single or multiple intensity and polarisation images. To compute the surface gradients, we present a global optimisation method based on a variational framework and a local optimisation method based on solving a set of nonlinear equations individually for each image pixel. These approaches are suitable for strongly non-Lambertian surfaces and those of diffuse reflectance behaviour and can also be adapted to surfaces of

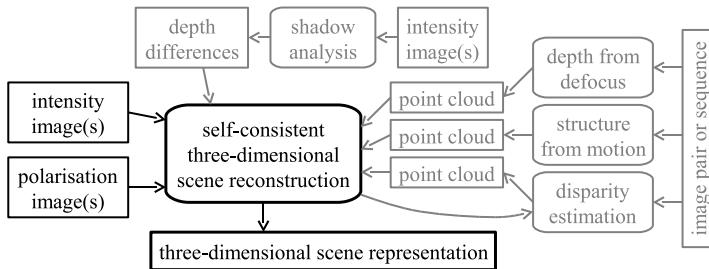


Fig. 4.18 Overview of the shape from photopolarimetric reflectance (SfPR) approach.

non-uniform albedo. We describe how independently measured absolute depth data are integrated into the shape from photopolarimetric reflectance framework in order to increase the accuracy of the three-dimensional reconstruction result. In this context we concentrate on dense but noisy depth data obtained by depth from defocus and on sparse but accurate depth data obtained by stereo or structure from motion analysis. We show that depth from defocus information should preferentially be used for initialising the optimisation schemes for the surface gradients. For integration of sparse depth information, we suggest an optimisation scheme that simultaneously adapts the surface gradients to the measured intensity and polarisation data and to the surface slopes implied by depth differences between pairs of depth points. In principle, arbitrary sources of depth information are possible in the presented framework. Experiments on synthetic and real-world data reveal that while depth from defocus is especially helpful for providing an initial estimate of the surface gradients and the albedo in the absence of a-priori knowledge, integration of stereo or structure from motion information significantly increases the three-dimensional reconstruction accuracy especially on large scales.

4.3.1 Shape from Photopolarimetric Reflectance

In our scenario we assume that the surface $z(x,y)$ to be reconstructed is illuminated by a point light source and viewed by a camera, both situated at infinite distance in the directions \mathbf{s} and \mathbf{v} , respectively (cf. Fig. 2.2). The xy plane is parallel to the image plane. Parallel unpolarised incident light and an orthographic projection model are assumed. For each pixel location (u, v) of the image we intend to derive a depth value z_{uv} . The surface normal \mathbf{n} , the illumination vector \mathbf{s} , the vector \mathbf{v} to the camera, the incidence and emission angles θ_i and θ_e , the phase angle α , and the surface albedo ρ_{uv} are defined according to Section 2.2.

In the framework of shape from photopolarimetric reflectance (SfPR) (d'Angelo and Wöhler, 2005a,b) as illustrated in Fig. 4.18, the light reflected from a surface point located at the world coordinates (x, y, z) with corresponding image coordinates (u, v) is described by the observed pixel intensity I_{uv} , the polarisation angle Φ_{uv} (i.e.

the direction in which the light is linearly polarised), and the polarisation degree D_{uv} . This representation is analogous to the one chosen in the context of shape from shading (cf. Section 2.2) and photometric stereo (cf. Section 2.3). The measurement of polarisation properties is thus limited to linear polarisation while circular or elliptic polarisation is not taken into account. It is assumed that models are available that express these photopolarimetric properties in terms of the surface orientation \mathbf{n} , illumination direction \mathbf{s} , and viewing direction \mathbf{v} . These models may either be physically motivated or empirical (cf. Sections 2.2.1 and 2.4.2) and are denoted here by R_I (intensity reflectance), R_Φ (polarisation angle reflectance), and R_D (polarisation degree reflectance). The aim of surface reconstruction in the presented framework is to determine for each pixel (u,v) the surface gradients p_{uv} and q_{uv} , given the illumination direction \mathbf{s} and the viewing direction \mathbf{v} , such that the modelled photopolarimetric properties of a pixel correspond to the measured values:

$$I_{uv} = R_I(p_{uv}, q_{uv}, \mathbf{s}, \mathbf{v}) \quad (4.14)$$

$$\Phi_{uv} = R_\Phi(p_{uv}, q_{uv}, \mathbf{s}, \mathbf{v}) \quad (4.15)$$

$$D_{uv} = R_D(p_{uv}, q_{uv}, \mathbf{s}, \mathbf{v}) \quad (4.16)$$

The reflectance functions (4.14)–(4.16) may depend on further, e. g. material-specific, parameters which possibly in turn depend on the pixel coordinates (u,v) , such as the surface albedo ρ_{uv} which influences the intensity reflectance R_I .

As long as a single light source is used, it is possible without loss of generality to define the surface normal in a coordinate system with positive x and zero y component of the illumination vector \mathbf{s} , corresponding to $p_s < 0$ and $q_s = 0$ with a surface normal $\mathbf{n} = (-p, -q, 1)^T$. Furthermore, for simplicity we always choose the z axis such that the viewing direction corresponds to $\mathbf{v} = (0, 0, 1)^T$. The surface normal $\tilde{\mathbf{n}} = (-\tilde{p}, -\tilde{q}, 1)^T$ in the world coordinate system, in which the azimuth angle of the light source is denoted by ψ , is related to \mathbf{n} by a rotation around the z axis, leading to

$$\tilde{p} = p \cos \psi - q \sin \psi \quad \text{and} \quad \tilde{q} = p \sin \psi + q \cos \psi. \quad (4.17)$$

It is generally favourable to define the reflectance functions R_I , R_Φ , and R_D in the coordinate system in which $q_s = 0$. If several light sources with different azimuth angles are used, it must then be kept in mind to take into account the transformation between the two coordinate systems according to Eq. (4.17).

In the following paragraphs we describe a global and a local approach to solve the problem of shape from photopolarimetric reflectance (SfPR), i.e. to adapt the surface gradients p_{uv} and q_{uv} to the observed photopolarimetric properties I_{uv} , Φ_{uv} , and D_{uv} by solving the (generally nonlinear) system of equations (4.14)–(4.16). The three-dimensional surface profile z_{uv} is then obtained by integration of the surface gradients according to the method proposed by Simchony et al. (1990) as described in Section 2.2.3.

4.3.1.1 Global Optimisation Scheme

The first solving technique is based on the optimisation of a global error function simultaneously involving all image pixels (d'Angelo and Wöhler, 2005a). This approach is described in detail in Section 2.2 (Horn and Brooks, 1989; Horn, 1989; Jiang and Bunke, 1997). One part of this error function is the intensity error term (2.20). As the pixel intensity information alone is not necessarily sufficient to provide an unambiguous solution for the surface gradients p_{uv} and q_{uv} , a regularisation constraint e_s is introduced which requires smoothness of the surface, i.e. for example small absolute values of the directional derivatives of the surface gradients. We therefore make use of the additional smoothness error term (2.21) (Horn and Brooks, 1989; Jiang and Bunke, 1997). In the scenarios regarded in this work, the assumption of a smooth surface is realistic. For wrinkled surfaces, where using Eq. (2.21) leads to an unsatisfactory result, it can be replaced by the departure from integrability error term (2.26) as discussed in detail in Section 2.2.3.

In our scenario, the incident light is unpolarised. For smooth metallic surfaces the light remains unpolarised after reflection at the surface. Rough metallic surfaces, however, partially polarise the reflected light, as shown e.g. by Wolff (1991). When observed through a linear polarisation filter, the reflected light has a transmitted radiance that oscillates sinusoidally as a function of the orientation of the polarisation filter between a maximum I_{\max} and a minimum I_{\min} . The polarisation angle $\Phi \in [0^\circ, 180^\circ]$ denotes the orientation under which maximum transmitted radiance I_{\max} is observed. The polarisation degree is defined by $D = (I_{\max} - I_{\min}) / (I_{\max} + I_{\min}) \in [0, 1]$ (cf. Section 2.4.2 for details). Like the reflectance of the surface, both polarisation angle and degree depend on the surface normal \mathbf{n} , the illumination direction \mathbf{s} , and the viewing direction \mathbf{v} . No sufficiently accurate physical model exists so far which is able to describe the polarisation behaviour of light scattered from a rough metallic surface. We therefore determine the functions $R_\Phi(\mathbf{n}, \mathbf{s}, \mathbf{v})$ and $R_D(\mathbf{n}, \mathbf{s}, \mathbf{v})$, describing the polarisation angle and degree of the material, respectively, for the phase angle α between the vectors \mathbf{s} and \mathbf{v} over a wide range of illumination and viewing configurations. To obtain analytically tractable relations rather than discrete measurements, phenomenological models are fitted to the obtained measurements (cf. Section 2.4.2).

To integrate the polarisation angle and degree data into the three-dimensional surface reconstruction framework, we define two error terms e_Φ and e_D which denote the deviations between the measured values and those computed using the corresponding phenomenological model, respectively:

$$e_\Phi = \sum_{l=1}^L \sum_{u,v} \left[\Phi_{uv}^{(l)} - R_\Phi \left(\theta_i^{(l)}(u, v), \theta_e(u, v), \alpha^{(l)} \right) \right]^2 \quad (4.18)$$

$$e_D = \sum_{l=1}^L \sum_{u,v} \left[D_{uv}^{(l)} - R_D \left(\theta_i^{(l)}(u, v), \theta_e(u, v), \alpha^{(l)} \right) \right]^2. \quad (4.19)$$

Based on the feature-specific error terms e_I , e_Φ , and e_D , a combined error term e is defined which takes into account both reflectance and polarisation properties:

$$e = e_s + \lambda e_I + \mu e_\Phi + \nu e_D. \quad (4.20)$$

Minimising error term (4.20) yields the surface gradients p_{uv} and q_{uv} that optimally correspond to the observed reflectance and polarisation properties, where the Lagrange parameters λ , μ , and ν denote the relative weights of the individual reflectance-specific and polarisation-specific error terms. With the discrete approximations $\left\{ \frac{\partial p}{\partial x} \right\}_{uv} = [p_{u+1,v} - p_{u-1,v}] / 2$ and $\left\{ \frac{\partial p}{\partial y} \right\}_{uv} = [p_{u,v+1} - p_{u,v-1}] / 2$ for the second derivatives of the surface z_{uv} and \bar{p}_{uv} as the local average over the four nearest neighbours of pixel (u, v) we obtain an iterative update rule for the surface gradients by setting the derivatives of the error term e with respect to p and q to zero, leading to

$$\begin{aligned} p_{uv}^{(n+1)} = \bar{p}_{uv}^{(n)} + \lambda \sum_{l=1}^L \left(I - R_I(\bar{p}_{uv}^{(n)}, \bar{q}_{uv}^{(n)}) \right) \frac{\partial R_I}{\partial p} + \mu \sum_{l=1}^L (\Phi - R_\Phi(\bar{p}_{uv}^{(n)}, \bar{q}_{uv}^{(n)})) \frac{\partial R_\Phi}{\partial p} + \\ \nu \sum_{l=1}^L (D - R_D(\bar{p}_{uv}^{(n)}, \bar{q}_{uv}^{(n)})) \frac{\partial R_D}{\partial p}, \end{aligned} \quad (4.21)$$

where n denotes the iteration index. A corresponding expression for q is obtained in an analogous manner (cf. also Section 2.2.2). The initial values $p_{uv}^{(0)}$ and $q_{uv}^{(0)}$ must be provided based on a-priori knowledge about the surface or on independently obtained depth data (cf. Section 4.3.3). The partial derivatives in Eq. (4.21) are evaluated at $(\bar{p}_{uv}^{(n)}, \bar{q}_{uv}^{(n)})$, respectively, making use of the phenomenological model fitted to the measured reflectance and polarisation data. The surface profile z_{uv} is then derived from the resulting gradients p_{uv} and q_{uv} by means of numerical integration of the gradient field according to the method suggested by Simchony et al. (1990).

Note that for the computation of the derivatives of R_I , R_Φ , and R_D with respect to the surface gradients p and q , as required to apply the iterative update rule (4.21), Eq. (4.17) has to be taken into account if the azimuth angles of the light sources are different from zero.

4.3.1.2 Local Optimisation Scheme

Provided that the model parameters of the reflectance and polarisation functions $R_I(p_{uv}, q_{uv})$, $R_\Phi(p_{uv}, q_{uv})$, and $R_D(p_{uv}, q_{uv})$ are known and measurements of intensity and polarisation properties are available for each image pixel, the surface gradients p_{uv} and q_{uv} can be obtained by solving the nonlinear system of equations (4.14)–(4.16) individually for each pixel (d'Angelo and Wöhler, 2005b). For this purpose we make use of the Levenberg-Marquardt algorithm (Press et al., 1992). In the overdetermined case, the root of Eqs. (4.14)–(4.16) is computed in the least-squares sense. The contributions from the different terms are then weighted ac-

cording to the corresponding measurement errors. In the application scenario regarded in Section 5.3, these standard errors have been empirically determined to $\sigma_I \approx 10^{-3} I_{\text{spec}}$ with I_{spec} as the intensity of the specular reflections, $\sigma_\phi \approx 0.1^\circ$, and $\sigma_D \approx 0.02$. The surface profile z_{uv} is again derived from the resulting gradients p_{uv} and q_{uv} by means of numerical integration of the gradient field (Simchony et al., 1990).

It is straightforward to extend this approach to photopolarimetric stereo because each light source provides an additional set of equations. Eq. (4.14) can only be solved, however, when the surface albedo ρ_{uv} is known for each surface point. A constant albedo can be assumed in many applications. If this assumption is not valid, albedo variations strongly affect the accuracy of surface reconstruction. In Section 2.3.2 it is shown that as long as the surface albedo can be assumed to be of the form (2.47), it is then possible to utilise two images $I_{uv}^{(1)}$ and $I_{uv}^{(2)}$ acquired under different illumination conditions. Eq. (4.14) can then be replaced by the ratio-based relation (2.48) such that the albedo cancels out (McEwen, 1985; Wöhler and Hafezi, 2005; Lena et al., 2006).

An advantage of the described local approach is that the three-dimensional reconstruction result is not affected by additional constraints such as smoothness of the surface but directly yields the surface gradient vector for each image pixel. A drawback, however, is the fact that due to the inherent nonlinearity of the problem, existence and uniqueness of a solution for p_{uv} and q_{uv} are not guaranteed for both the albedo-dependent and the albedo-independent case. However, in the experiments presented in Section 4.3.4 and Chapter 5 we show that in practically relevant scenarios a reasonable solution for the surface gradient field and the resulting depth z_{uv} is obtained even in the presence of noise.

4.3.2 Estimation of the Surface Albedo

For the specular surfaces regarded for the experimental evaluations based on synthetic data (cf. Section 4.3.4) and on real-world objects in the context of industrial quality inspection (cf. Chapter 5), we utilise the three-component BRDF according to Eq. (2.15), corresponding to the reflectance function

$$R_I(\theta_i, \theta_e, \alpha) = \rho \left[\cos \theta_i + \sum_{n=1}^N \sigma_n \cdot (2 \cos \theta_i \cos \theta_e - \cos \alpha)^{m_n} \right] \quad (4.22)$$

with $N = 2$. The term in round brackets corresponds to $\cos \theta_r$ with θ_r as the angle between the direction \mathbf{v} from the surface to the camera and the direction of mirror-like reflection (cf. Eq. (2.12)). For $\theta_r > 90^\circ$ only the diffuse component is considered. For a typical rough metallic surface, the measured reflectance function is shown in Fig. 2.4, where the material-specific parameters according to Eq. (4.22) are given by $\sigma_1 = 3.85$, $m_1 = 2.61$, $\sigma_2 = 9.61$, and $m_2 = 15.8$. The specular lobe is described by σ_1 and m_1 and the specular spike by σ_2 and m_2 , respectively.

One possible way to determine a uniform surface albedo ρ is its estimation based on the specular reflections in the images used for three-dimensional reconstruction, which appear as regions of maximum intensity $I_{\text{spec}}^{(l)}$ as long as the reflectance behaviour is strongly specular, i.e. at least one of the parameters σ_n is much larger than 1. Note that the pixel grey values of these regions must not be oversaturated. For these surface points we have $\theta_r = 0$ and $\theta_i^{(l)} = \alpha^{(l)}/2$. Relying on the previously determined parameters σ_n , Eq. (4.22) yields

$$\rho = \frac{1}{L} \sum_{l=1}^L I_{\text{spec}}^{(l)} \cdot \left[\cos\left(\frac{\alpha^{(l)}}{2}\right) + \sum_{n=1}^N \sigma_n(\alpha^{(l)}) \right]^{-1}. \quad (4.23)$$

In principle, a single image is already sufficient to determine the value of ρ as long as specular reflections are present in it. Note that in Eq. (4.23) the dependence of the parameters of the reflectance function on the phase angle α is explicitly included.

An albedo estimation according to Eq. (4.23) is not possible when the maximum intensity in the image does not correspond to specular reflection with $\theta_r = 0$. In the global optimisation scheme the surface albedo ρ can then be estimated in each iteration step n simultaneously along with the surface gradients. This is achieved by solving Eq. (4.22) for ρ_{uv} individually for each pixel (u, v) based on the values of $p_{uv}^{(n)}$ and $q_{uv}^{(n)}$ according to

$$\rho_{uv}^{(n)} = \frac{I_{uv}}{\tilde{R}\left(p_{uv}^{(n)}, q_{uv}^{(n)}\right)} \quad (4.24)$$

with $\tilde{R}(p, q)$ defined according to Eq. (2.47). The uniform albedo ρ is then obtained by computing an appropriate average of the computed values of $\rho_{uv}^{(n)}$. For the strongly specular surfaces regarded in our experiments, we found that the median of $\rho_{uv}^{(n)}$ provides a more robust estimate of ρ than the mean, since already a small number of pixels with inaccurately estimated surface gradients (which occur especially at the beginning of the iteration procedure) leads to a significant shift of the mean value while leaving the median value largely unaffected. If no a-priori information about the surface gradients is available, the initial guess of ρ , which in turn depends on the initial guess $p_{uv}^{(0)}$ and $q_{uv}^{(0)}$ of the surface gradients, has a strong influence on the solution found by the global optimisation scheme. Such a-priori information can be obtained based on independently measured depth data as described in the next section.

4.3.3 Integration of Depth Information

Since the obtained solution of shape from shading and, to a lesser extent, SfPR may be ambiguous as long as single images are regarded, integrating additional information into the surface reconstruction process is supposed to improve the reconstruction result. For example, a sparse three-dimensional point cloud of the object surface

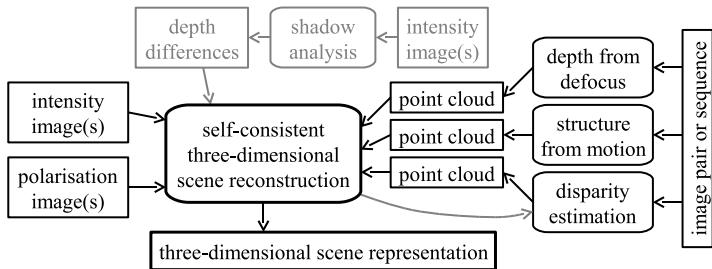


Fig. 4.19 Overview of the shape from photopolarimetric reflectance and depth (SfPRD) approach.

can be reconstructed by stereo vision, laser triangulation, or shadow analysis. Previous approaches either merge the results of stereo and shape from shading (Cryer et al., 1995) or embed the shape from shading algorithm into stereo (Samaras et al., 2000) or structure from motion (Lim et al., 2005) algorithms.

This section describes how independently acquired additional depth information can be integrated into the SfPR framework outlined in Section 4.3.1. We found that the fusion between SfPR and depth-measuring algorithms is especially useful if the depth data are dense but display a considerable amount of noise, or if they are accurate but only available for a sparse set of surface points. Hence, we concentrate on dense but noisy depth information, examining as an example the monocular depth from defocus technique, and on reasonably accurate but sparse depth information, based on the examples of stereo image analysis and structure from motion (cf. Fig. 4.19).

4.3.3.1 Fusion of SfPR with Depth from Defocus

To obtain a dense depth map of the surface, we employ the two-image depth from defocus method described in Section 3.2.2. Once the characteristic curve $\sigma(z - z_0)$ which relates the PSF radius σ to the depth offset $(z - z_0)$ is known (cf. Fig. 3.5), it is possible to extract a dense depth map from a pixel-synchronous pair of images of a surface of unknown shape, provided that the images are acquired at the same focus setting and with the same apertures as the calibration images. The resulting depth map z_{uv}^{DfD} , however, tends to be very noisy as illustrated in Fig. 3.7 (variables representing results obtained by depth from defocus are marked by the index “DfD”). It is therefore favourable to fit a plane $\tilde{z}_{uv}^{\text{DfD}}$ to the computed depth points, since higher-order information about the surface is usually not contained in the noisy depth from defocus data. This procedure reveals information about the large-scale properties of the surface (d’Angelo and Wöhler, 2005c). Approximate surface gradients can then be obtained by computing the partial derivatives $p_{uv}^{\text{DfD}} = \partial \tilde{z}_{uv}^{\text{DfD}} / \partial x$ and $q_{uv}^{\text{DfD}} = \partial \tilde{z}_{uv}^{\text{DfD}} / \partial y$.

In many cases there exists no unique solution for the surface gradients p_{uv} and q_{uv} within the SfPR framework, especially for highly specular reflectance functions.

This applies both to the global (Section 4.3.1.1) and to the local (Section 4.3.1.2) optimisation scheme. Therefore, the obtained solution tends to depend strongly on the initial values $p_{uv}^{(0)}$ and $q_{uv}^{(0)}$. As we assume that no a-priori information about the surface is available, we initialise the optimisation scheme with $p_{uv}^{(0)} = p_{uv}^{\text{DfD}}$ and $q_{uv}^{(0)} = q_{uv}^{\text{DfD}}$, thus making use of the large-scale surface gradients obtained by depth from defocus analysis. The ambiguity of the solution of the global optimisation scheme is even more pronounced when no a-priori knowledge about both the surface gradients and the albedo is available. In such cases, which are often encountered in practically relevant scenarios, an initial albedo value is computed according to Eq. (4.24) based on the initial surface gradients p_{uv}^{DfD} and q_{uv}^{DfD} . We found experimentally that it is advantageous to keep this albedo value constant during the iteration process as long as no additional constraints can be imposed on the surface, since treating the albedo as a further free parameter in the iteration process increases the manifold of local minima of the error function.

The depth from defocus data are derived from two images acquired with large and small aperture, respectively. In practice, it is desirable but often unfeasible to use the well-focused image acquired with small aperture for three-dimensional reconstruction—the image brightness then tends to become too low for obtaining reasonably accurate polarisation data. The surface reconstruction algorithm thus may have to take into account the position-dependent PSF. We incorporate the depth from defocus information into the global optimisation scheme since it is not possible to introduce PSF information (which applies to a local neighbourhood of a pixel) into an approach based on the separate evaluation of each individual pixel. The error terms (2.20), (4.18), and (4.19) of the SfPR scheme described in Section 4.3.1 are modified according to

$$e_I^{\text{PSF}} = \sum_{u,v} [I_{uv} - G_{uv} * R_I(p_{uv}, p_{uv}, q_{uv}, \alpha)]^2 \quad (4.25)$$

$$e_\Phi^{\text{PSF}} = \sum_{u,v} [\Phi_{uv} - G_{uv} * R_\Phi(p_{uv}, q_{uv}, \alpha)]^2 \quad (4.26)$$

$$e_D^{\text{PSF}} = \sum_{u,v} [D_{uv} - G_{uv} * R_D(p_{uv}, q_{uv}, \alpha)]^2 \quad (4.27)$$

describing the mean square deviation between the observed intensity and polarisation values and the modelled reflectances convolved with the PSF G_{uv} extracted from the image as described in Section 3.2.2. This approach is related to the shape from shading scheme for blurred images introduced by Joshi and Chaudhuri (2004). In that work, however, the PSF radius is estimated simultaneously with the surface gradients, while we independently determine the PSF radius for every position in the image during the depth from defocus analysis. The iterative update rule (4.21) then becomes

$$p^{(n+1)} = \bar{p}^{(n)} + \lambda (I - G * R_I) G * \frac{\partial R_I}{\partial p} + \mu (\Phi - G * R_\Phi) G * \frac{\partial R_\Phi}{\partial p} +$$

$$v(D - G * R_D)G * \frac{\partial R_D}{\partial p}, \quad (4.28)$$

where the dependence of the surface gradients and the PSF on u and v has been omitted for clarity. An analogous expression is readily obtained for q .

4.3.3.2 Integration of Accurate but Sparse Depth Information

One possible method to obtain depth information about the surface is stereo image analysis. In our experiments we utilise the correlation-based blockmatching method described in Section 1.5.2. Apart from blockmatching techniques, one might think of employing a dense stereo algorithm which computes a depth value for each image pixel independent of the presence of texture (Horn, 1986; Scharstein and Szeliski, 2002; Hirschmüller, 2006). However, parts of the surface may show no surface texture at all, or corresponding parts of the stereo image pair do not display a similar structure. The latter behaviour e.g. occurs as a consequence of specular reflectance properties leading to a different appearance of the respective surface part in the stereo images. In such cases of missing or contradictory texture information, dense stereo algorithms usually interpolate the surface across the ambiguous image parts, leading to an inaccurate three-dimensional reconstruction result for the corresponding region. Hence, we prefer to compute depth points only in places where point correspondences can be established unambiguously and accurately and to compute dense depth data in a subsequent step based on an integration of the available photometric or photopolarimetric information. Furthermore, if the surface is smooth and textureless, a light pattern projected on it may generate the texture necessary to increase the number of reliable point correspondences (Calow et al., 2002). The metallic surfaces regarded in Chapter 5, however, are sufficiently rough and textured to generate a three-dimensional point cloud without applying structured illumination.

Another technique well suitable to determine a three-dimensional point cloud of the surface to be combined with the SfPR analysis is structure from motion (cf. Section 1.2). The unknown scaling factor may be determined by incorporating depth from defocus information according to Section 4.1 or based on a-priori knowledge about the scene such as the average pixel scale. For surface materials with relatively weak specular reflectance components, active scanning techniques (Battile et al., 1998) may also be favourably applied to obtain a three-dimensional point cloud.

Horovitz and Kiryati (2004) have shown that methods directly enforcing sparse depth constraints during the surface gradient integration suffer from the fact that the sparse depth values only have a local influence and lead to spikes in the reconstructed surface. Hence, they propose a local approach, assigning a radial weighting function to each depth point, and a global approach which consists of interpolating a bias correction surface to the depth differences between gradient reconstruction and depth points. But in their framework the locality of the influence of the depth points on the gradient field is only partially removed.

Our approach to incorporate sparse depth information into the global optimisation scheme presented in Section 4.3.1.1 consists of defining a depth error term based on the surface gradient field and depth differences between sparse three-dimensional points (d'Angelo and Wöhler, 2006). The depth difference between two three-dimensional points at image positions (u_i, v_i) and (u_j, v_j) is given by

$$(\Delta z)_{ij} = z_{u_j v_j} - z_{u_i v_i}. \quad (4.29)$$

The corresponding depth difference of the reconstructed surface gradient field is calculated by integration along a path C_{ij} between the coordinates (u_i, v_i) and (u_j, v_j) :

$$(\Delta z)_{ij}^{\text{surf}} = \int_{C_{ij}} (p \, dx + q \, dy). \quad (4.30)$$

In our implementation the path C_{ij} is approximated by a list of K discrete pixel positions (u_k, v_k) with $k = 1, \dots, K$. While in principle any path C_{ij} between the points (u_i, v_i) and (u_j, v_j) is possible, the shortest integration path, a straight line between (u_i, v_i) and (u_j, v_j) , is used here. Longer paths tend to produce larger depth difference errors because the gradient field is not guaranteed to be integrable.

Using these depth differences, it is possible to extend the global optimisation scheme introduced in Section 4.3.1 by adding the error term e_z which minimises the squared distance between all N depth points according to

$$e_z = \sum_{i=1}^N \sum_{j=i+1}^N \frac{((\Delta z)_{ij} - (\Delta z)_{ij}^{\text{surf}})^2}{\|(u_j, v_j) - (u_i, v_i)\|_2}, \quad (4.31)$$

where $\|\dots\|_2$ denotes the Euclidean distance in the image plane in pixel units. The iterative update rule Eq. (4.21) then becomes

$$\begin{aligned} p_{uv}^{(n+1)} &= \bar{p}_{uv}^{(n)} + \lambda \frac{\partial e_I}{\partial p} + \mu \frac{\partial e_\Phi}{\partial p} + \nu \frac{\partial e_D}{\partial p} \\ &\quad + 2 \chi \sum_{i=1}^N \sum_{j=i+1}^N \left[\frac{(\Delta z)_{ij} - (\Delta z)_{ij}^{\text{surf}}}{\|(u_j, v_j) - (u_i, v_i)\|_2} \right] \frac{\partial (\Delta z)_{ij}^{\text{surf}}}{\partial p} \Big|_{u,v}, \end{aligned} \quad (4.32)$$

An analogous expression is obtained for q . The derivatives of $(\Delta z)_{ij}^{\text{surf}}$ with respect to p and q may only be nonzero if the pixel (u_k, v_k) belongs to the path C_{ij} and are zero otherwise. They are computed based on the discrete gradient field. The derivative depends on the direction $(d_u^{(k)}, d_v^{(k)})$ of the integration path at pixel location (u_k, v_k) with $d_u^{(k)} = u_{k+1} - u_k$ and $d_v^{(k)} = v_{k+1} - v_k$ according to

$$\frac{\partial (\Delta z)_{ij}^{\text{surf}}}{\partial p} \Bigg|_{u_k, v_k} = d_u^{(k)} p_{u_k v_k}$$

$$\left. \frac{\partial(\Delta z)_{ij}^{\text{surf}}}{\partial q} \right|_{u_k, v_k} = d_v^{(k)} q_{u_k v_k}. \quad (4.33)$$

The update of the surface gradient at position (u, v) is then normalised with the number of paths to which the corresponding pixel belongs. Error term (4.31) leads to the evaluation of $N(N - 1)/2$ lines at each update step and becomes prohibitively expensive for a large number of depth measurements. Therefore only a limited number of randomly chosen lines is used during each update step. Due to the discrete pixel grid, the width of each line can be assumed to correspond to one pixel. It is desirable that a large fraction of the image pixels are covered by the lines. For randomly distributed points and square images of size $w \times w$ pixels, we found that about 70 percent of all image pixels are covered by the lines when the number of lines corresponds to $\sim 10w$.

This method is termed shape from photopolarimetric reflectance and depth (SfPRD). It is related to the approach by Wöhler and Hafezi (2005) described in Section 4.2 combining shape from shading and shadow analysis using a similar depth difference error term, which is, however, restricted to depth differences along the light source direction. The method proposed by Fassold et al. (2004) directly imposes depth constraints selectively on the sparse set of surface locations with known depth. As a consequence, in their framework the influence of the depth point on the reconstructed surface is restricted to its immediate local neighbourhood. Horovitz and Kiryati (2004) reduce this effect by applying a weighted least squares extension of depth from gradient field computation and by adding an interpolation surface to the reconstructed shape. In their framework, the influence of the three-dimensional points on the reconstructed surface is better behaved but still decreases considerably with increasing distance. In contrast, our method effectively transforms sparse depth data into dense depth difference data as long as a sufficiently large number of paths C_{ij} is taken into account. The influence of the depth error term is thus extended across a large number of pixels by establishing large-scale surface gradients based on depth differences between three-dimensional points.

4.3.4 Experimental Evaluation Based on Synthetic Data

To examine the accuracy of the three-dimensional surface reconstruction methods described in Sections 4.3.1 and 4.3.3 in comparison to ground truth data and to reveal possible systematic errors, we test our algorithms on synthetically generated surfaces.

Our first evaluation regards the integration of depth from defocus into the SfPR framework based on the synthetically generated surface shown in Fig. 4.20a. For simplicity, to generate the spatially varying blur we assume for the PSF radius σ the relation $\sigma \propto |z - z_0|$ according to Section 3.2.1 (Pentland, 1987). To obtain a visible surface texture in the photopolarimetric images, we introduce random fluctuations of z_{uv} of the order 0.1 pixels. We assume a perpendicular view on the surface along the z

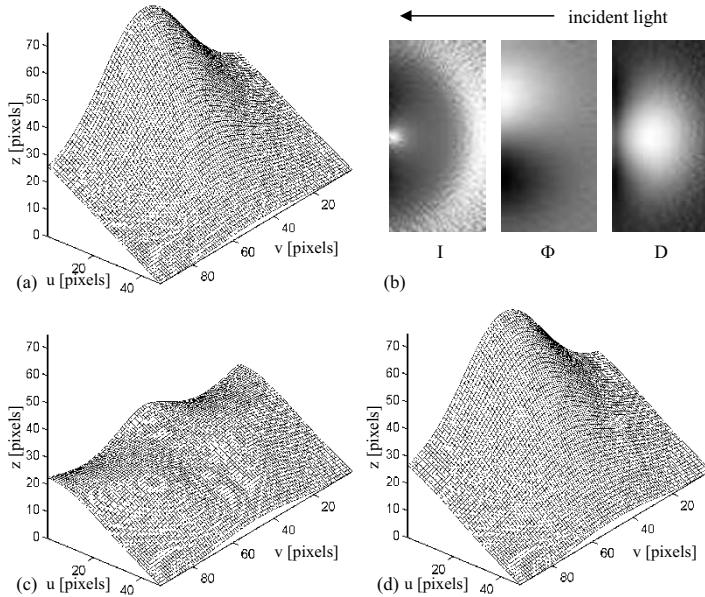


Fig. 4.20 Three-dimensional reconstruction of a synthetically generated surface. (a) Ground truth. (b) Intensity I , polarisation angle Φ , and polarisation degree D images. The three-dimensional reconstruction result was obtained based on photopolarimetric analysis (c) without and (d) with depth from defocus information.

Table 4.3 Evaluation results obtained based on the synthetically generated surface shown in Fig. 4.20a. The error values for z are given in pixels.

Utilised information	RMSE (without DfD)			RMSE (with DfD)		
	z	p	q	z	p	q
Reflectance	11.6	0.620	0.514	9.62	0.551	0.514
Pol. angle	17.0	0.956	0.141	6.62	0.342	0.069
Pol. degree	4.72	0.138	0.514	4.73	0.135	0.514
Pol. angle and degree	1.83	0.121	0.057	1.71	0.119	0.056
Reflectance and pol. angle	12.0	0.528	0.099	2.52	0.280	0.055
Reflectance and pol. degree	10.9	0.575	0.514	8.46	0.418	0.517
Reflectance and polarisation (angle and degree)	10.2	0.277	0.072	0.91	0.091	0.050

axis with $\mathbf{v} = (0, 0, 1)^T$. The surface is illuminated from the right hand side at a phase angle of $\alpha = 75^\circ$. The surface albedo ρ was computed from Eq. (4.23) based on the specular reflections. We set $p_{uv}^{(0)} = p_{uv}^{\text{DfD}}$ and $q_{uv}^{(0)} = q_{uv}^{\text{DfD}}$ in Eq. (4.21) when using depth from defocus information, while otherwise, the PSF G is set to unity and the surface gradients are initialised with zero values due to the lack of prior information. Fig. 4.20c–d illustrates that the main benefit of depth from defocus analysis comes from the improved initialisation which prevents that the SfPR algorithm converges towards a local, suboptimal minimum of the error function. The best results are

Table 4.4 Three-dimensional reconstruction results for the synthetic ground truth surface shown in Fig. 4.21a. If not identified otherwise, the results were obtained with the global optimisation approach. The error values for z are given in pixels.

Method, utilised information	Figure	RMSE (without noise)			RMSE (with noise)		
		z	p	q	z	p	q
I_1	4.21d	4.21d	1.11	0.046	0.077	1.11	0.047
ϕ_1	—	—	2.13	0.102	0.059	3.92	0.163
I_1, I_2	4.21e	4.21e	0.22	0.012	0.018	0.21	0.014
I_1, I_2 (local)	4.21j	4.21j	0.00	0.000	0.000	0.19	0.046
I_1, ϕ_1	4.21f	4.21f	0.17	0.012	0.007	0.19	0.040
I_1, ϕ_1 (local)	4.21k	4.21k	0.00	0.000	0.000	0.28	0.134
I_1, D_1	—	—	1.11	0.044	0.077	1.13	0.098
I_1, D_1 (local)	—	—	2.12	0.088	0.178	7.17	0.837
I_1, ϕ_1, D_1	—	—	0.01	0.001	0.001	0.52	0.103
I_1, ϕ_1, D_1 (local)	—	—	0.00	0.000	0.000	0.31	0.149
I_1, I_2, ϕ_1	—	—	0.11	0.009	0.005	0.20	0.034
I_1, I_2, ϕ_1 (local)	—	—	0.00	0.000	0.000	0.35	0.074
I_1, I_2, ϕ_1, ϕ_2	—	—	0.01	0.001	0.001	0.21	0.056
I_1, I_2, ϕ_1, ϕ_2 (local)	—	—	0.00	0.000	0.000	0.24	0.057
z	4.21g	4.21g	0.14	0.012	0.013	0.20	0.034
I_1, z	4.21h	4.21h	0.11	0.008	0.009	0.15	0.023
I_1, ϕ_1, z	4.21i	4.21i	0.09	0.006	0.005	0.13	0.042
I_1, I_2, ϕ_1, z	—	—	0.09	0.006	0.005	0.15	0.036
$I_1, I_2, \phi_1, \phi_2, z$	—	—	0.07	0.004	0.003	0.11	0.052

obtained by utilising a combination of polarisation angle and degree, of reflectance and polarisation angle, or a combination of all three features (cf. Table 4.3).

To examine the behaviour of the local and global optimisation schemes and their combination with sparse depth data, dependent on how many images based on which reflectance and polarisation features are used, we apply the developed algorithms to the synthetically generated surface shown in Fig. 4.21a (d’Angelo and Wöhler, 2008). We again assume a perpendicular view on the surface along the z axis, corresponding to $\mathbf{v} = (0, 0, 1)^T$. The scene is illuminated sequentially by $L = 2$ light sources under an angle of 15° with respect to the horizontal plane at azimuth angles of $\psi^{(1)} = -30^\circ$ and $\psi^{(2)} = +30^\circ$, respectively. This setting results in identical phase angles $\alpha^{(1)} = \alpha^{(2)} = 75^\circ$ for the two light sources. A set of 500 random points is extracted from the ground truth surface, to which Gaussian noise is added as described below prior to using them as sparse depth data for three-dimensional reconstruction.

The reflectance functions of the rough metallic surface measured according to Sections 2.2.1 and 2.4.2 were used to render the synthetic images shown in Fig. 4.21c. Gaussian noise is applied with a standard deviation of 5×10^{-4} for the intensity I , where the maximum grey value is about 6×10^{-2} , 1° for the polarisation angle Φ , and 0.4 pixels for the depth values (z between 0 and 6 pixels). The weights for the error terms according to Eq. (4.32) are set to $\lambda = 450$, $\mu = 40$, $\nu = 100$, and $\chi = 1$. The surface gradients are initialised with zero values.

Fig. 4.21 shows the reconstruction results on noisy synthetic images, where the plots (d)–(f), (j), and (k) were obtained by applying SfPR alone, while the plots

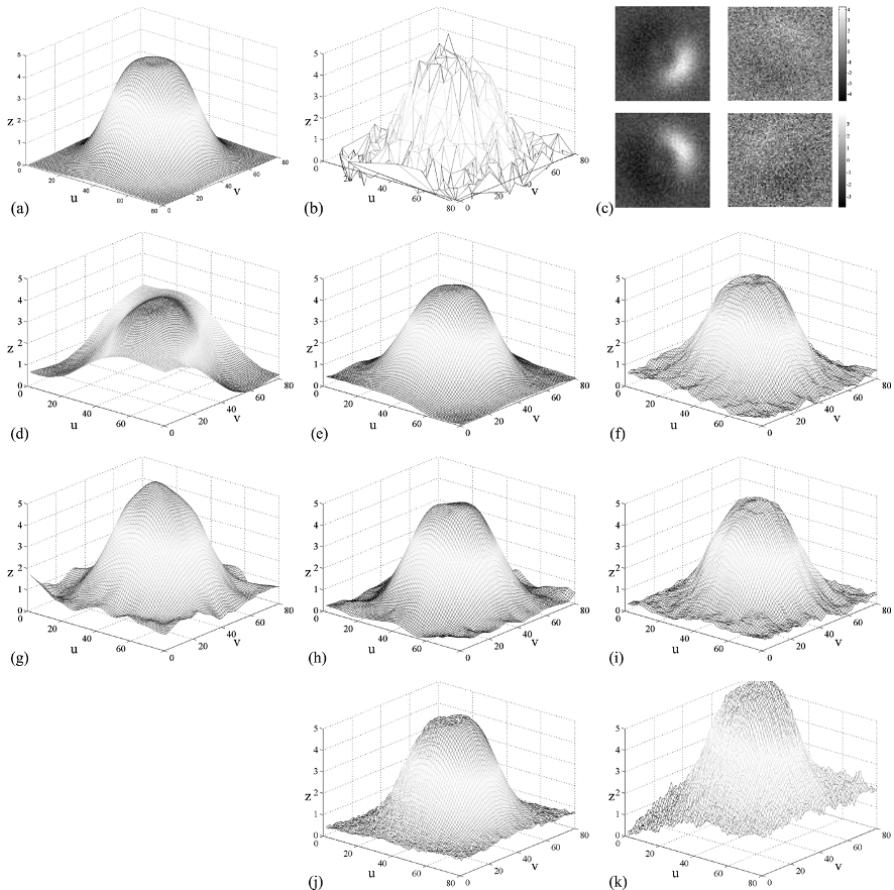


Fig. 4.21 Three-dimensional reconstruction results for a synthetically generated surface. (a) Ground truth. (b) Noisy depth data. (c) Noisy intensity and polarisation angle images, based on measured reflectance functions of a raw forged iron surface. The reconstruction result for noisy images of a surface with uniform albedo, obtained by SfPR with global optimisation without integration of sparse depth data, is shown in (d) using a single intensity image, in (e) using both intensity images, and in (f) using one intensity and one polarisation angle image. (g) Reconstructed surface obtained based on noisy sparse depth data alone. (h) Reconstruction result using sparse depth data and intensity. (i) Reconstruction result using sparse depth data, intensity, and polarisation angle. For comparison, the reconstruction result obtained based on SfPR with local optimisation and without sparse depth data is shown in (j) using both intensity images and in (k) using one intensity and one polarisation angle image.

(g)–(i) depict the results obtained based on integration of sparse depth data into the SfPR framework. The respective reconstruction errors are given in Table 4.4. It is apparent that the shape from shading reconstruction using a single light source fails to reconstruct the surface (Fig. 4.21d), while the surface shape can be reconstructed approximately using a single intensity and polarisation angle image (Fig. 4.21f).

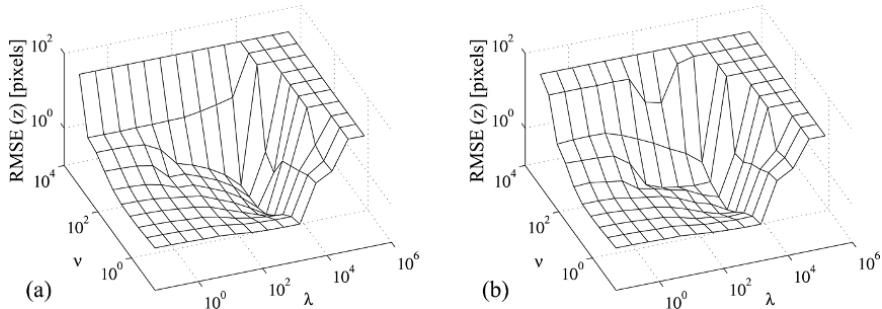


Fig. 4.22 Dependence of the reconstruction error of the SfPR approach on the weight parameters λ and μ according to Eqs. (4.21) and (4.32) for the synthetic example shown in Fig. 4.21a without noise (left) and with noise (right).

To reach similar reconstruction accuracy without polarisation information, illumination from two different directions is required (Fig. 4.21e). Table 4.4 illustrates that using intensity and polarisation degree in the three-dimensional reconstruction process leads to poor accuracy both for the global and the local approach, while using intensity and polarisation angle yields a high accuracy which does not further increase when the polarisation degree is additionally used. The reason for this behaviour is the fact that intensity and polarisation degree contain somewhat redundant information, since both display a maximum in or near the specular direction ($\theta_r = 0^\circ$) and decrease in a qualitatively similar lobe-shaped manner for increasing value of θ_r . The dependence on surface orientation, however, is much stronger for the intensity than for the polarisation degree, while the measurement error tends to be significantly lower for the intensity. The local optimisation approach according to Section 4.3.1.2 provides a very accurate reconstruction for the noise-free case, but performs worse than the global approach on noisy data (cf. Fig. 4.21j–k). This property can be observed clearly by comparing the corresponding reconstruction errors of p and q given in Table 4.4. With intensity and polarisation angle images, the reconstruction result becomes very accurate. Similarly accurate reconstruction results, however, are already obtained based on a single intensity and polarisation image.

Fig. 4.21g shows the reconstruction result using only the sparse depth values, effectively smoothing and interpolating the sparse depth values shown in Fig. 4.21b. The overall shape is correct, but smaller details like the flattened top of the object are missing in the reconstructed three-dimensional profile. Adding intensity and polarisation terms improves the results and captures the finer details which are not visible in the sparse depth data (cf. Fig. 4.21h–i).

The values for the weight parameters of the error terms according to Eqs. (4.21) and (4.32) are related to the magnitudes of the intensity and polarisation features and their measurement uncertainties. The influence of the weight parameters on the reconstruction accuracy has been evaluated using the previously described synthetic

data. As a typical example, Fig. 4.22 shows the root mean square depth error of the reconstructed surface profile obtained from one intensity and one polarisation angle image for different weight parameters λ and μ . For noise-free image data, the reconstruction error decreases with increasing λ and μ until the algorithm starts to diverge at fairly well-defined critical values. For noisy input images (cf. Fig. 4.21c) the reconstruction error displays a weaker dependence on λ and μ and a less pronounced minimum. This is a favourable property since small changes in the weight parameters do not lead to large differences in the reconstruction accuracy as long as the values chosen for λ and μ are well below their critical values for which the algorithm begins to diverge.

4.3.5 Discussion

In this section we have presented an image-based three-dimensional surface reconstruction method relying on simultaneous evaluation of intensity and polarisation features and its combination with depth data.

The shape from photopolarimetric reflectance (SfPR) technique is based on the analysis of single or multiple intensity and polarisation images. The surface gradients are determined based on a global optimisation method involving a variational framework and based on a local optimisation method which consists of solving a set of nonlinear equations individually for each image pixel. These approaches are suitable for strongly non-Lambertian surfaces and surfaces of diffuse reflectance behaviour.

The shape from photopolarimetric reflectance and depth (SfPRD) method integrates independently measured absolute depth data into the SfPR framework in order to increase the accuracy of the three-dimensional reconstruction result. In this context we concentrated on dense but noisy depth data obtained by depth from defocus and on sparse but more accurate depth data obtained e.g. by stereo analysis or structure from motion. However, our framework is open for independently measured three-dimensional data obtained from other sources such as laser triangulation.

We have shown that depth from defocus information can be used for determining the large-scale properties of the surface and for appropriately initialising the surface gradients. At the same time it provides an estimate of the surface albedo. For integration of sparse depth information, we have suggested an optimisation scheme that simultaneously adapts the surface gradients to the measured intensity and polarisation data and to the surface slopes implied by depth differences between pairs of depth points. This approach transforms sparse depth data into dense depth difference data, leading to a non-local influence of the depth points on the reconstructed surface profile.

Our experimental evaluation based on synthetic data illustrates that including polarisation information into the three-dimensional reconstruction scheme significantly increases the accuracy of the reconstructed surface. The main benefit arises from taking into account polarisation angle data, while intensity and polarisation

degree tend to contain redundant information. We found that taking into account dense but noisy depth from defocus data may be helpful to estimate the surface albedo and to avoid local minima of the error function. The integration of sparse but accurate depth data significantly increases the three-dimensional reconstruction accuracy especially on large scales.

4.4 Stereo Image Analysis of Non-Lambertian Surfaces

A general drawback of all methods for stereo image analysis mentioned in Section 1.5 is the fact that they implicitly assume Lambertian reflectance properties of the object surfaces. Two images of the surface are acquired from different viewpoints, and two image parts are assumed to correspond to the same physical surface point if their appearances are similar. This is only the case for Lambertian surfaces, where the surface brightness is independent of the viewpoint. However, even in the Lambertian case geometric distortions between the two images occur, which may be taken into account by estimating an affine transformation between the views (Shi and Tomasi, 1994). A method for three-dimensional reconstruction of objects from image sequences that accounts for the changing camera viewpoint by a combination of structure from motion and photometric stereo is presented by Lim et al. (2005). The reflectance behaviour of the surface, however, is still explicitly assumed to be Lambertian.

Since specular reflections are viewpoint dependent, they may cause large intensity differences at corresponding image points. As a consequence of this behaviour, stereo analysis is often unable to establish correspondences at all, or the inferred disparity values tend to be inaccurate, or the established correspondences do not belong to the same physical surface point. Only few methods that specifically address the three-dimensional reconstruction of specular surfaces have been proposed in the literature. A technique based on deflectometry, i.e. analysis of the deformation of fringe patterns projected on the surface, is introduced by Lowitzsch et al. (2005). It is restricted to mirror-like surfaces. Bhat and Nayar (1996) infer a relationship between stereo vergence, i.e. the angle between the optical axes of the stereo cameras, surface roughness, and the likelihood of a correct stereo match from the physics of specular reflection and stereo geometry. This approach attempts to minimise the influence of specular reflections on the images by choosing an appropriate configuration of the stereo camera pair but does not consider the effect of specular reflections in the stereo matching process.

A method for separating the specular from the Lambertian reflectance component based on the dichromatic reflection model (DRM) is described by Klette et al. (1999). This model can be used for inhomogeneous dielectric materials with a surface structure which can be described by an interface and an optically neutral medium containing colour pigments. Body reflection, i.e. reflection of incident light at the optically neutral medium, is diffuse, while reflection at the interface is specular. The resulting scene radiance is characterised by the interface reflection colour

and the body reflection colour. Klette et al. (1999) show that the colour values of the light reflected from the surface lie in a plane in the RGB colour space, the so-called dichromatic plane. Clustering the resulting points in the dichromatic plane based on colour histograms allows to separate the two reflectance components and to remove the specular reflectance component from the image. Three-dimensional surface reconstruction is then performed based on the diffuse reflectance component alone.

The multi-image shape from shading method by Lohse et al. (2006) can in principle be used with an arbitrary but precisely known reflectance function. Their method directly relates image grey values to a three-dimensional surface model defined in the world coordinate system and to the parameters of the reflectance function. The three-dimensional shape of the surface, which is determined directly in object space, is estimated along with the photometric parameters from the image grey values in a least-mean-squares adjustment, where a uniform surface albedo is assumed. This method is favourably used under an oblique viewing geometry since point correspondences are established only implicitly by comparing the observed with the modelled pixel brightnesses. Depth differences more strongly translate into offsets in the image coordinates and thus have a more pronounced influence on the utilised error function when the surface is observed under an oblique angle. However, the experimental evaluation by Lohse et al. (2006) does not involve specular surfaces.

BRDF independent stereo image analysis can be achieved by exploiting the Helmholtz reciprocity condition (2.8), stating that the value of the BRDF does not change when the camera and the light source are exchanged (Magda et al., 2001; Zickler et al., 2002, 2003a,b). By attaching a point light source to each camera, reciprocal image pairs are acquired such that the image taken by the first camera is illuminated by the second light source and vice versa. Due to reciprocity, identical amounts of light are received by the first and the second camera, respectively. This method, however, requires that the light sources are located in the optical centres of the cameras, since otherwise the reconstruction result becomes inaccurate. Furthermore, the optical systems need to be calibrated with respect to the light sources.

Other methods for BRDF independent stereo image analysis are based on radiometric illumination variations, i.e. the light sources are kept at fixed locations while the emitted intensity distribution is changed. This is the case e.g. for coded structured light techniques (Battile et al., 1998) and also the spacetime stereo framework introduced by Davis et al. (2005). Wolff and Angelopoulou (1994) propose a method to obtain a dense set of point correspondences which is based on photometric ratios. Two stereo image pairs of the scene are taken from the same viewpoint but under different illumination conditions, including geometric and radiometric variations. A diffuse reflectance model for smooth dielectric surfaces is utilised, which is more complex than the Lambertian reflectance law but does not represent specular reflections. Wolff and Angelopoulou (1994) show that the pair of ratio images obtained by dividing two stereo image pairs acquired under different illumination conditions by each other is invariant with respect to camera gain, viewpoint, and albedo as long as no specular reflections occur. Stereo correspondences between the pair of ratio images are established by a direct pixelwise comparison of the ratio values. Jin et al.

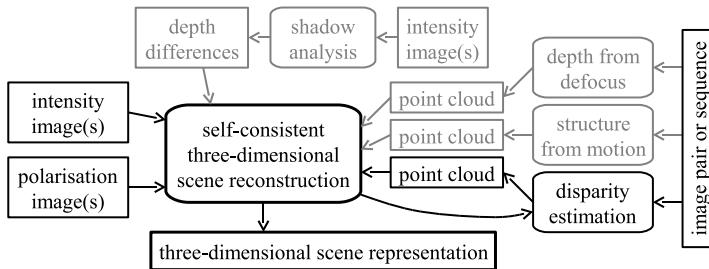


Fig. 4.23 Overview of the specular stereo approach.

(2003) show that a multi-view rank constraint on reflectance complexity is implied by a reflectance function consisting of a diffuse and an additive specular component. They utilise this constraint for three-dimensional reconstruction of non-Lambertian surfaces. For stereo image pairs acquired under geometrically constant but radiometrically variable illumination conditions, a different invariant termed light transport constancy (LTC) is exploited by Wang et al. (2007) which can be formulated as a rank constraint on the matrix denoting the irradiances observed by the cameras under the different illumination conditions. They show that establishing stereo correspondences based on ratio images similar to Wolff and Angelopoulou (1994) is mathematically equivalent to evaluating the rank constraint. Due to the fact that Wang et al. (2007) only allow radiometric illumination variations, their method is not restricted to diffusely reflecting surfaces but can be applied to surfaces with arbitrary BRDFs. In their experiments, radiometric illumination variations are generated by two stationary light projectors.

This section describes a method for three-dimensional reconstruction of surfaces with non-Lambertian reflectance properties based on stereo image analysis (Wöhler, 2008; Wöhler and d'Angelo, 2009). Using the SfPRD technique described in Section 4.3 as a basis, geometric cues are combined with photometric and polarimetric information into an iterative framework that allows to establish stereo correspondences in accordance with the specular reflectance behaviour and at the same time to determine the surface gradient field based on the known photometric and polarimetric reflectance properties (cf. Fig. 4.23). Illumination by a single point light source is sufficient, no variable illumination is required. Disparities are refined based on a comparison between observed and modelled pixel brightnesses. The approach yields a dense three-dimensional reconstruction of the surface which is consistent with all observed geometric and photopolarimetric data.

4.4.1 Iterative Scheme for Disparity Estimation

We utilise a correlation-based blockmatching stereo algorithm (cf. Section 1.5.2) to obtain depth information about the surface. The images are rectified to standard

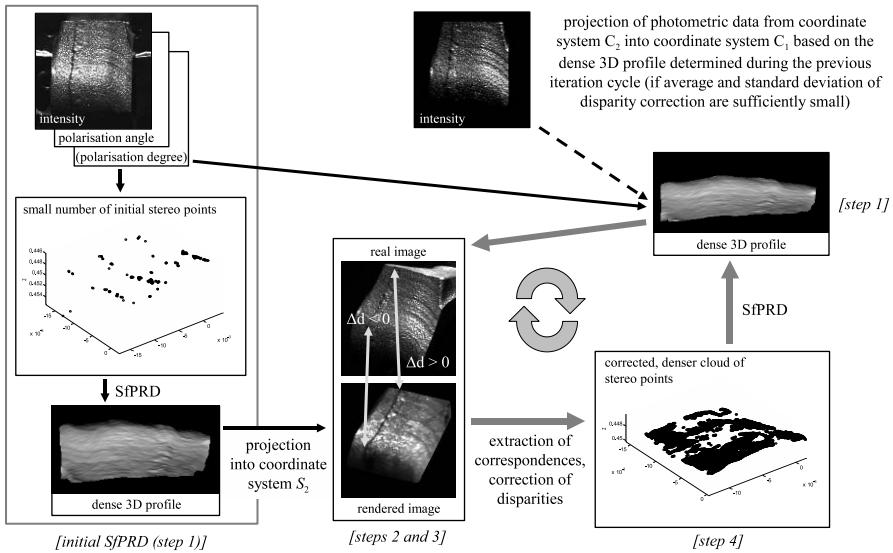


Fig. 4.24 Schematic description of the specular stereo algorithm.

stereo geometry, resulting in epipolar lines corresponding to the image rows (cf. Section 1.5). Directly applying the stereo algorithm to an image pair of a rough metallic surface usually results in a fairly sparse disparity map due to limited texture, repeating patterns, or different appearance of corresponding surface parts in the stereo images as a consequence of the strongly specular reflectance behaviour.

The coordinate systems of the two cameras are denoted by the indices C_1 (left camera) and C_2 (right camera), the corresponding rectified coordinate systems by R_1 (left rectified camera) and R_2 (right rectified camera). The transformations between these coordinate systems and therefore also the viewing directions $C_1\mathbf{v}_1$ and $C_1\mathbf{v}_2$ of the cameras are known from the extrinsic camera calibration, e.g. in coordinate system C_1 . We assume that the surface is illuminated with a point light source situated at infinite distance in a direction given by the vector $C_1\mathbf{s}$. The intensity and polarisation angle reflectance functions R_I and R_Φ are assumed to be known from a reference measurement. The proposed stereo image analysis method for non-Lambertian surfaces is termed specular stereo. It consists of the following steps (cf. also Fig. 4.24):

1. *Compute a three-dimensional surface profile based on SfPRD:* A three-dimensional surface profile is computed with the SfPRD method based on the intensity and polarisation data of camera 1 and the depth points $C_1\mathbf{r}_j^{(m)}$ obtained by stereo analysis, where m denotes the iteration cycle index. Each pixel is regarded as a three-dimensional point $C_1\mathbf{x}_i^{(m)}$. For each pixel the surface normal $C_1\mathbf{n}_i^{(m)}$ is known as a result of the SfPRD method. The three-dimensional point cloud is transformed into the rectified coordinate system S_2 of camera 2 accord-

ing to

$${}^{S_2}\mathbf{x}_i^{(m)} = \frac{S_2}{C_1} \mathcal{T} \left({}^{C_1}\mathbf{x}_i^{(m)} \right), \quad (4.34)$$

where $\frac{S_2}{C_1} \mathcal{T}$ denotes the transformation (a rotation and a translation) from coordinate system C_1 into coordinate system S_2 . The same transformation is performed for the surface normals ${}^{C_1}\mathbf{n}_i^{(m)}$ and the illumination vector ${}^{C_1}\mathbf{s}$, resulting in the vectors ${}^{S_2}\mathbf{n}_i^{(m)}$ and ${}^{S_2}\mathbf{s}$.

2. *Render a synthetic image for rectified camera 2:* Based on the known reflectance function, a synthetic image $R_I^{(m)}({}^{S_2}u, {}^{S_2}v)$ is rendered, which represents the pixel intensities expected for the rectified coordinate system S_2 .
3. *Determine disparity corrections:* Deviations between the estimated and the true surface profile are now revealed by a position-dependent lateral offset $\Delta d_j^{(m)}({}^{S_2}u_j^{(m)}, {}^{S_2}v_j^{(m)})$ between the rendered and the observed image of rectified camera 2. In each iteration cycle m , the blockmatching stereo algorithm re-determines the pixels $({}^{S_2}u_j^{(m)}, {}^{S_2}v_j^{(m)})$ for which correspondences between the rendered and the observed image in the rectified coordinate system S_2 can be established. Due to the chosen standard geometry, a depth error of a pixel in the image of camera 1 translates into an offset along the corresponding epipolar line, i.e. image row, in the rectified image of camera 2. The value of $\Delta d_j^{(m)}({}^{S_2}u_j^{(m)}, {}^{S_2}v_j^{(m)})$ corresponds to the disparity error of the pixel at $({}^{S_2}u_j^{(m)}, {}^{S_2}v_j^{(m)})$ in the rectified image of camera 2. We determine the offset based on the same correlation-based blockmatching approach as utilised for the initial stereo image analysis.
4. *Compute corrected three-dimensional points:* The positions $({}^{S_2}u_j^{(m)}, {}^{S_2}v_j^{(m)})$ and disparities $d_j^{(m)}$ of all pixels for which the blockmatching technique is able to determine a value $\Delta d_j^{(m)}$ are updated according to

$$\begin{aligned} {}^{S_2}u_j^{(m), \text{corr}} &= {}^{S_2}u_j^{(m)} - \Delta d_j^{(m)} \\ {}^{S_2}v_j^{(m), \text{corr}} &= {}^{S_2}v_j^{(m)} \\ d_j^{(m), \text{corr}} &= d_j^{(m)} + \Delta d_j^{(m)}. \end{aligned} \quad (4.35)$$

The corrected three-dimensional point cloud ${}^{S_2}\mathbf{r}_j^{(m+1)}$ is obtained from the corrected pixel positions $({}^{S_2}u_j^{(m), \text{corr}}, {}^{S_2}v_j^{(m), \text{corr}})$ and disparities $d_j^{(m), \text{corr}}$ determined according to Eq. (4.35), relying on the basic equations of stereo analysis in standard epipolar geometry (Horn, 1986). Transformed into the coordinate system of camera 1, the corrected three-dimensional points are denoted by ${}^{C_1}\mathbf{r}_j^{(m+1)}$. Finally, the iteration cycle index m is incremented: $m \leftarrow m + 1$.

5. Iterate steps 1–4 until the average and the standard deviation of the disparity corrections $\Delta d_j^{(m)}$ are of the order of 1 pixel. Note that the disparities $d_j^{(m)}$ are measured between the observed rectified images with coordinate systems S_1 and

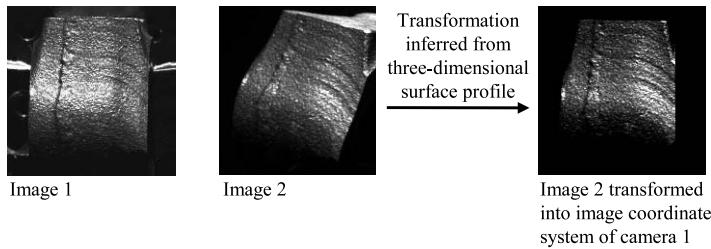


Fig. 4.25 Transformation of image 2 into the image coordinate system of camera 1 after the average and the standard deviation of the disparity correction Δd_j have decreased to less than about one pixel.

S_2 , while the disparity corrections $\Delta d_j^{(m)}$ are measured between the rendered and the observed image in the rectified coordinate system S_2 .

Once this degree of self-consistency is reached, it is favourable to additionally take into account the photopolarimetric information of camera 2. In our experiments (cf. Section 5.3.4), we do not acquire a second polarisation image—it would be necessary to perform the difficult task of absolutely calibrating the rotation angles of two polarising filters with respect to each other—but merely use the intensity information of camera 2. In Eq. (4.20), the intensity error e_I then consists of two parts according to $e_I = e_I^{(1)} + e_I^{(2)}$. For this purpose, the image of camera 2 is transformed during step 1 of the iteration scheme from coordinate system C_2 to C_1 (cf. Fig. 4.25). In iteration cycle m , the appropriate transformation is obtained based on the three-dimensional surface profile defined by the points $C_2 \mathbf{x}_i^{(m-1)}$ obtained during the previous iteration cycle. The resulting transformed image of camera 2 and the image of camera 1 are pixel-synchronous at reasonable accuracy due to the already achieved smallness of the disparity corrections.

Including the accordingly transformed intensity information of camera 2 into the optimisation scheme corresponds to a photometric stereo approach which exploits the effect of different viewpoints on the measured brightness of the surface, while the direction of illumination remains constant. This technique does not provide additional photometric constraints for Lambertian surfaces as their brightness is independent of the viewing direction, which is the main reason why traditional photometric stereo (Horn, 1986) relies on multiple illumination directions rather than multiple viewpoints.

The described iterative scheme for specular stereo analysis is visualised in Fig. 4.24 for the connection rod example regarded in detail in Section 5.3.4. The initial SfPRD step according to the left part of Fig. 4.24 (marked as step 1) yields a dense 3D surface profile which results in a rendered image in the rectified camera coordinate system S_2 (step 2) that does not correspond very well with the rectified image observed by camera 2. Determining correspondences between the rendered and the observed image (step 3) and generating an accordingly corrected 3D point

cloud (step 4) yields a dense 3D surface profile which displays a similar small-scale structure as the initial 3D profile but a lower large-scale slope. Repeating steps 1–4 during subsequent iterations and taking into account the intensity information of camera 2 further refines the 3D surface profile.

The specular stereo method explicitly models the appearance of the surface in the two cameras based on the known reflectance properties. Especially, the behaviour of specular reflections subject to changing viewpoint and the resulting effects on the estimation of disparities are taken into account. Furthermore, it is possible to utilise the method for large baseline stereo, where the stereo baseline is comparable to the object distance, as the differences in perspective introduced by the strongly different viewpoints are explicitly considered as well. All available photopolarimetric and geometric information is utilised for the purpose of 3D surface reconstruction.

The iterative optimisation scheme described in this section assumes convergence. However, the solution is not guaranteed to converge since the photopolarimetric and geometric information may in principle be contradictory e.g. due to camera calibration errors or inaccurate knowledge of the illumination direction. However, we observed that in all experiments regarded in Section 5.3.4 convergence is achieved after 4–8 iteration cycles. Self-consistent measures for assessing the 3D reconstruction accuracy and convergence behaviour are discussed in Section 5.3.4.3.

4.4.2 Qualitative Behaviour of the Specular Stereo Algorithm

Fig. 4.26 shows the observed versus the rendered images of the specular surface of the connection rod. While initially in camera 1 the rendered image is very similar to the observed image since the initial SfPRD step aims for obtaining a small value of e_I according to Eq. (2.20), the differences between the observed image in camera 2 and the correspondingly rendered image are evident. During the initial SfPRD step, the photometric information of image 2 is not taken into account. The differences in surface shape occur as a consequence of the large-scale surface slope in vertical image direction being inaccurately estimated by the initial SfPRD step—for only less than 0.5 percent of the image pixels an initial disparity value can be determined. Differences in surface brightness (the right third of the surface profile appears bright in the observed image but fairly dark in the rendered image) are due to the same large-scale inaccuracies of the initial three-dimensional shape estimate which result in inaccurate surface gradients and thus estimated surface brightness.

After 8 cycles of the iterative scheme proposed in Section 4.4.1, the geometric appearance of the surface and the distribution of surface brightness in both rendered images closely resemble the corresponding observed images. Now for 18 percent of the image pixels a disparity value can be determined. This example qualitatively illustrates the convergence behaviour of the proposed algorithm. Section 5.3.4 provides a more detailed experimental evaluation in the application scenario of industrial quality inspection.

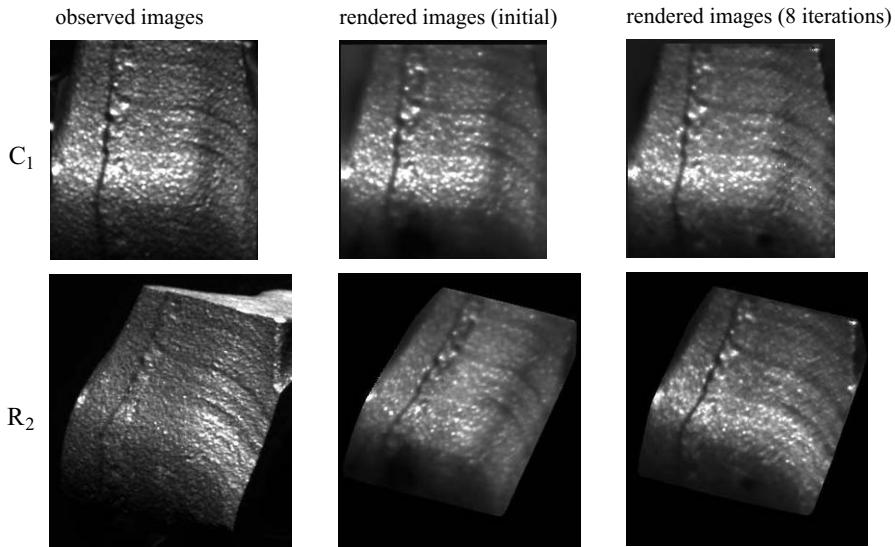


Fig. 4.26 Observed (left) vs. rendered (middle, right) images for the connection rod example in camera coordinate system C_1 and in rectified coordinate system R_2 , respectively, shown before the first iteration cycle and after eight iteration cycles.

4.5 Three-dimensional Pose Estimation Based on Combinations of Monocular Cues

This section describes integrated approaches to the problem of three-dimensional pose estimation (cf. Section 1.6.1). Most appearance-based approaches to three-dimensional pose estimation explicitly rely on point features or edges. However, in the presence of cluttered background or low contrast between object and background, edge information tends to be an unreliable cue for pose estimation. The first proposed technique (cf. Section 4.5.1) applies the principles of combining several sources of complementary information for three-dimensional surface reconstruction, which have been derived in the previous sections of this chapter, to the problem of three-dimensional pose estimation (Barrois and Wöhler, 2007). This appearance-based method relies on the comparison of the input image to synthetic images generated by an OpenGL-based renderer using model information about the object provided by CAD data. The comparison provides an error measure which is minimised by an iterative optimisation algorithm. Although all six degrees of freedom are estimated, the described approach requires only a monocular camera, circumventing disadvantages of multiocular camera systems such as the need for extrinsic camera calibration. Our framework is open for the inclusion of independently acquired depth data. A first evaluation is performed based on a simple but realistic example object.

The second proposed method (cf. Section 4.5.2) combines the contracting curve density (CCD) algorithm described in Section 1.6.2 with depth from defocus data as a regularisation term for the optimisation process. Here it is assumed that the object boundary determined by the CCD algorithm is an ideal edge in the world which displays a spatially varying amount of defocus in the image, depending on the distance of the object points to the camera. This approach is not very robust in realistic applications but illustrates the mathematical principle how additional, independent information apart from pixel grey values can be integrated into the CCD algorithm. As an evaluation, we present three-dimensional pose estimation results obtained with a monocular camera regarding a simple example object, using the implementation by Krauß (2006).

Classical monocular pose estimation approaches have in common that they are not able to estimate the distance to the object at reasonable accuracy, since the only available information is the scale of a known object in the resulting image. Scale information yields no accurate results since for small distance variations the object scale does not change significantly. In comparison, for a convergent stereo setup with a baseline similar to the object distance, for geometrical reasons a depth accuracy of the same order as the lateral translational accuracy is obtainable. For this reason, a variety of three-dimensional pose estimation methods relying on multiple images of the scene have been proposed (cf. Section 1.6 for an overview). However, from the practical point of view, using a monocular camera system is often favourable (cf. Section 5.1) while nevertheless a high pose estimation accuracy may be required e.g. to detect subtle deviations between the true and the desired object pose. In this section we therefore regard a monocular configuration, which can be extended towards a multiocular system as conceptionally outlined by Krüger (2007).

4.5.1 Appearance-based Pose Estimation Relying on Multiple Monocular Cues

4.5.1.1 Photometric and Polarimetric Information

The first source of information we exploit is the intensity reflected from the object surface. For this purpose, we make use of the two-component specular reflectance function inspired by Nayar et al. (1991) and defined by Eq. (4.22). The unknown surface albedo ρ is estimated by the optimisation algorithm described below. Although we regard objects of uniform surface albedo in our experiments, our framework would in principle allow to render and investigate objects with a textured surface by using texture mapping in combination with an estimation of the factor ρ . The other parameters of the reflectance function are determined empirically as described in Section 2.2.1, regarding a sample of the corresponding surface material attached to a goniometer.

The determined parameters of the reflectance function and a CAD model of the object are used to generate a synthetic image of the observed scene. For this pur-

pose, an OpenGL-based renderer has been implemented. The surface orientation is required for each point of the object surface to compute a reflectance value according to Eq. (4.22) but OpenGL does not directly provide this information. Hence, the technique developed by Decaudin (1996) is used to calculate the surface normal for every pixel based on three rendered colour images obtained by a red, a green, and a blue virtual light source appropriately distributed in space. Afterwards, the reflectance function (4.22) is used to compute the predicted intensity for each pixel. We obtain a photorealistic image $R_I(p_{uv}, q_{uv})$ which can be compared with the input image I_{uv} , resulting in the intensity error term

$$e_I = \sum_{u,v} [I_{uv} - R_I(p_{uv}, q_{uv})]^2. \quad (4.36)$$

The summation is carried out for the rendered pixels representing the object surface. A disadvantage of the technique proposed by Decaudin (1996) is the fact that no shadow information is generated for the scene. Hence, shadows are computed in a further raytracing step after the photorealistic rendering process.

Furthermore, we introduce an analogous error term e_Φ taking into account the polarisation angle Φ_{uv} of the light reflected from the object surface. We utilise the polarisation reflectance function $R_\Phi(p_{uv}, q_{uv})$ according to Eq. (2.60) as defined in Section 2.4.2 with empirically determined parameters. The renderer then predicts the polarisation angle for each pixel, resulting in the error term

$$e_\Phi = \sum_{u,v} [\Phi_{uv} - R_\Phi(p_{uv}, q_{uv})]^2. \quad (4.37)$$

In principle, a further error term denoting the polarisation degree might be introduced at this point. However, in all our experiments we found that the polarisation degree is an unreliable feature with respect to three-dimensional pose estimation of objects with realistic surfaces, as it depends more strongly on small-scale variations of the microscale roughness of the surface than on the surface orientation itself.

4.5.1.2 Edge Information

To obtain information about edges in the image, we compute a binarised edge image from the observed intensity image using the Canny edge detector (Canny, 1986). In a second step, a distance transform image C_{uv} is obtained by computing the Chamfer distance for each pixel (Gavrila and Philomin, 1999). As our approach compares synthetically generated images with the observed image, we use a modified Chamfer matching technique. The edges in the rendered image are extracted with a Sobel edge detector, resulting in a Sobel magnitude image E_{uv} , which is not binarised. To obtain an error term which gives information about the quality of the match, a pixel-wise multiplication of C_{uv} by E_{uv} is performed. The advantage of omitting the binarisation is the continuous behaviour of the resulting error function with respect to the pose parameters, which is a favourable property regarding the optimisation

stage. If the edge image extracted from the rendered image were binarised, the error function would become discontinuous, making the optimisation task more difficult. Accordingly, the edge error term e_E is defined as

$$e_E = - \sum_{u,v} C_{uv} E_{uv}, \quad (4.38)$$

where the summation is carried out over all image pixels (u, v) . The minus sign in Eq. (4.38) arises from the fact that our optimisation scheme aims for a determination of the minimum of the error function.

4.5.1.3 Defocus Information

We utilise the depth from defocus technique described in Section 3.2.3.2 to estimate depth values from the amount of defocus. This approach requires two pixel-synchronous images, one of which is acquired with a small aperture ($\kappa = 8$), while the second one is acquired with a large aperture ($\kappa = 2$). This procedure may be automated using a lens equipped with a motorised iris. For the first image we assume that no perceivable amount of defocus is present. The images are partitioned into windows of 32×32 pixels size. The PSF radius Σ in frequency space is computed by fitting a Gaussian to the ratio of the amplitude spectra of the corresponding windows of the first and the second image, respectively. Only the range of intermediate spatial frequencies is regarded in order to reduce the influence of noise on the resulting value of Σ . The depth–defocus function according to Eq. (3.17) is calibrated using the combined chequerboard and random dot pattern shown in Fig. 3.9.

4.5.1.4 Total Error Optimisation

To start the optimisation process, an initial object pose has to be provided. With this pose a first set of images (intensity, polarisation angle, edges, and depth map) is rendered. Each measured cue provides an error term, denoted by e_I , e_Φ , e_E , and e_D , respectively. We use these error terms to compute an overall error e_T which is minimised in order to obtain the object pose. As the individual error terms are of different orders of magnitude, we introduce the weight factors β_I , β_Φ , β_E , and β_D to appropriately take into account the individual terms in the total error e_T according to

$$e_T = \beta_I e_I + \beta_\Phi e_\Phi + \beta_E e_E + \beta_D e_D. \quad (4.39)$$

The individual error terms are not independent of each other, such that they have to be minimised simultaneously via minimisation of the total error e_T . This may become a fairly intricate nonlinear optimisation problem. The value of each weight factor is chosen inversely proportional to the typical relative measurement error, respectively. However, we found that the influence on the observed intensity, polarisation, edge, and depth cues is different for small variations of each pose parameter

Table 4.5 Influence of small changes of the pose parameters on the observed photopolarimetric, geometric, and depth cues.

	Intensity, polarisation	Edges	Depth
Rotation angles	strong	weak	weak
Lateral translation (x,y)	weak	strong	weak
Translation in z	weak	weak	strong

(cf. Table 4.5). For example, a slight lateral translation has a strong influence on the edges in the image but may leave the observed intensity and polarisation angle largely unchanged. On the other hand, under certain viewing conditions, rotations around small angles are hardly visible in the edge image while having a significant effect on the observed intensity or polarisation behaviour.

For minimisation of the overall error e_T we use an iterative gradient descent approach. We have chosen this algorithm because of its stable convergence behaviour, but other optimisation methods are possible. Since it is impossible to calculate analytically the derivatives of the total error term with respect to the pose parameters as the error term is computed based on rendered images, the gradient is evaluated numerically. If a certain cue does not provide useful information (which may e.g. be the case for polarisation data when the surface material only weakly polarises the reflected light, or for edges in the presence of cluttered background), this cue can be neglected in the optimisation procedure by setting the corresponding weight factor in Eq. (4.39) to zero. It is shown experimentally in this section and Section 5.1 that pose estimation remains possible when relying on merely two or three different cues.

Our framework requires a-priori information about the object pose for initialisation of the nonlinear optimisation routine, such that it is especially useful for the purpose of pose refinement. In comparison, the template matching based approach by von Bank et al. (2003) yields five pose parameters without a-priori knowledge about them, while the sixth parameter, the distance to the object, is assumed to be exactly known. In the addressed application domain of industrial quality inspection, a-priori information about the pose is available from the CAD data of the part itself and the workpiece to which it is attached. Here it is not necessary to detect the part in an arbitrary pose but to measure small differences between the true pose parameters and those desired according to the CAD data. Hence, when applied in the context of industrial quality inspection, our method should be initialised with the pose given by the CAD data, and depending on the tolerances stored in the CAD data, a production fault is indicated when the deviation of one or several pose parameters exceeds the tolerance value. The experimental evaluation presented in the next section shows that our framework is able to detect small differences between the true and the desired object pose.

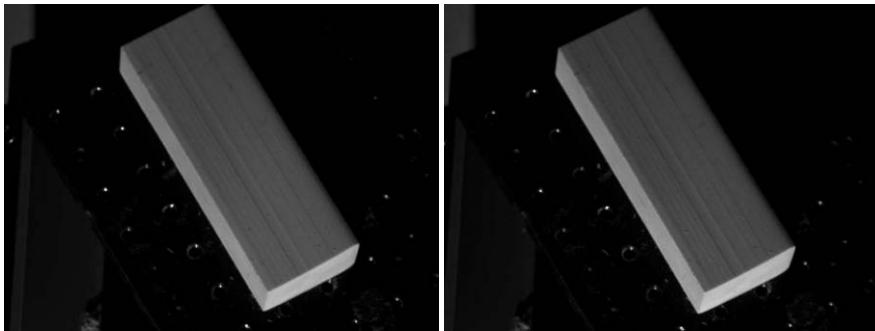


Fig. 4.27 Input intensity images for pose 1 (left) and pose 2 (right) of the rubber (cf. Table 4.6).

4.5.1.5 Experimental Evaluation Based on a Simple Real-world Object

For a first evaluation of the performance of the presented approach we estimated the pose of a simple cuboid-shaped real-world object (a rubber) and compared the results to the independently derived ground truth. The images were taken with a Baumer industrial CCD camera of 1032×776 pixels image size, equipped with a $f = 25$ mm lens. The approximate distance to the object was 0.5 m. The coordinate system was chosen such that the x and y axes correspond to the horizontal and vertical image axis, respectively, while the z axis is parallel to the optical axis. The scene was illuminated with a LED point light source located at a known position. The algorithm was initialised with four different poses, differing by several degrees in the rotation angles and a few millimetres in translation. As the result of pose estimation we adopted the minimisation run yielding the lowest residual error according to Eq. (4.39).

The reflectance function R_I was determined with a goniometer by estimating the parameters according to Eq. (4.22). At the same time we found that the polarisation degree of the light reflected from the surface is so small that it cannot be reliably determined. Hence, the input data for pose estimation are limited to intensity, edges, and depth.

For our evaluation, we attached the rubber with its lateral surface to the goniometer table and oriented it in two different poses relative to the camera. The angular difference between the two poses is only a few degrees (cf. Fig. 4.27). For the determination of the ground truth, we replaced the rubber for each pose by a chequerboard of known geometry. The chequerboard was attached to the goniometer table, and its pose was estimated using the rig finder algorithm described by Krüger et al. (2004), which is based on a bundle adjustment approach for camera calibration purposes. Due to the simple cuboid shape of the rubber the chequerboard pattern could be aligned at high accuracy into the same direction as the lateral surfaces of the rubber, such that the chequerboard pose could be assumed to be identical with the pose of the rubber.

Table 4.6 Estimated pose and ground truth (GT) for the rubber example.

Parameter	Pose 1	GT 1	Pose 2	GT 2
roll ($^{\circ}$)	13.3	13.5	16.7	16.3
pitch ($^{\circ}$)	-18.2	-18.9	-18.6	-19.7
yaw ($^{\circ}$)	59.4	58.6	59.2	58.5
t_x (mm)	-3.6	-3.2	-2.8	-2.5
t_y (mm)	2.3	2.3	1.3	1.7
t_z (mm)	451.5	454.3	457.5	453.9

The results of this experiment are shown in Table 4.6. The deviations between the measured and the true pose parameters are only a few tenths of a degree for the rotation angles and a few tenths of a millimetre for the lateral translations. The translation in z is determined at an accuracy of about 4 mm (which is about an order of magnitude lower than the lateral accuracy) or 1 percent. This is a reasonable result, given that only monocular image data are available.

4.5.1.6 Discussion

The three-dimensional pose estimation approach described in this section is based on photometric, polarimetric, edge, and defocus cues. A correspondingly defined error function is minimised by comparing the observed data to their rendered counterparts, where an accurate rendering of intensity and polarisation images is performed based on the material-specific reflectance functions determined with a goniometer. If a certain cue cannot be reliably measured or does not yield useful information, it can be neglected in the optimisation procedure.

A pose estimation accuracy comparable to the one obtained in the simple rubber example is achieved for more difficult objects in the context of industrial quality inspection as described in Section 5.1. It turns out that this accuracy is comparable to or higher than that of the monocular template matching approach by von Bank et al. (2003) (cf. Section 1.6.1), which exclusively relies on edge information and estimates only five degrees of freedom. The depth from defocus method has proven to be a useful instrument for the estimation of object depth in the close range at an accuracy of about 1 percent. Beyond depth from defocus, the described pose estimation framework is open for depth data independently obtained e.g. by active range measurement.

4.5.2 Contour-based Pose Estimation Using Depth from Defocus

4.5.2.1 Integration of Depth from Defocus into the CCD Algorithm

In the context of three-dimensional pose estimation of rigid objects, the contracting curve density (CCD) algorithm as described in Section 1.6.2.3 can be favourably

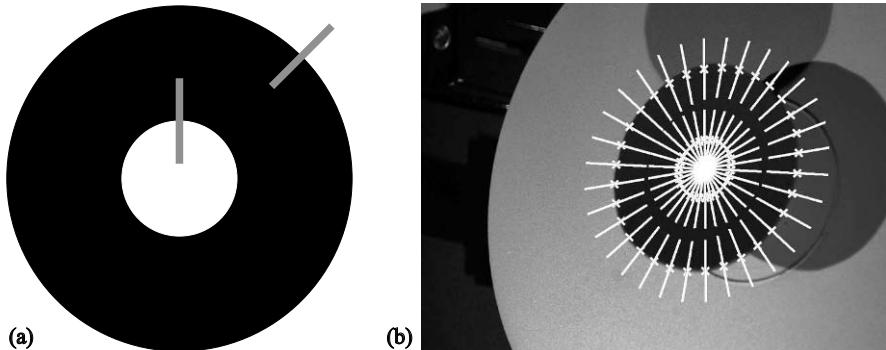


Fig. 4.28 (a) Model of the planar ring-shaped object used for the evaluation of the combined CCD and depth from defocus technique. Two normals along which grey value statistics and defocus are estimated are indicated as grey lines. (b) Typical pose estimation result. The reprojected object contour is marked by crosses, while the normals are denoted by lines.

used to determine the object boundary based on a three-dimensional contour model of the object. According to the CCD approach, a large number of intensity profiles is extracted along the estimated object boundary in the direction orthogonal to it. For all normals the grey value statistics inside and outside the contour are computed, denoted by $S(\mathbf{m}_T, \Sigma_T)$ with \mathbf{m}_T and Σ_T as the average and covariance of the model parameters T .

Additionally, the amount of defocus is estimated for each profile by fitting a sigmoid edge model of the form $I(u) = a \tanh[\zeta(u - u_0)] + b$ to the grey values. It is shown in Section 3.2.3 that the value of ζ is directly related to the PSF radius σ by $\zeta = (\sqrt{2/\pi})/\sigma$ (cf. Section 1.4.8 for different representations of an ideal edge blurred by a Gaussian PSF). The distance z is determined by a calibration procedure as outlined in Section 3.2.3.

To integrate depth from defocus information into the CCD algorithm, Eq. (1.142) expressing the a posteriori probability of the model parameters is extended by a term $p(\zeta|T)$ denoting the probability of the observed defocus values, given the model parameters T . The a posteriori probability of the model parameters T , given the observed image grey values denoted by I and the observed defocus values denoted by ζ , then becomes

$$p(T|I, \zeta) = p(I|T) \cdot p(\zeta|T) \cdot p(T). \quad (4.40)$$

Here, depth values inferred from the model parameters T for points on the object contour are translated into values for ζ based on the previously determined depth-defocus function. The probability $p(\zeta|T)$ is assumed to be Gaussian, and the defocus values are assumed to be independent of each other along the contour. This leads to the expression

$$p(\zeta|\mathbf{T}) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_{\zeta_i}} \exp\left(-\frac{(\hat{\zeta}_i - \zeta_i)^2}{2\sigma_{\zeta_i}^2}\right) \quad (4.41)$$

where i denotes the index of the normal, ζ_i the measured defocus for normal i , and $\hat{\zeta}_i$ the defocus inferred from the model parameters \mathbf{T} . The value of σ_{ζ_i} corresponds to the uncertainty of the measured value of ζ_i and can usually be assumed to be identical for all normals. As in the original form of the CCD algorithm, it is numerically favourable to perform the optimisation based on the log-likelihood

$$X_{\text{CCD+DFD}} = -2 \ln [p(I|S(\mathbf{m}_T, \Sigma_T)) \cdot p(\mathbf{T}|\hat{\mathbf{m}}_T, \hat{\Sigma}_T) \cdot p(\zeta|\mathbf{T})]. \quad (4.42)$$

Eqs. (4.40) and (4.42) provide a general Bayesian formalism according to which further information such as polarisation data (Krauß, 2006) or independently measured depth data can be integrated into the CCD algorithm.

4.5.2.2 Experimental Evaluation Based on a Simple Object

The ring-shaped planar object regarded for an evaluation of the combined CCD and depth from defocus approach is shown in Fig. 4.28a. This object is attached to the goniometer such that for the goniometer angles $\gamma_1 = \gamma_2 = 0^\circ$ the surface normal of the planar object is parallel to the optical axis of the camera. For image acquisition, we employed a Baumer CCD camera of 1032×768 pixels image size equipped with a lens of $f = 12$ mm, situated at a distance of 480 mm to the object centre. The resolution of the images is $186 \mu\text{m}$ per pixel, and the horizontal field of view amounts to 22.6° . An example image with a typical pose estimation result is shown in Fig. 4.28b.

In orthographic projection, the image of the contour of the circular object shown in Fig. 4.28a becomes an ellipse with a major axis d and a minor axis $d \cos \delta$, where δ is the angle between the surface normal of the object and the optical axis. The image of the contour is then symmetric with respect to the angle δ , i.e. a monocular pose estimation cannot recover the direction under which the object is tilted with respect to the optical axis. This symmetry is still approximately valid for perspective projection with a moderate field of view as encountered in our experiments. Furthermore, it is difficult to recover the value of δ at reasonable accuracy based on the image geometry alone if the angle δ is small, since $\cos \delta$ then hardly changes with varying value of δ .

In our experimental evaluation we compared the pose estimation result of the CCD algorithm to that obtained with the combined CCD and depth from defocus approach. The goniometer angles γ_1 and γ_2 provide a highly accurate ground truth for the orientation of the surface normal of the planar object. We examine the dependence of the pose error on the deviation between the true object orientation and the pose angles used as initial values for the optimisation procedure. The pose angles were defined in the same coordinate system as the goniometer angles γ_1 and γ_2 .

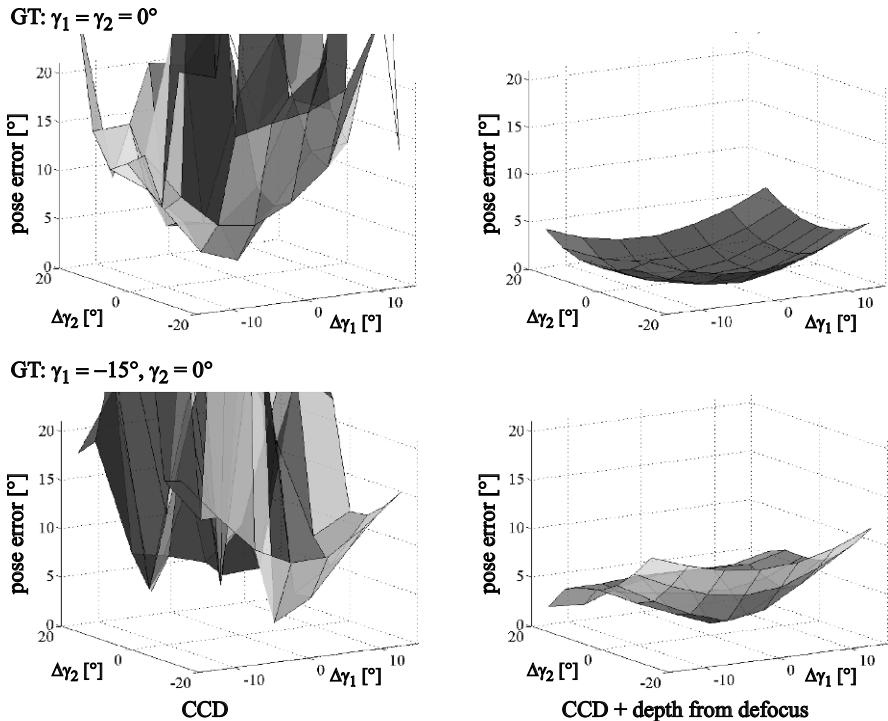


Fig. 4.29 Dependence of the pose error on the initial values of the optimisation procedure for two different object poses. The result of the CCD algorithm is shown on the left, the result of the combined CCD and depth from defocus approach on the right, respectively.

The pose error as indicated in Fig. 4.29 is defined as the angle between the surface normal of the object and the optical axis.

For the goniometer angles $\gamma_1 = \gamma_2 = 0^\circ$ (upper row in Fig. 4.29), where the normal to the object plane is parallel to the optical axis, the CCD algorithm alone yields small pose errors only for initial poses that deviate by no more than a few degrees from the true pose. For very inaccurate initial values the CCD algorithm diverges. Integrating depth from defocus information into the CCD algorithm yields much more accurate pose results also for initial poses that deviate by more than 20° from the true pose, and no divergence of the combined approach is observed. A similar result is obtained when the object is tilted by 15° to the optical axis, corresponding to the goniometer angles $\gamma_1 = -15^\circ$ and $\gamma_2 = 0^\circ$ (bottom row in Fig. 4.29). The CCD algorithm alone merely achieves an inaccurate estimation of the pose of the object with a pose error of more than 5° even when the optimisation procedure is initialised with the true pose. For most configurations where the initial angles deviate by more than 10° from the ground truth the CCD algorithm diverges. The combined CCD and depth from defocus method yields pose errors of only a few degrees for a broad range of initial values for γ_2 as long as γ_1 is initialised with the correct sign. Hence,

for this geometrically difficult monocular pose estimation problem the integration of depth from defocus data leads to a significant improvement of the accuracy and convergence behaviour of the CCD algorithm.

4.5.2.3 Discussion

The monocular three-dimensional pose estimation algorithm described in this section is based on a Bayesian framework for the integration of independently obtained additional information into the CCD algorithm. As an example, we have examined the combination of the CCD algorithm with depth from defocus data extracted from the contour normals. Our experimental evaluation, which is based on a fairly simple planar object, illustrates that the convergence behaviour of the combined approach is strongly improved, compared to the CCD algorithm alone. Under more realistic conditions as encountered e.g. in the domain of industrial quality inspection, the utilised depth from defocus technique is not appropriate as object contours tend to deviate from the ideal edge model due to specular reflections or shading effects. Hence, the experiments described in this section are intended to illustrate the principles of the combined approach and its behaviour in the presence of objects of limited complexity.

Chapter 5

Applications to Industrial Quality Inspection

Industrial quality inspection is an important application domain of three-dimensional computer vision methods. Traditional vision-based industrial quality inspection systems primarily rely on two-dimensional detection and pose estimation algorithms e.g. relying on the detection of point and line features, the extraction of blob features from binarised images, or two-dimensional grey value correlation techniques (Demant, 1999). More advanced vision-based quality inspection systems employ three-dimensional methods in order to detect production faults more reliably and robustly. In this section we regard applications in the automobile industry of the methods for three-dimensional pose estimation of rigid and articulated objects described in Section 1.6.

A typical area of interest is checking for completeness of a set of small parts attached to a large workpiece, such as plugs, cables, screws, and covers mounted on a car engine. A different task is the inspection of the position and orientation of parts, e.g. for checking if they are correctly mounted but also for grasping and transporting them with an industrial robot.

Section 5.1 regards the three-dimensional pose estimation of rigid parts in the context of quality inspection of car engine components. The approach of object detection by pose estimation without a-priori knowledge about the object pose is analysed in Section 5.1.1, while the technique of pose refinement based on an appropriate initial pose is regarded in Section 5.1.2. Here we concentrate on the methods introduced by von Bank et al. (2003) and by Barrois and Wöhler (2007), respectively (cf. Sections 1.6.1 and 4.5.1).

The three-dimensional pose estimation of tubes and cables in the scenario of car engine production is analysed in Section 5.2 based on the method by d'Angelo et al. (2004) outlined in Section 1.6.2. For each scenario we compare our evaluation results to results reported in the literature for systems performing similar inspection tasks. Section 5.3 describes applications of the integrated approaches introduced in Sections 4.1–4.4 (Wöhler and Hafezi, 2005; d'Angelo and Wöhler, 2005a,b,c, 2006, 2008) to the three-dimensional reconstruction of rough metallic surfaces of automotive parts.

Table 5.1 Properties of the three oil cap template hierarchies (ranges of pose angles ρ , ε , λ , and grid sizes in degrees). Hierarchy 1 consists of 4550 templates, hierarchies 2 and 3 of 1331 templates, respectively.

Hierarchy	ρ range	$\Delta\rho$	ε range	$\Delta\varepsilon$	λ range	$\Delta\lambda$
1	$0^\circ \dots 180^\circ$	2°	$18^\circ \dots 72^\circ$	6°	$-12^\circ \dots +12^\circ$	6°
2	$0^\circ \dots 20^\circ$	2°	$30^\circ \dots 50^\circ$	2°	$-10^\circ \dots +10^\circ$	2°
3	same as 2, but without writing		modelled			

5.1 Inspection of Rigid Parts

The first typical quality inspection scenario involving the three-dimensional pose estimation of rigid parts is detection by pose estimation (cf. Section 5.1.1), corresponding to a simultaneous detection of the presence and estimation of the pose of an object of known three-dimensional geometry. In the second scenario, pose refinement (cf. Section 5.1.2), a reasonably accurate initial pose of the object is required which is then refined further.

Many applications of pose estimation methods for quality inspection purposes impose severe constraints on the hardware to be used with respect to robustness and easy maintenance. Hence, it is often difficult or even impossible to utilise stereo camera systems since they have to be recalibrated regularly, especially when the sensor unit is mounted on an industrial robot. In this section we therefore describe applications of the monocular pose estimation methods by von Bank et al. (2003) and by Barrois and Wöhler (2007) (cf. Sections 1.6.1 and 4.5.1) in the automobile production environment.

5.1.1 Object Detection by Pose Estimation

The inspection task addressed in this section is to detect the presence and to estimate the three-dimensional pose of the oil cap shown in Fig. 5.1 (von Bank et al., 2003). To generate real-world images with well-defined ground truth poses, we made use of a calibrated robot system. The accuracy of calibration with respect to the world coordinate system is about 0.1° with respect to camera orientation and 0.1 mm with respect to camera position. As the engine itself is not part of the robot system, the relation between world coordinate system and engine coordinate system has to be established separately, which reduces the accuracies stated above by about an order of magnitude.

First, the difference between the measured and the true pose of the correctly assembled oil cap is determined depending on the camera viewpoint and the illumination conditions. The scene is illuminated by a cylindric lamp around the camera lens (confocal illumination) and a halogene spot. The background of the scene may be fairly cluttered. The distance to the object amounts to 200 mm and is assumed to be known, such that only five degrees of freedom are to be estimated. For this

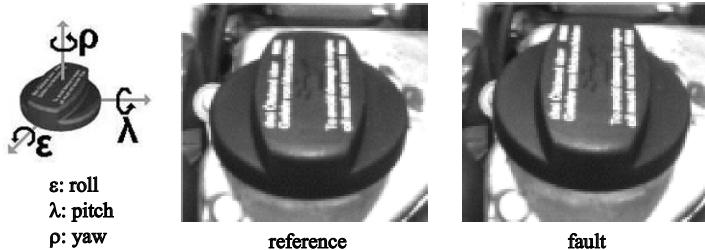


Fig. 5.1 Left: Definition of the roll, pitch, and yaw angles ϵ , λ , and ρ . Centre and right: Reference pose of the oil cap and typical fault.

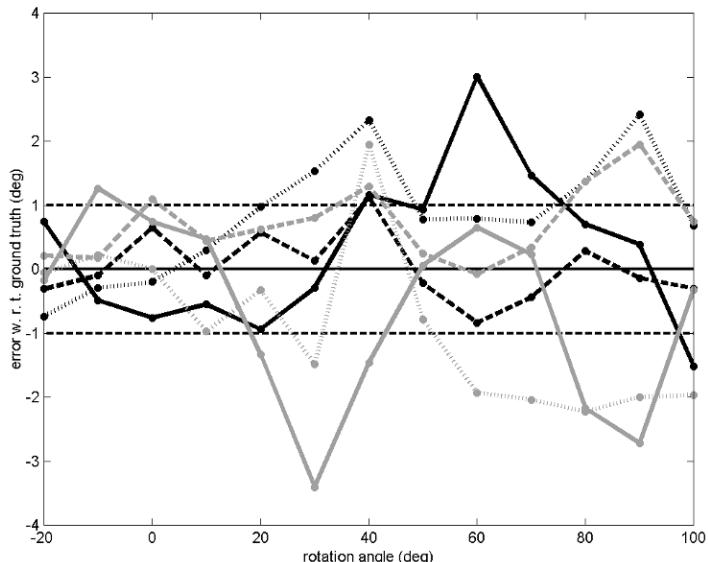


Fig. 5.2 Deviations of ϵ (solid lines), ρ (dashed lines), and λ (dotted lines) from their ground truth values. Illumination is with cylindric lamp only (black lines) and with both cylindric and halogene lamp (gray lines). The true roll angle is constantly set to $\epsilon = 70^\circ$.

examination we use template hierarchy 1 (cf. Table 5.1). The angle ϵ denotes the roll angle, λ the pitch angle, and ρ the yaw angle. For camera viewpoints with $-10^\circ \leq \rho \leq 10^\circ$ and $50^\circ \leq \epsilon \leq 60^\circ$, the measured pose lies within the calibration accuracy interval of 1° for all three angles. Fig. 5.2 shows that for $\epsilon = 70^\circ$, this is even true for $-20^\circ \leq \rho \leq 20^\circ$. This implies that from a correspondingly chosen viewpoint, the algorithm is highly sensitive with respect to deviations from the reference pose. Hence, it is possible to determine the pose of the oil cap to an accuracy of about 1° . For comparison, the state-of-the-art technique for pose estimation of industrial parts presented by Bachler et al. (1999) yields pose errors of about 3° even when the object is put on a uniform background. Significantly changing the illumination conditions by switching off the halogene lamp does not affect the pose

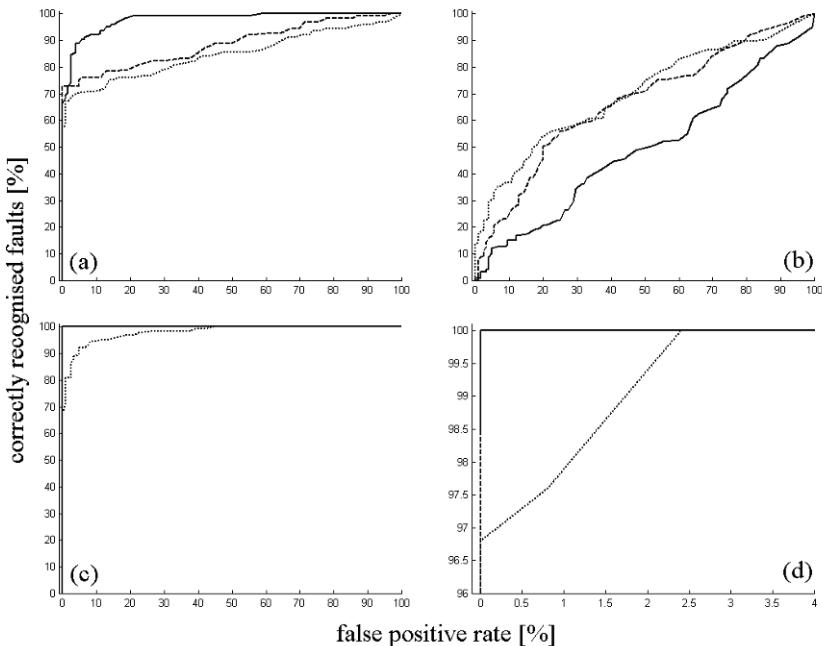


Fig. 5.3 ROC curves of the inspection system, based on the measured values of (a) the roll angle ε , (b) the yaw angle ρ , (c) the pitch angle λ , (d) all three pose angles. Solid, dashed, and dotted lines denote three different camera viewpoints. Note that the axis scaling in (d) is different from that in (a)–(c).

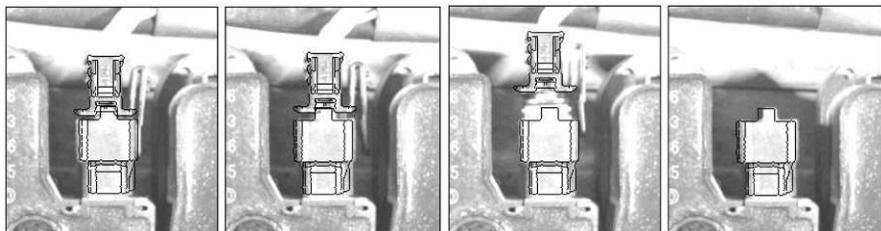


Fig. 5.4 Ignition plug inspection: Reference configuration (left) and three fault configurations with the corresponding matching results, using two templates. Image scale is 0.2 mm per pixel.

estimation results. The computation time of the system amounts to about 200 ms on a Pentium IV 2.4 GHz processor.

The described system relies on the template-based approach by von Bank et al. (2003) described in Section 1.6.1. As it aims for distinguishing incorrect from correct poses, i. e. performing a corresponding classification of the inspected object, the rate of correctly recognized faults (the rate of incorrectly assembled oil caps which are recognized as such by the inspection system) is determined versus the rate of correctly assembled objects erroneously classified as

incorrectly assembled (false positive rate). This representation of the system behaviour is termed receiver operating characteristics (ROC) curve. We determined the recognition behaviour of the system for three different camera viewpoints. Here we concentrate on a typical fault situation showing angle differences $\Delta\rho = 0^\circ$, $\Delta\epsilon = 2.5^\circ$, $\Delta\lambda = -3.5^\circ$ with respect to the reference pose. In the production environment, the engine and thus the attached oil cap is positioned with a tolerance of about 1 cm with respect to the camera. This positional variation was simulated by acquiring 125 different images of each examined fault situation from 125 camera positions inside a cube of 1 cm size which are equally spaced by 2.5 mm in each coordinate direction. This offset is taken into account appropriately in the pose estimation based on the measured position of the oil cap in the image. As a first step, a fault is assigned based on each of the three angles separately if the corresponding angle deviates from the reference value by more than a given threshold. By varying this threshold, a ROC curve is generated for each angle separately as shown in Fig. 5.3a–c. We then generate a combined ROC curve by assuming that the oil cap is assembled incorrectly if the deviation of at least one of the pose angles is larger than the corresponding threshold. These thresholds are then adjusted such that the area under the ROC curve becomes maximum. This combination generally yields a ROC curve showing very few misclassifications on the acquired test set, as illustrated in Fig. 5.3d. Both with template hierarchy 1, which covers a wide range of pose angles with a large grid size, and with hierarchy 2, covering a region on the viewing sphere close to the reference view with a small grid size (cf. Table 5.1), very high recognition rates close to 100 percent are achieved. With hierarchy 3, which is identical to hierarchy 2 except that the writing on top of the oil cap has been omitted, the performance decreases, but not significantly: At a false positive rate of zero, still a rate of correctly recognized faults of 98.4 percent is achieved.

In the second scenario, dealing with the inspection of an ignition plug, we regard three fault configurations in addition to the reference configuration: The clip is not fixed, the plug is loose, and the plug is missing (Fig. 5.4). The connector and the plug are modelled as two separate objects such that the offset of the plug in vertical direction can be used to distinguish fault configurations from the reference configuration. The matching results in Fig. 5.4 show that the vertical position of the plug relative to the connector can be determined at an accuracy of about 0.5 mm, which is sufficient to faithfully distinguish correctly from incorrectly assembled ignition plugs.

Lacey et al. (2002) provide accuracy values for the TINA system, which performs a three-dimensional pose estimation of objects in the context of object manipulation by a robot arm. They report a translational accuracy of better than 5 mm when estimating all six degrees of freedom based on a pair of stereo images of the scene. No values for the accuracy of the estimated pose angles are given. The oil cap described above is also used by Krüger (2007) as a test object to evaluate the performance of a multiocular template matching technique (which is an extension of the method described in this section), as well as a feature pose map algorithm (also known as generalised Hough transform) and the gradient sign table approach explained in Section 1.6.2. The pixel resolution of the regarded images is comparable to that

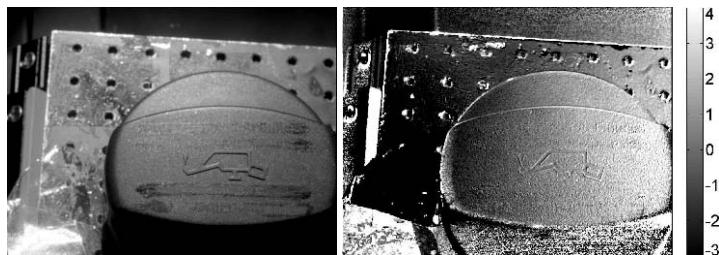


Fig. 5.5 Example of a high-dynamic range intensity image (left, grey values are displayed logarithmically) and a polarisation angle image (right, grey value map is scaled in degrees).

of the images examined in this section. The three algorithms are evaluated using the trinocular Digidlops camera system. For all three methods, Krüger (2007) obtains accuracies of about 2° for the rotation angles and between 3 and 7 mm for the depth. In contrast, the depth is not estimated but assumed to be known when using the monocular technique by von Bank et al. (2003) examined above. No values are given by Krüger (2007) for the translational accuracy in the directions parallel to the image plane. Hence, the accuracies of the estimated pose angles are comparable for the monocular template matching approach that estimates five degrees of freedom and the extended, trinocular template matching technique that determines all six degrees of freedom.

5.1.2 Pose Refinement

In this section we examine the application of the appearance-based approach by Barrois and Wöhler (2007) described in Section 4.5.1 to three-dimensional pose estimation of automotive parts. This method combines monocular photopolarimetric, edge, and defocus information. As a first example, we again regard the oil cap. The experimental setup is the same as in Section 4.5.1 (Baumer industrial CCD camera, 1032×776 pixels image size, $f = 25$ mm, object distance around 0.5 m). The system performs pose refinement with four initial poses differing by several degrees in the rotation angles and a few millimetres in translation. The minimisation run yielding the lowest residual error is adopted as the pose estimation result. Since due to its complex shape this object cannot be attached to the goniometer table in a reproducible manner, we determined the ground truth pose in this experiment based on a stereoscopic bundle adjustment tool which exploits manually established point correspondences between a calibrated stereo image pair and the CAD model of the object. As in the first experiment, the goniometer was used to determine the intensity and polarisation angle reflectance functions R_I and R_ϕ . The light reflected by the plastic surface of the oil cap is partially polarised by 10–20 percent, such that the polarisation angle can be used in our pose estimation framework in addition to intensity, edges, and depth. The intensity images of the two regarded poses are

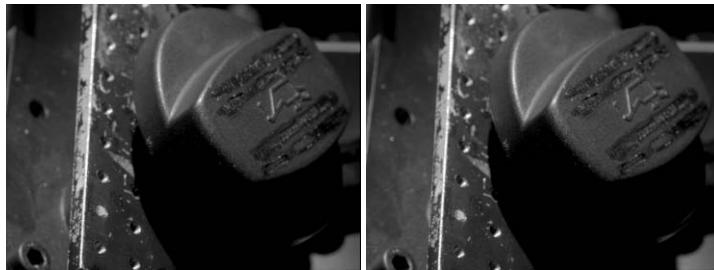


Fig. 5.6 Intensity images of the oil cap for pose 1 (left) and pose 2 (right). Grey values are displayed in logarithmic scale.

Table 5.2 Estimated pose and ground truth (GT) for the oil cap example.

Parameter	Pose 1	GT 1	Pose 2	GT 2
roll (°)	233.2	234.5	230.7	232.1
pitch (°)	1.3	2.3	0.9	2.4
yaw (°)	57.3	55.2	56.8	56.0
t_x (mm)	14.7	14.7	15.0	14.8
t_y (mm)	2.1	2.8	2.0	2.5
t_z (mm)	512.9	509.2	512.7	509.2

shown in Fig. 5.6, illustrating that at some places especially near the right image border the edges are not well-defined, such that the pose estimation algorithm to a large extent has to rely on intensity and polarisation information. The comparison to the ground truth is shown in Table 5.2, demonstrating that the object pose can be determined with an accuracy of 1° – 2° for the rotation angles, some tenths of a millimetre for the lateral translations, and several millimetres or about 1 percent for the object distance. We observed that small deviations of the rotation angles can be compensated by correspondingly adjusting the albedo factor ρ , leading to a lower accuracy of the rotation angles, compared to the rubber example. Due to the somewhat ill-defined edges the pose estimation fails when only edge information is used, as no convergence of the minimisation routine is achieved.

For the oil cap example, it is possible to directly compare the results of the method by Barrois and Wöhler (2008) to those of the monocular edge-based template matching method proposed by von Bank et al. (2003), since in that work the same object and the same CAD model are regarded. The deviation of the rotation angles estimated by von Bank et al. (2003) from the corresponding ground truth is typically around 1° – 2° but may also become larger than 3° . In contrast to the method described in this study, it is assumed by von Bank et al. (2003) that the distance to the object is known, i.e. only five rather than six degrees of freedom are estimated by von Bank et al. (2003). On the other hand, that method does not require a-priori information about the object pose.

In a further experiment we regard another automotive part, a door hinge, consisting of cast metal with a rough and strongly specular surface (cf. Fig. 5.7). For the pose we have chosen for our experiment, the light from the point light source

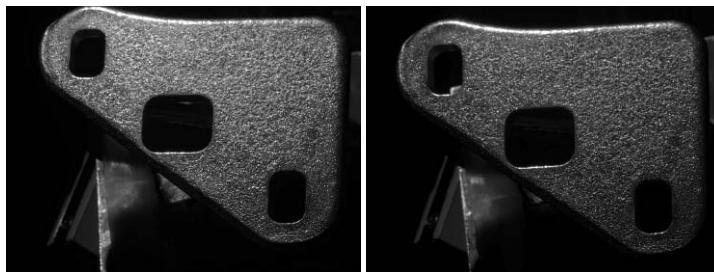


Fig. 5.7 Intensity images of the door hinge for pose 1 (left) and pose 2 (right). Greylevels are displayed in logarithmic scale.



Fig. 5.8 Distance-transformed edge image of the door hinge.

Table 5.3 Estimated pose differences and ground truth for the door hinge example.

Parameter difference	Result	GT
$\Delta\epsilon$ ($^{\circ}$) (roll)	4.15	4.23
$\Delta\lambda$ ($^{\circ}$) (pitch)	2.06	1.69
$\Delta\rho$ ($^{\circ}$) (yaw)	0.22	0.58
Δt_x (mm)	0.71	0.06
Δt_y (mm)	1.88	2.33
Δt_z (mm)	3.82	0.16

is specularly reflected into the camera. The Canny edge detector yields a very large number of edges (cf. Fig. 5.8), thus providing no reliable information about the object pose. As a consequence, our approach fails when we attempt to perform a pose estimation of the hinge based on the extracted edge information. The surface of the hinge does not perceptibly polarise the reflected light. Hence, we only use intensity and defocus data as input information for our algorithm. The obtained results illustrate that our algorithm also works in the absence of some of the input cues and that it is suitable for pose estimation of objects with a strongly specular surface.

In this experiment, the chequerboard method could not be used for determining the ground truth since the hinge could not be attached to the goniometer in a reproducible manner, such that it was not possible to place it in a known position relative to the chequerboard and the goniometer. Similarly, the bundle adjustment tool based on manually established point correspondences could not be used since unlike the oil cap, the hinge does not display well-defined corner points. Hence, we compare the estimated poses to the difference imposed by the two chosen goniometer settings, which are given at high accuracy. The estimated pose differences and the corresponding ground truth values are shown in Table 5.3. Although not all geometric, photometric, and depth cues are available, the obtained results are comparable to or better than those obtained in the previous experiments (some tenths of a degree for the rotation angles, some tenths of a millimetre for the lateral translation, and some millimetres for the object distance).

The examined monocular pose estimation framework is based on photometric, polarimetric, edge, and defocus cues. The pose is obtained by minimising a correspondingly defined error function. The observed data are compared to their rendered counterparts, where an accurate rendering of intensity and polarisation images is performed based on the material-specific reflectance functions determined with a goniometer. If a certain cue cannot be reliably measured or does not yield useful information, it can be neglected in the optimisation procedure. Beyond depth from defocus, in principle our pose estimation framework is open for depth data obtained e.g. by active range measurement. The inferred pose refinement accuracy is comparable to or higher than that of the monocular template matching approach by von Bank et al. (2003) analysed in Section 5.1.1, which exclusively relies on edge information. This result is achieved despite the fact that our method additionally provides an estimate of the distance to the object, while the method by von Bank et al. (2003) assumes that the object distance is known.

The depth from defocus method has turned out to be a useful instrument for the estimation of object depth in the close range at an accuracy of about 1 percent. We have demonstrated the usefulness of our method under conditions typically encountered in industrial quality inspection scenarios such as the assembly of complex parts. Here, the desired pose of the whole workpiece or part of it is given by the CAD data and the inspection system has to detect small differences between the actual and the desired pose.

At this point it is interesting to compare the accuracy of our method with that achieved by other pose refinement approaches. In the monocular system by Socher (1997), objects are detected based on colour and edge features, and object poses are refined in a subsequent step by minimising an error function that measures the difference between the observed features and the reprojected model features. The image resolution is about 0.4 mm per pixel, the translational and rotational accuracies are 4 mm and 5°, respectively. The monocular pose refinement system by Kölzow (2002) combines information about the quality of local feature matches into a global error measure for different types of features. One regarded example object is the same oil cap as the one examined in this section. For this object, the pose refinement essentially relies on edges and point features. At an image resolu-

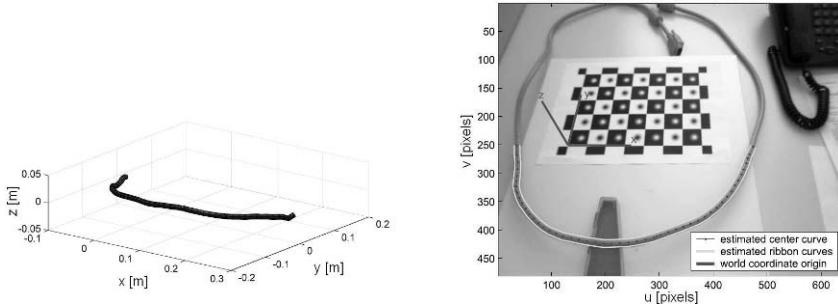


Fig. 5.9 Cable Scene. It is evident that the reconstructed cable is located above the table plane ($z = 0$). The world coordinate system has been defined by estimating the pose of the chequerboard.

tion of about 0.7 mm per pixel, the system achieves an average translational and rotational deviation between the measurements and the ground truth of 1 mm and 1° , respectively, with an uncertainty of a single measurement of 1 mm and 2° . However, it should be noted that the evaluation by Kölzow (2002) exclusively relies on synthetically rendered rather than real-world images. The method by Stößel (2007) is designed to estimate the pose parameters of articulated objects in monocular images. When applied to the oil cap, it yields a mean rotational error between 0.1° and 0.3° . Due to the stochastic nature of the employed optimisation algorithm, however, the result of an optimisation run is non-deterministic. Hence, the uncertainty of an individual pose estimation result amounts to $0.9\text{--}1.3^\circ$. The system by Hel-Or and Werman (1996) performs a pose refinement of articulated objects in stereo images based on a user-defined initial pose, relying on explicit correspondences between the image and an object model. They report a mean error of 1° for the rotational angles with an uncertainty of an individual measurement of 1° . Neither Hel-Or and Werman (1996) nor Stößel (2007) report accuracy values for the translational degrees of freedom. The system by Yoon et al. (2003) performs a pose estimation of industrial parts in real time in stereo images. The objects are located at a distance of 600–800 mm, a resolution of about 0.9 mm per pixel can be inferred from the presented example images. They report an accuracy of 1.5° and 2.4 mm for the rotational and translational degrees of freedom, respectively.

This comparison illustrates that our pose refinement approach based on multiple monocular cues is able to estimate the rotational degrees of freedom at an accuracy which is of the same order of magnitude as the mean error of the method proposed by Stößel (2007) even under viewing directions where geometric cues alone do not allow to recover the object pose reliably. It should be noted, however, that the uncertainty of an individual measurement by Stößel (2007) is nearly an order of magnitude higher than the mean error. Presumably, the high accuracy of the estimated rotational degrees of freedom achieved by our method is due to the integration of photopolarimetric information. The translational accuracy of our method in directions parallel to the image plane, which is largely determined by the available edge information, is comparable to or slightly higher than that observed for other pose

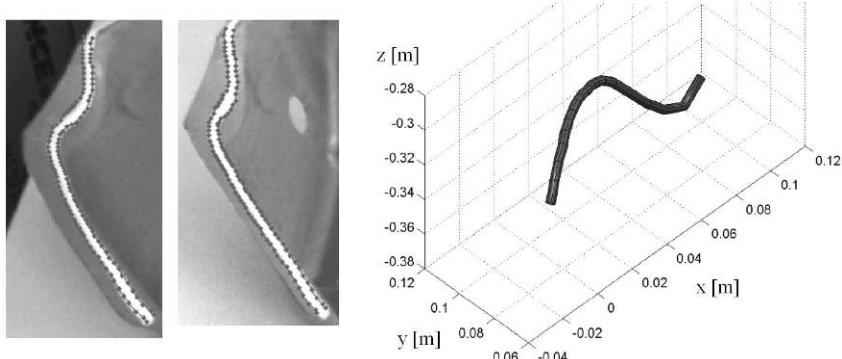


Fig. 5.10 Three-dimensional reconstruction of a glue line on the non-planar surface of a car body part (left), resulting three-dimensional ziplock ribbon snake (right).

refinement systems, which all rely to a considerable extent on edge information. The accuracy of the values we obtain for the object depth with our monocular depth from defocus technique comes close to that of the system by Yoon et al. (2003), which is based on the evaluation of stereo image pairs.

5.2 Inspection of Non-rigid Parts

In this section the three-dimensional active contour algorithm described in Section 1.6.2.2 is applied to the reconstruction of a cable, using images acquired with a Digiclops trinocular camera system at a distance to the object of about 1 m as shown in Fig. 5.9. A three-dimensional ziplock ribbon snake has been used in these examples. The approximate ribbon radius was used as model information, and constraints for the upper and lower bound of the ribbon width were applied. The segmented contour part is displayed from a slightly different viewing angle to show the spatial segmentation. The accuracy of reconstruction compared to ground truth obtained with a calliper gauge is about 1 mm, corresponding to 1.5 pixels, in this example.

A different inspection scenario is the three-dimensional reconstruction of a glue line on the non-planar free-form surface of a car body part as shown in Fig. 5.10. Although the object is not flexible, a pose estimation method for rigid objects cannot be applied due to the lack of an accurate geometry model. Hence, the multiocular ziplock ribbon snake algorithm is an adequate approach to reconstruct the three-dimensional shape of the glue line. Again, the images were acquired with the Digiclops system, and the object was situated at a distance of about 0.7 m to the camera. The accuracy of the reconstruction result compared to ground truth obtained with a calliper gauge amounts to 1 mm, corresponding to 1.5 pixels. A similar scenario in which this technique has been successfully applied is the inspection of injection tubes as shown in Fig. 5.11, where the inspection system checks if the tube has been

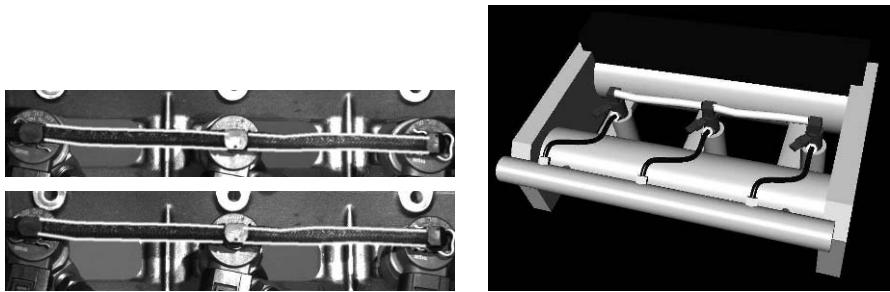


Fig. 5.11 Three-dimensional reconstruction of two injection tubes attached to an engine. Left: Pair of stereo images. The reprojected three-dimensional contours are indicated as white lines. Right: Reconstruction result (white), rendered into the CAD model of the corresponding part of the engine.

correctly mounted to the engine block. Pairs of stereo images are acquired by two calibrated Baumer CCD cameras with 1032×768 pixels image size located at a distance of 1 m to the object, equipped with lenses of focal length $f = 12$ mm. Our algorithm reliably detects a displacement of the tube trajectory of about 1 mm.

A scenario that goes beyond mere quality inspection is illustrated in Fig. 5.12 in a laboratory environment. A bar code is desired to be read by a monocular camera (camera 2 in Fig. 5.12a) but is partially occluded by a cable. The trajectory of the cable is reconstructed with the multiocular ziplock snake method (cf. Fig. 5.12b–c) using the Digiclops trinocular camera system (camera 1 in Fig. 5.12a). As soon as the three-dimensional pose estimation has been performed as indicated in Fig. 5.12c, the industrial robot grasps the cable and holds it aside as apparent in Fig. 5.12d. For this purpose, the robot needs to be calibrated with respect to the coordinate system of the trinocular camera. In the image of the monocular camera, the bar code is now unoccluded and can be read (cf. Fig. 5.12e). Finally, the robot places the cable back to its original position as shown in Fig. 5.12f. A localisation accuracy of 1 mm is required for mechanical reasons to enable the robot to grasp the cable, which is faithfully achieved by the multiocular ziplock snake approach.

In the application scenario of three-dimensional reconstruction of tubes, a simplified variant of the MOCCD algorithm is used by Krüger and Ellenrieder (2005). An extensive evaluation based on the full MOCCD approach according to Hahn et al. (2007) as described in Section 1.6.2 is provided by Krüger (2007). The three-dimensional reconstruction accuracy of the multiocular ziplock snake method is compared to that of the MOCCD and the gradient sign table approach (cf. Section 1.6.2). The regarded scenario is the so-called dangling cable problem, i.e. the three-dimensional position of the non-rigid object and the direction of its trajectory is known *a priori* for only one of its ends. In the application scenarios regarded in this section to evaluate the multiocular ziplock snake approach, this information is available for both ends of the objects.

Krüger (2007) arrives at the conclusion that multiocular ziplock snakes are more robust with respect to missing edges due to the error function involv-

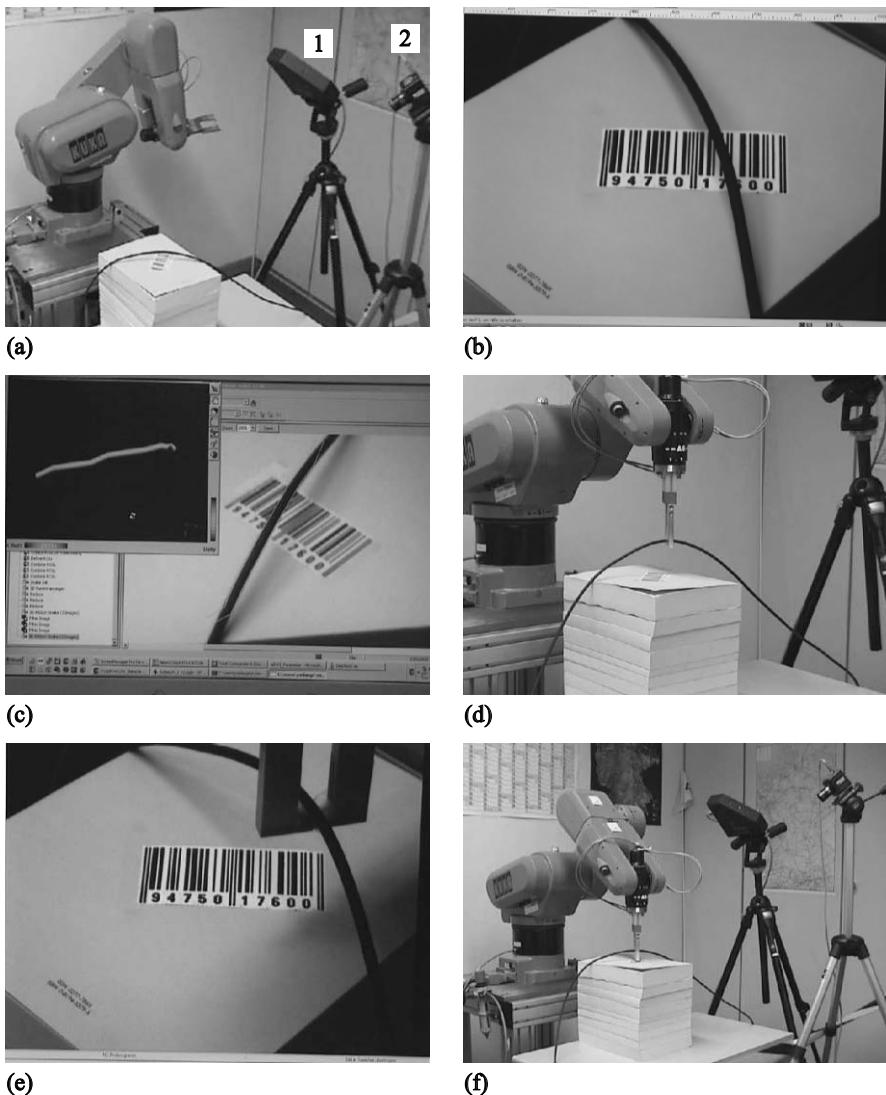


Fig. 5.12 Industrial robot grasping a cable. (a) The pose of the cable is estimated based on trinocular image data acquired with a Digiclops system (1). The image of the bar code is taken by a monocular camera (2). (b) The bar code cannot be read due to the cable partially occluding it. (c) The three-dimensional pose of the cable is estimated. (d) The robot grasps the cable and holds it aside. (e) Now the bar code can be read. (f) The robot places the cable back into its original position.

ing a blurred gradient image. Using the Digiclops trinocular camera system for image acquisition and setting the distance to the object to about 1 m, the multiocular ziplock snake and the MOCCD yield comparable average devia-

tions between the reconstructed three-dimensional trajectories and the ground truth of 1–3 mm as long as the contrast between object and background is high. For a low contrast or when shading effects are apparent at the edges of the object, the MOCCD may become more accurate than the multiocular ziplock snake by up to an order of magnitude, which is due to the fact that effectively the behaviour of the MOCCD error function is similar to that of an adaptive edge filter. On the other hand, the MOCCD algorithm only operates for objects larger than a certain minimum size in the image. Multiocular ziplock snakes are more suitable for objects below the minimal size required by the MOCCD or if only one edge of the object is visible. Furthermore, Krüger (2007) finds that the gradient sign tables approach is of similar accuracy as the multiocular ziplock snake and the MOCCD. Its advantages are its somewhat higher robustness with respect to edges of low contrast and in some cases its higher processing speed.

5.3 Inspection of Metallic Surfaces

This section describes applications of the previously described three-dimensional surface reconstruction methods to the inspection of metallic surfaces mainly of automotive parts. Where possible, the results are compared to independently derived ground truth values obtained by a laser profilometer or by tactile measurement.

Traditionally, photometric three-dimensional surface reconstruction techniques such as shape from shading are regarded as being unsuitable for industrial applications (Mühlmann, 2002). State-of-the-art commercial three-dimensional surface reconstruction systems for industrial quality inspection purposes are based on active scanning techniques such as projection of coded structured light. It is shown in this section, however, that especially for metallic surfaces, taking into account photometric image information during the reconstruction process under well-controlled conditions yields three-dimensional reconstruction results that are at least comparable to those obtained with active scanning devices while requiring significantly less complex and expensive instrumental configurations.

5.3.1 *Inspection Based on Integration of Shadow and Shading Features*

As a first (non-automotive) example, we regard the three-dimensional reconstruction of the lateral surface of a sharpener consisting of aluminium. Due to the perpendicular view and strongly oblique illumination, the specular components in the reflectance function (2.15) are negligible as $\cos \theta_r \ll 1$, such that Lambertian reflectance is a reasonable approximation to the true reflectance behaviour. The result of the combined shadow and shading approach according to Section 4.2, relying on one shadow and one shading image, is shown in Fig. 5.13. The surface dis-

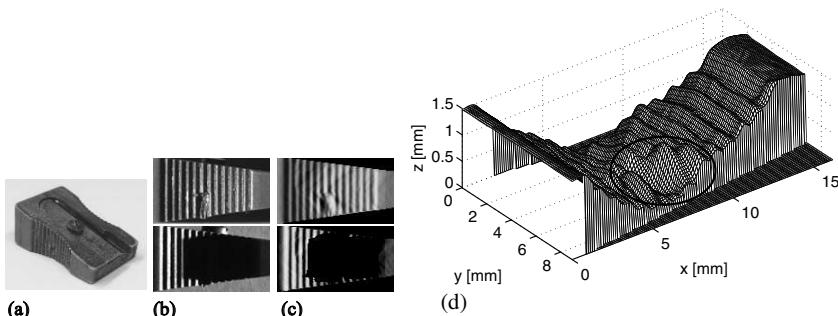


Fig. 5.13 Three-dimensional reconstruction of the lateral surface of a sharpener, based on the integration of shadow and shading features according to Section 4.2. (a) The sharpener. (b) Images used for three-dimensional reconstruction. (c) Images rendered based on the reconstruction result. (d) Three-dimensional surface reconstruction. A surface defect of 0.5 mm depth is marked by an ellipse.

plays a regularly striped structure and a small surface defect, which are well visible in the reconstruction result. The depth of the surface defect marked by an oval in Fig. 5.13d of approximately 0.5 mm corresponds within 0.1 mm to the value derived by tactile measurement. This example illustrates that even in the absence of accurate knowledge about the surface reflectance the combined shadow and shading approach yields a quantitatively fairly accurate three-dimensional reconstruction result.

5.3.2 Inspection of Surfaces with Non-uniform Albedo

An application in which the three-dimensional reconstruction approach has to cope with a non-uniform surface albedo is the inspection of uncoated steel sheet surfaces. The surface shown in Fig. 5.14 displays a dark line in the left half of the image, which is an acceptable surface property that may be due to small particles deposited on the surface during the production process, and a constriction in the right half of the image, which is a defective thinning of the sheet metal that may lead to mechanical instabilities.

The three-dimensional reconstruction of the surface is performed relying on the ratio-based intensity error term (2.50) and the ratio-based iterative update rule according to Eq. (2.51). The specular component of the surface reflectance was found to be weak, such that it could be neglected for the viewing and illumination geometry applied to acquire the images shown in Fig. 5.14. Under a perpendicular view on the surface, two images were acquired at phase angles of 77.9° and 59.7° . This configuration corresponds to elevation angles of the light source of 12.1° and 30.3° above the average surface plane, respectively. In the ratio image, the dark

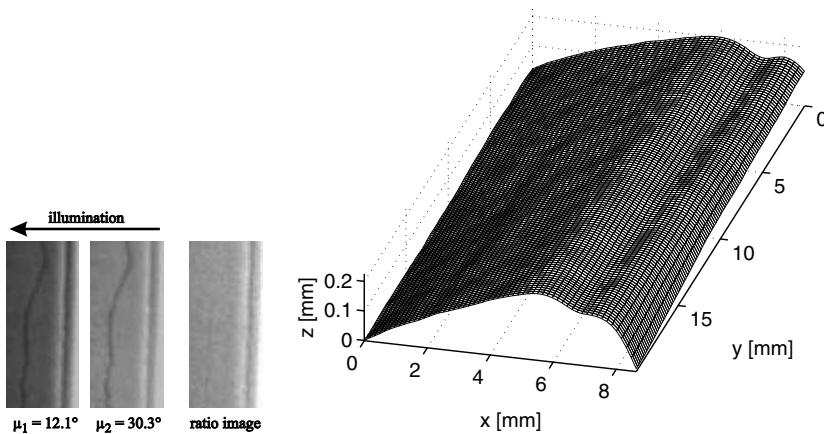


Fig. 5.14 Three-dimensional reconstruction of a metallic surface of non-uniform albedo (dark line traversing the left half of the image from top to bottom). A constriction is visible at the right image border. The input images (left) are acquired at illumination angles of 12.1° and 30.3° , respectively. The surface reconstruction (right) is obtained by applying the ratio-based iterative update rule (2.51).

line disappears while the constriction stands out clearly. The reconstructed surface profile shows that the constriction with a depth of less than 0.1 mm is recovered while no artifacts occur at the location of the dark line. Hence, the ratio-based approach achieves to separate pixel brightness variations due to non-uniform albedo from those caused by changing surface gradients.

As another example, Fig. 5.15 shows the reconstructed surface of a steel sheet with a shallow deformation along with the corresponding albedo map ρ_{uv} . The surface profile was obtained by extending the fusion scheme for shading and shadow features outlined in Section 4.2 to several light sources as pointed out in Section 2.3. The reconstruction procedure consisted of applying the ratio-based iterative update rule Eq. (2.51) including the shadow-related error term (4.10), and subsequently the iterative update rule Eq. (4.13) with the error term (2.43) for multiple light sources. As the average inclination of the reconstructed part of the surface with respect to the ground plane amounts to several degrees, a reference plane was fitted to the reconstructed profile. In Fig. 5.15c, this reference plane has been subtracted from the reconstruction result. In the albedo map, variations appear that correspond to dark and bright spots and lines. Again, intensity changes in the image which are caused by shading effects have been separated from those due to variations of the surface albedo.

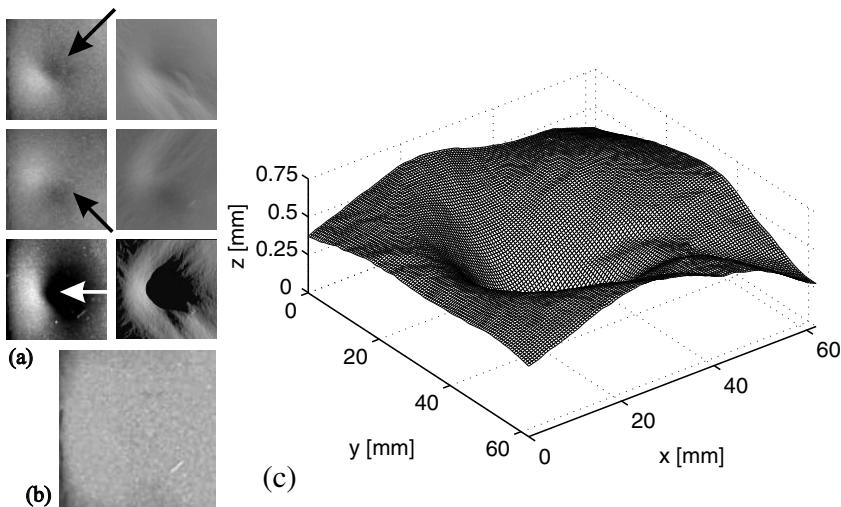


Fig. 5.15 Three-dimensional surface reconstruction of a steel sheet with a deformation. (a) Two shading images and one shadow image with their simulated counterparts. Illumination direction is as indicated by the arrows. (b) Albedo map computed according to Eq. (2.49). (c) Reconstructed surface profile, shown with respect to the fitted reference plane. The depth of the deformation corresponds to 0.36 mm.

5.3.3 Inspection Based on SfPR and SfPRD

The surface reconstruction algorithms described in Section 4.3 were applied to the raw forged iron surface of a connection rod (d'Angelo and Wöhler, 2005b, 2006, 2008). In this section we describe the obtained three-dimensional reconstruction results and compare them to a ground truth cross-section of the same surface, measured with a scanning laser focus profilometer. The second inspected part is a slightly damaged section of the raw surface of a flange also consisting of forged iron. Its surface shows several small deformations. We compare the depth of these deformations inferred from our three-dimensional reconstruction results to ground truth values obtained by tactile measurement.

We utilise a convergent stereo setup consisting of two CCD cameras of 1032×776 pixels image resolution, equipped with lenses of 25 mm focal length. The baseline distance of the cameras is 320 mm and the average distance to the object amounts to 480 mm. The resulting field of view corresponds to 10° . The size of the image sections used for three-dimensional reconstruction is 240×240 pixels. The surface is illuminated by one single or subsequently by two LED point light sources. The intensity I , polarisation angle Φ , and sparse depth Z are used for three-dimensional reconstruction. In addition to the fact that intensity and polarisation degree essentially provide redundant information as pointed out in Section 4.3.4, for the regarded rough metallic surfaces the behaviour of D is strongly affected by small-scale variations of the surface roughness. Accordingly, the value of the

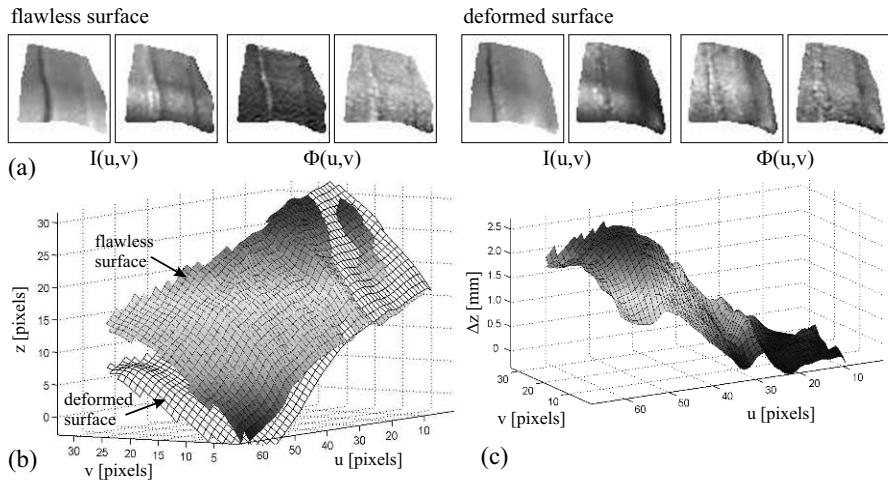


Fig. 5.16 Application of the local SfPR algorithm to the raw forged iron surface of the connection rod (d'Angelo and Wöhler, 2005b). (a) Images of a flawless and of a deformed surface. (b) Comparison of the three-dimensional surface profiles obtained with the local optimisation scheme. (c) Difference between the two profiles.

polarisation degree D for specular reflection varies across the surface by up to 20 percent. Hence, the polarisation degree does not represent a useful feature for three-dimensional reconstruction in this application context.

This unfavourable behaviour of the polarisation degree is known from previous work in the domain of photopolarimetry. A modified Fresnel equation for the polarisation degree as a function of incidence angle is derived by Morel et al. (2005) for smooth, specularly reflecting metallic surfaces based on the assumption that the absolute value of the complex diffraction index of the surface material is much larger than 1. However, Germer et al. (2000) demonstrate that the polarisation degree strongly depends on the microscopic surface roughness even for smooth, polished and etched steel surfaces. Their measurements cannot be explained by a simple physical model, but it is necessary to take into account microroughness and subsurface scattering effects. The experimental results by Germer et al. (2000) give an impression of the difficulties encountered when attempting to compute the polarisation degree of light reflected from a metallic surface based on physical models. Based on our experiments regarding raw forged iron materials, however, we found that in contrast to the polarisation degree, the polarisation angle is not perceptibly influenced by slight variations of the surface roughness. As a consequence, the polarisation degree is a feature which is useful for determination of the surface orientation only for smooth dielectric surfaces, which can be accurately described in terms of the Fresnel equations (Atkinson and Hancock, 2005b).

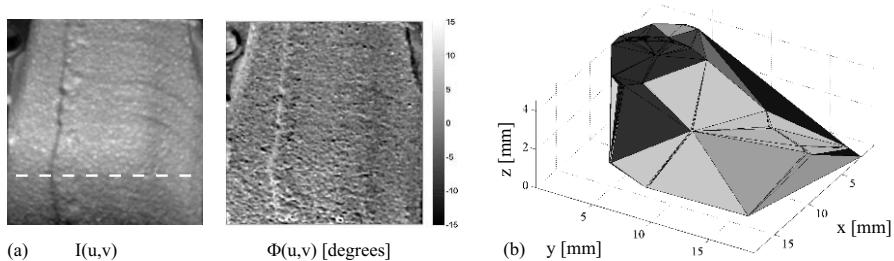


Fig. 5.17 (a) Reflectance and polarisation angle images of the raw forged iron surface of the connection rod. (b) Triangulated stereo reconstruction result.

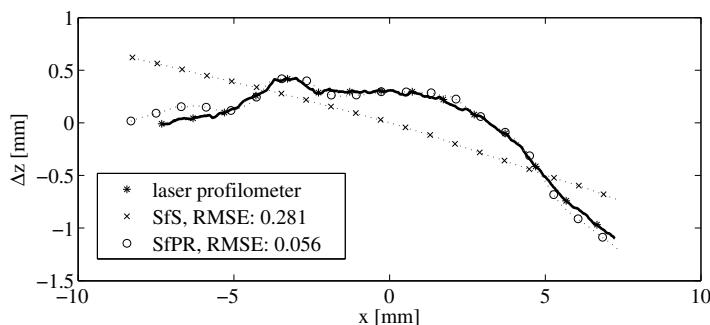


Fig. 5.18 Cross-sections of the raw forged iron surface of the connection rod, compared to ground truth (RMSE in mm), obtained with the SfPR and the shape from shading approach (SfS), no stereo information, initialisation with zero surface gradients, albedo estimated based on specular reflections according to Eq. (4.23).

5.3.3.1 Results Obtained with the SfPR Technique

For three-dimensional surface reconstruction of the raw forged iron surface of the connection rod with the local SfPR approach according to d'Angelo and Wöhler (2005b) described in Section 4.3.1.2 we employed two intensity images and one polarisation angle image. Fig. 5.16a shows a flawless part and a part that displays a surface deformation. We utilised the ratio-based and thus albedo-independent intensity error term according to Eq. (2.48). The deviation between the flawless and the deformed surface becomes evident in Fig. 5.16b–c. The comparison between the ground truth and the corresponding cross-section extracted from the reconstructed profile yields a RMSE of 220 μm .

Two experiments concerning the application of the global SfPR approach to the connection rod surface were performed (d'Angelo and Wöhler, 2008). In the first experiment, we initialised the surface gradients by zero values and determined the uniform surface albedo ρ_0 according to Eq. (4.23), relying on specular reflections. Cross-sections extracted from the corresponding reconstructed surface profiles and

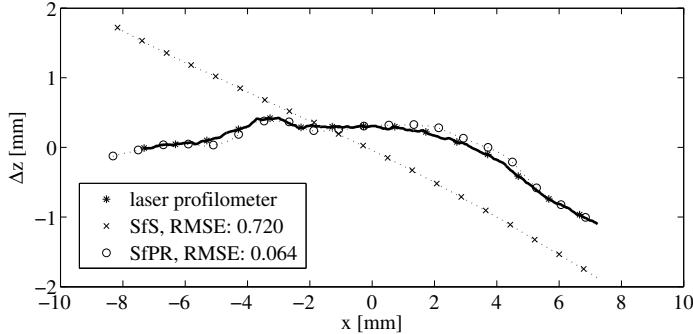


Fig. 5.19 Cross-sections of the raw forged iron surface of the connection rod, compared to ground truth (RMSE in mm). SfPR and shape from shading approach (SfS), no stereo information, initialisation with surface gradients obtained by depth from defocus, albedo estimated based on all image pixels according to Eq. (4.24).

their comparison to ground truth are shown in Fig. 5.18. The RMSE values are 56 µm for the SfPR approach and 281 µm for the shape from shading approach which neglects polarisation information. While the SfPR approach yields a very accurate reconstruction of the surface, the shape from shading approach estimates a largely uniform value of the surface gradient perpendicular to the direction of incident light due to the minor influence of this gradient on the error function.

In the second experiment, we initialised the global optimisation scheme with the surface gradients p_{uv}^{DfD} and q_{uv}^{DfD} inferred from the depth from defocus result as described in Section 4.3.3.1. To calibrate the depth from defocus algorithm, a linear function was fitted to the $(\Sigma, (z - z_0))$ data points (cf. Section 3.2.2). The calibration curve is shown in Fig. 3.5, and for illustration purposes, the raw and median-filtered depth from defocus measurements are shown in Fig. 3.7 (cf. Section 3.2.2). The profile shows many spurious structures especially on small scales, such that shape details are not reliably recovered. Three-dimensional reconstruction was performed based on a combination of intensity and polarisation angle (Fig. 5.17a). The albedo ρ_0 was estimated based on all image pixels according to Eq. (4.24) with the surface gradients set to p_{uv}^{DfD} and q_{uv}^{DfD} and was kept constant during the iteration process. Cross-sections extracted from the corresponding reconstructed surface profiles and their comparison to ground truth are shown in Fig. 5.19. The RMSE values are 64 µm for the SfPR approach and 720 µm for the shape from shading approach. The shape from shading approach again does not estimate correctly the surface gradients perpendicular to the direction of incident light, which results in a large RMSE value. Including polarisation information yields largely the same result as obtained with the albedo estimated from specular reflections, but without requiring the presence of specular reflections in the image.

For the flange as shown in Fig. 5.20a the global SfPR approach was initialised with zero surface gradients, and the uniform surface albedo ρ_0 was determined according to Eq. (4.23), relying on specular reflections. The three-dimensional recon-

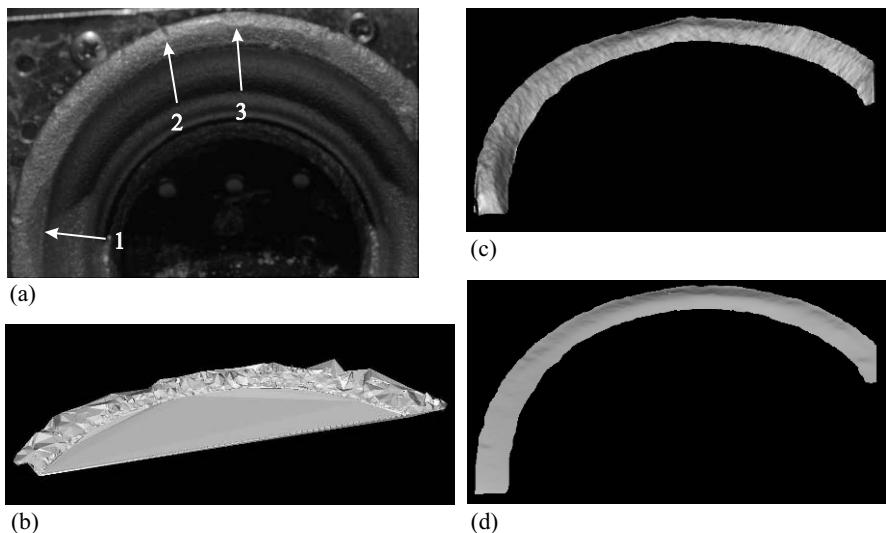


Fig. 5.20 (a) High dynamic range image of the flange, displayed at logarithmic scale. Three-dimensional reconstruction is performed for the ring-shaped surface part. The depths of the indicated dents were measured on the reconstructed surface profile and compared to ground truth data. (b) Triangulated stereo reconstruction result. (c) SfPR, no stereo information. (d) Shape from shading, no stereo information. Albedo estimated based on specular reflections according to Eq. (4.23).

struction is performed for the ring-shaped part only as the neighbouring parts are situated in the shadow and only visible due to secondary reflections (Fig. 5.20a is a high dynamic range image displayed at logarithmic scale). What is more, the surface normals of the neighbouring parts are nearly orthogonal to the viewing direction. Our goniometer setup for measuring the intensity and polarisation reflectance functions does not cope with such an extreme viewing geometry, such that in this range the reflectance function values are unknown. Furthermore, photometric surface reconstruction techniques are most favourably applied when the view on the surface is largely perpendicular (McEwen, 1991). Although the small-scale deformations of the surface are clearly apparent in the SfPR result (Fig. 5.20c) and to a lesser extent also in the shape from shading result (Fig. 5.20d), large-scale deviations from the essentially flat true surface shape are apparent.

5.3.3.2 Results Obtained with the SfPRD Technique

Our convergent stereo setup was calibrated with the automatic camera calibration system described by Krüger et al. (2004). After acquisition of the images, they were rectified to standard epipolar geometry. Effectively, this results in typical disparity values of around 4000 pixels at the object distance in the rectified stereo image pairs. Experiments with synthetic data have shown that the standard deviation of the

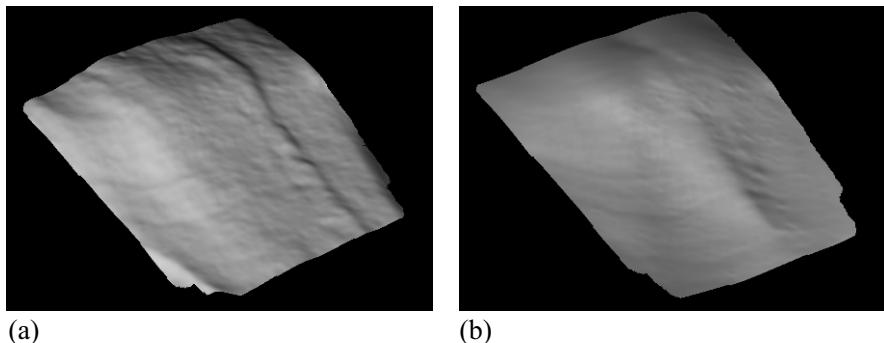


Fig. 5.21 Reconstructed surface profile of the connection rod. (a) SfPR and (b) shape from shading approach combined with stereo information, albedo estimated during the iteration process based on all image pixels according to Eq. (4.24).

disparity amounts to 0.3 pixels, resulting in a standard deviation of 30 μm of the resulting depth points. One of the stereo cameras is equipped with a rotating linear polarisation filter and is used to acquire the images required for SfPR according to Section 4.3.1. Due to the highly specular reflectance of the metallic surfaces, usually only a sparse set of depth points can be reliably extracted using the blockmatching stereo algorithm.

For the raw forged iron surface of the connection rod, Fig. 5.17a shows the intensity and polarisation angle image and Fig. 5.17b the triangulated stereo reconstruction result (d'Angelo and Wöhler, 2006). The surface albedo was estimated based on Eq. (4.24) during each step of the iteration process. We found that the RMSE between the corresponding cross-section extracted from our reconstructed three-dimensional profile and the ground truth amounts to 45 μm (Figs. 5.21a and 5.22). If the shape from shading approach is used such that polarisation information is not taken into account, the RMSE is 163 μm (cf. Fig. 5.21b). The RMSE of the combined shape from shading and stereo approach is larger than that of the stereo reconstruction alone, corresponding to 80 μm , since no stereo data are available for the rightmost 3.2 mm of the cross-section. Neglecting this margin yields a much smaller RMSE of 50 μm . For the examined strongly specular surface, Figs. 5.21a and 5.21b illustrate that in contrast to the shape from shading approach, the SfPR method reveals a large amount of small scale surface detail. The results of the comparison to ground truth data are summarised in Table 5.4.

Fig. 5.23 shows the three-dimensional reconstruction of the flange surface calculated using one intensity and one polarisation angle image along with stereo depth information. The triangulated set of stereo depth points is shown in Fig. 5.20b. As in the previous example, the surface albedo was estimated during the iteration process according to Eq. (4.24). In contrast to the first experiment, it was not possible in the second experiment to determine accurate ground truth values for a cross-section through the surface because the laser profilometer is not suitable for acquir-

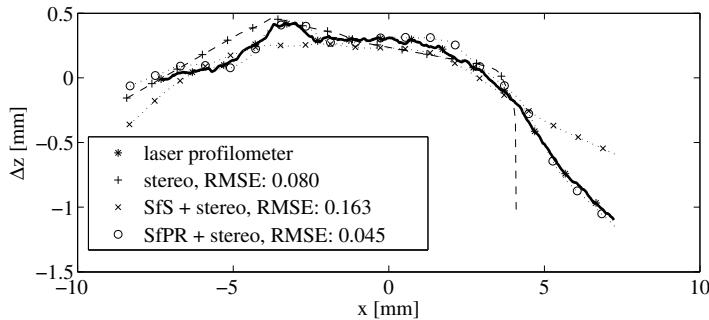


Fig. 5.22 Cross-sections of the raw forged iron surface, compared to ground truth (RMSE in mm). SfPR and shape from shading approach combined with stereo information, albedo estimated during the iteration process based on all image pixels according to Eq. (4.24).

Table 5.4 Three-dimensional reconstruction results for the raw forged iron surface of the connection rod, obtained based on comparison of the cross-section shown in Fig. 5.17a. Albedo determination marked as “initial” denotes that the albedo was estimated prior to the iteration process either based on specular reflections or based on depth from defocus data and was not changed afterwards, while “adapted” denotes an estimation of the albedo during the iteration process.

Utilised information	Albedo determination	Figure	z	RMSE (μm)
SfS	Eq. (4.23), initial	5.18	281	
SfPR	Eq. (4.23), initial	5.18	56	
SfS, DfD	Eq. (4.24), initial	5.19	1153	
SfPR, DfD	Eq. (4.24), initial	5.19	55	
Stereo	—	5.22	80	
SfS, stereo	Eq. (4.24), adapted	5.22	50, (163)	
SfPR, stereo	Eq. (4.24), adapted	5.22	45	

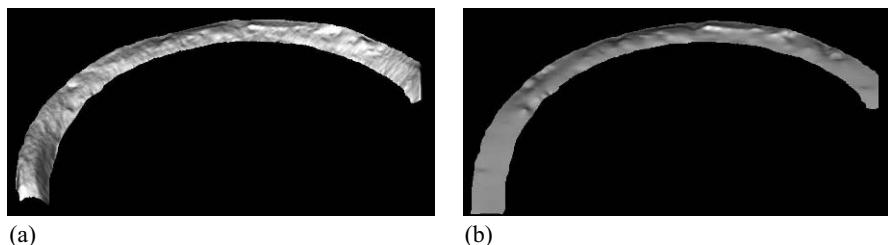


Fig. 5.23 Reconstructed surface profile of the flange. (a) SfPRD and (b) shape from shading approach combined with stereo information, albedo estimated during the iteration process based on all image pixels according to Eq. (4.24).

ing measurements of such a large and curved surface section. Instead, we regarded the depths of the three dents indicated in Fig. 5.20a, for which ground truth values were obtained by tactile measurement and compared to the reconstructed depth differences. Due to the small size of the surface defects the accuracy of the tactile depth

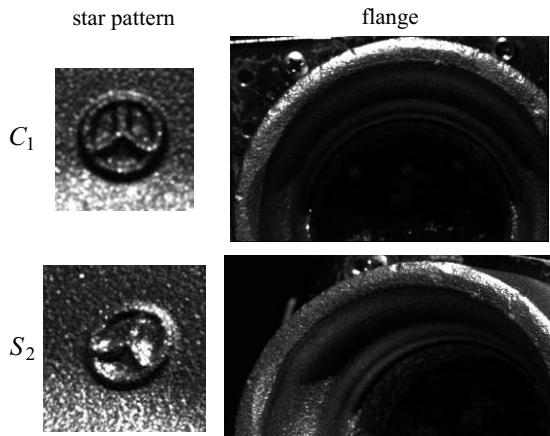


Fig. 5.24 Stereo image pairs of the star pattern (left) and the ring-shaped flange (right). These examples illustrate that due to the specular reflectance behaviour of the surface, corresponding surface parts do not necessarily have a similar appearance in the images.

measurement only amounts to 0.05 mm. The true depth of dent 1 is 1.2 mm, the reconstructed depth 1.3 mm. Dents 2 and 3 each have a true depth of 0.25 mm, while the corresponding depth on the reconstructed surface profile amounts to 0.30 mm and 0.26 mm, respectively. On large scales, our three-dimensional reconstruction correctly displays a flat surface. These comparisons indicate a good correspondence between the true surface and our reconstruction results.

5.3.4 Inspection Based on Specular Stereo

In this section the specular stereo method described in Section 4.4 (Wöhler, 2008; Wöhler and d’Angelo, 2009) is applied to the three-dimensional reconstruction of rough metallic surfaces displaying a weak diffuse reflection component along with considerable specular lobes and spikes. The same experimental setup as in Section 5.3.3.2 is used. We examine the surface of the connection rod, a cast iron surface displaying a star pattern, and the ring-shaped surface part of the flange. The first and the third example have also been regarded in Section 5.3.3.2.

5.3.4.1 Qualitative Discussion of the Three-dimensional Reconstruction Results

Fig. 4.26 displays a stereo image pair of the connection rod, while in Fig. 5.24 the stereo image pairs of the star pattern and the ring-shaped flange examples are shown. The appearance of the surface of the star pattern differs so strongly between the two

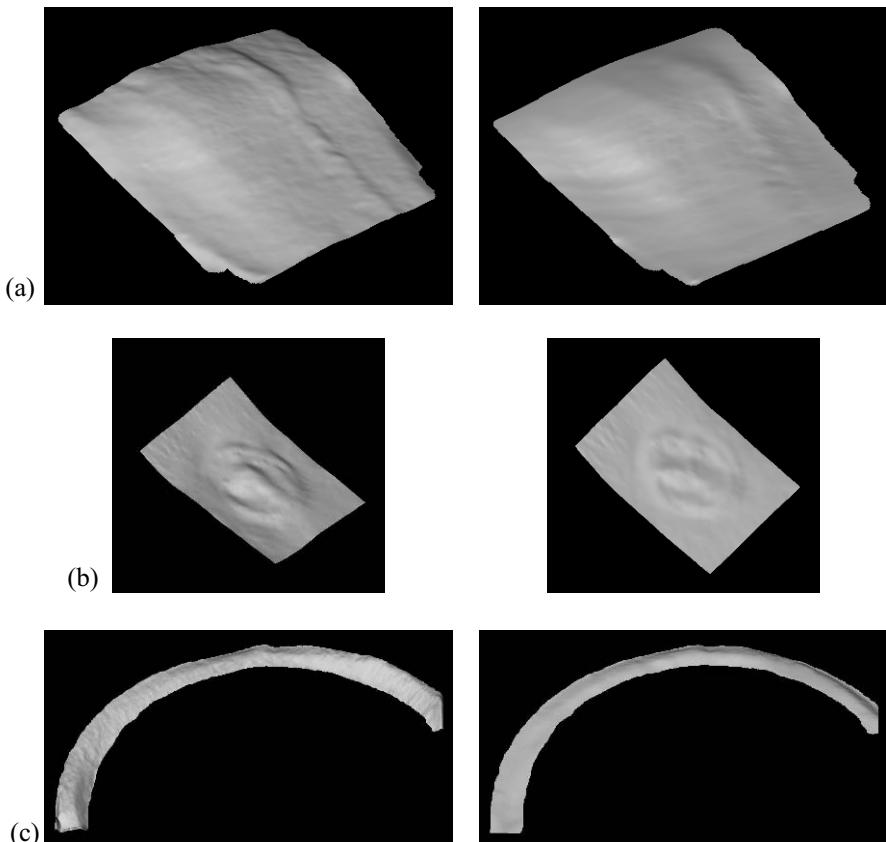


Fig. 5.25 Three-dimensional reconstruction results for (a) the connection rod, (b) the star pattern, and (c) the ring-shaped flange. The images in the left column display the results obtained based on intensity and polarisation angle as photometric information, while for the images in the right column the polarisation information was neglected.

images that initially only one single stereo point can be determined.¹ In the flange example, the brightness distribution across the surface is also fairly different for the stereo images, but the appearance of the surface is sufficiently similar to obtain initial three-dimensional points for about 10 percent of the image pixels situated on the ring section.

Fig. 5.25 shows the final three-dimensional reconstruction results for the three regarded examples. In each case, the three-dimensional profile obtained using polarisation angle data from camera 1 is compared to the result obtained without po-

¹ In all experiments, the same blockmatching threshold was used. Slightly decreasing this threshold for the star pattern example would have resulted in more than just one initial 3D point. However, this somewhat extreme configuration was used intentionally in order to illustrate that this information is sufficient to obtain convergence of the specular stereo scheme.

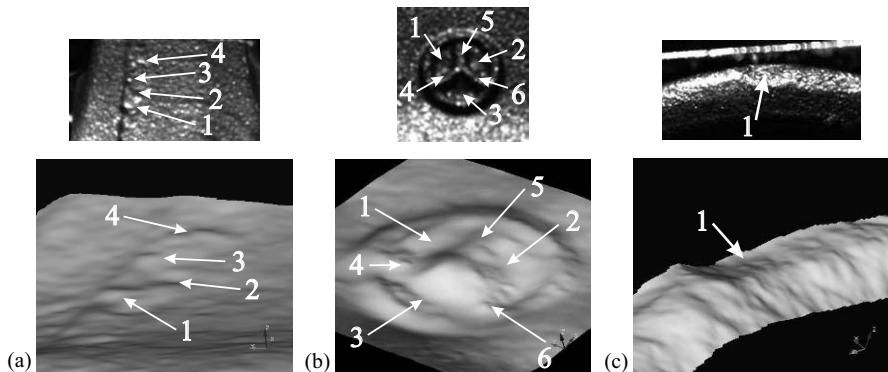


Fig. 5.26 Close-up views of the three-dimensional surface profiles in comparison to the input images. (a) Small positive relief structures (1, 3, 4) and a depression (2) on the surface of the connection rod. (b) Low parts (1, 2, 3) and ridges (4, 5, 6) of the star pattern. (c) Small depression on the surface of the ring-shaped flange.

larisation data, exclusively relying on intensity and stereo information. We found that neglecting polarisation data leads to an underestimation of the surface slopes in the direction perpendicular to the illumination direction, such that details like the bending in the leftmost part of the flange surface in Fig. 5.25c tend to disappear. This effect is due to the illumination by a single light source (Horn, 1989). These results illustrate that the pixel intensity $I(u, v)$ and the polarisation angle $\Phi(u, v)$ contain complementary information about the surface gradients $p(u, v)$ and $q(u, v)$, which is the main reason for the fact that a better 3D reconstruction is achieved when intensity and polarisation angle information is used.

Fig. 5.26 shows close-up views of the 3D surface profiles in comparison to the input images, clearly indicating that mesoscopic and macroscopic structures which appear as dark and bright spots in the input images can be reconstructed. Fig. 5.26a displays several small positive relief structures (1, 3, 4) and a depression (2) on the surface of the connection rod. In Fig. 5.26b the low parts of the star pattern are indicated by 1, 2, and 3 and the ridges by 4, 5, and 6. Fig. 5.26c illustrates how the specular stereo method recovers a shallow depression on the surface of the ring-shaped flange (cf. also Section 5.3.4.2).

5.3.4.2 Comparison to Ground Truth Data

For the connection rod example a highly accurate reference profile was measured with a laser profilometer for a cross-section of the surface as indicated in Fig. 5.27. The root-mean-square deviation between the result of the specular stereo method and the ground truth amounts to $55 \mu\text{m}$, which is somewhat smaller than the average lateral extension of the image pixels of $86 \mu\text{m}$.

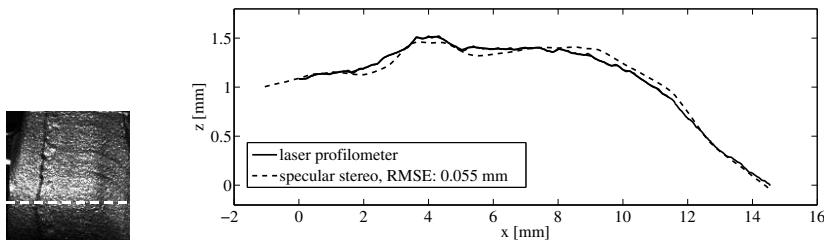


Fig. 5.27 Comparison between a cross-section through the three-dimensional surface profile obtained with specular stereo, relying on intensity and polarisation angle data, and the ground truth determined with a laser profilometer (connection rod example).

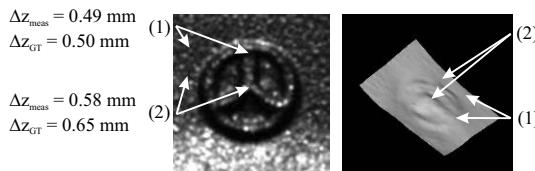


Fig. 5.28 Comparison between height differences on the surface obtained by specular stereo and the corresponding ground truth values determined based on tactile measurement (star pattern example).

For the star pattern example it was not possible to obtain ground truth values with the laser profilometer as the depth differences across the profile are too large and the surface slopes are too steep. Hence, we determined height differences at two representative locations on the surface by tactile measurement. At both locations, the correspondence between the specular stereo reconstruction and the ground truth is better than 0.1 mm. For the ring-shaped flange example, a similar technique was used to examine the surface reconstruction accuracy.

For the shallow depression marked in the lower right point cloud in Fig. 5.29, a depth of 0.23 mm was obtained by specular stereo, which is in good agreement with the ground truth value of 0.25 mm obtained by tactile measurement. The same depth value was inferred from the dense surface profile shown in Fig. 5.25 obtained based on the stereo point cloud and the available photometric and polarimetric data.

5.3.4.3 Self-consistency Measures for Three-dimensional Reconstruction Accuracy

The proposed specular stereo algorithm yields several self-consistency measures which provide an impression of how accurately the surface is reconstructed. These quantities describe how consistent the geometric and the photopolarimetric data are with each other, given the result of the specular stereo algorithm.

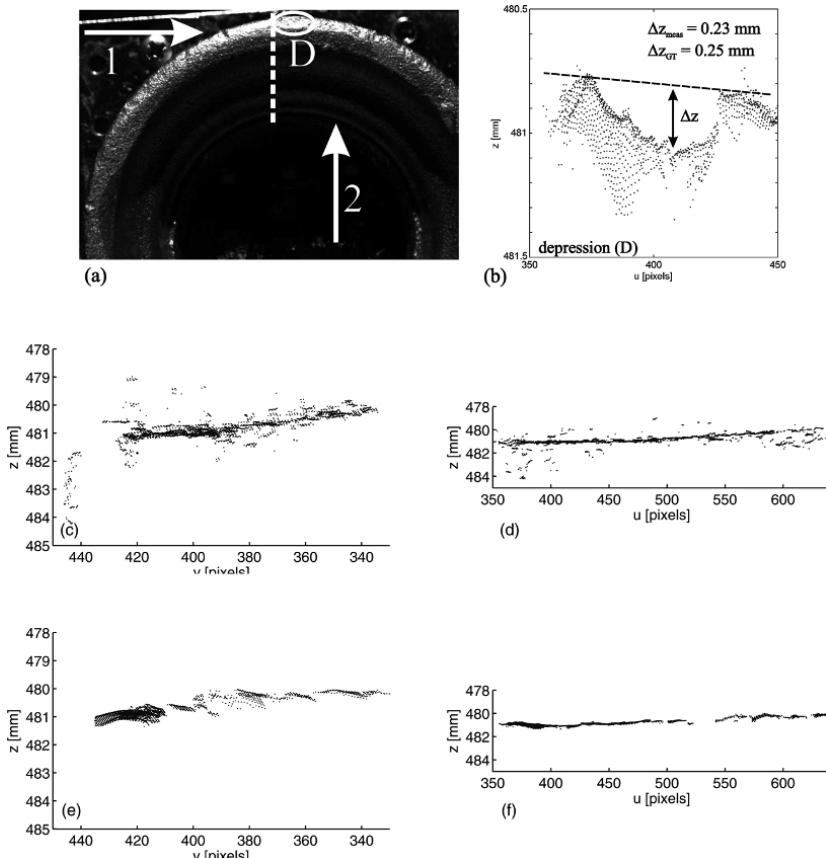


Fig. 5.29 Three-dimensional point clouds generated by the initial blockmatching stereo and by the specular stereo algorithm (flange example). For clarity, in all diagrams only the part of the point cloud right of the dashed line marked in (a) is shown. (a) Viewing directions into the point cloud. A shallow depression in the surface is marked by “D”. (b) Enlarged part of the three-dimensional point cloud generated by specular stereo as seen from direction 2, showing a side view of the shallow depression. (c) View from direction 1 and (d) from direction 2 into the initial three-dimensional point cloud. (e) View from direction 1 and (f) from direction 2 into the three-dimensional point cloud generated by the specular stereo algorithm.

An intuitive measure for reconstruction accuracy is the appearance of the three-dimensional point cloud obtained by stereo analysis. In Fig. 5.29 two views into the three-dimensional point cloud of the ring-shaped flange are shown for the initial blockmatching stage and for the final result of the specular stereo method. The initial blockmatching stage yields a large number of three-dimensional points (cf. centre row in Fig. 5.29) which appear to form a plane surface. The depth values are fairly noisy, and some outliers are apparent which deviate from the plane by several

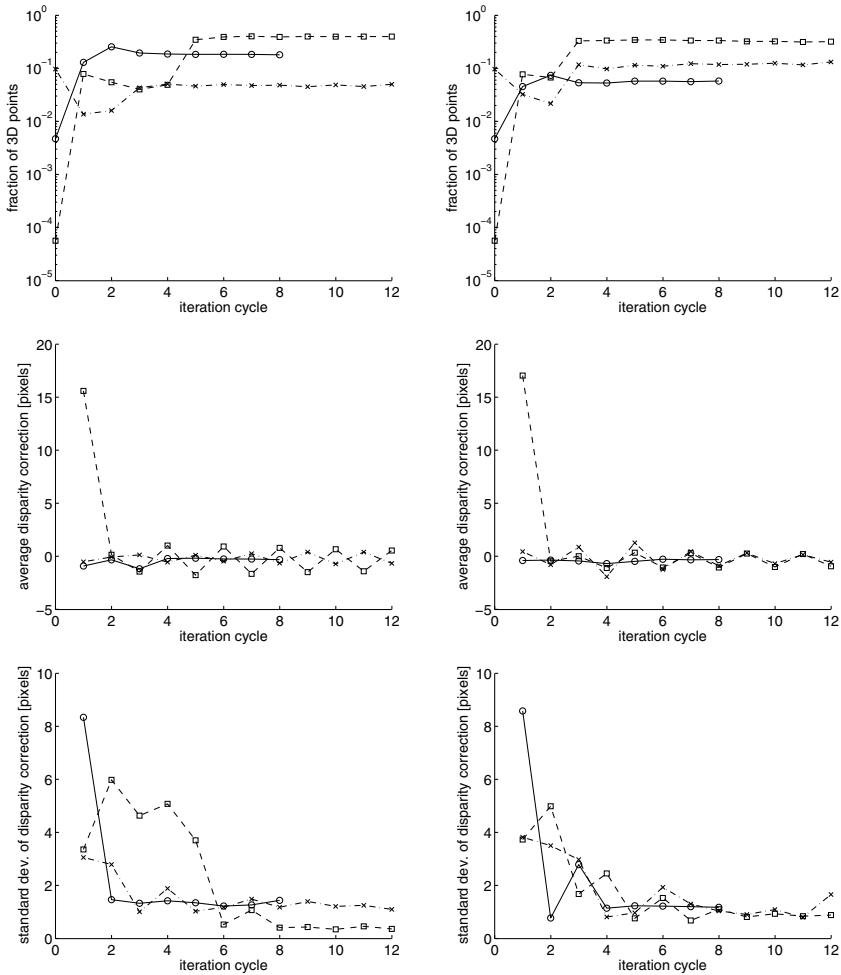


Fig. 5.30 Fraction of three-dimensional points relative to the number of pixels (top row), average disparity correction (centre row), and standard deviation of the disparity correction (bottom row) during the iteration process for the connection rod example (solid curve, circles), the star pattern example (dashed curve, squares), and the flange example (dashed-dotted curve, crosses). Left column: Results obtained based on intensity and polarisation angle information. Right column: Results obtained based on intensity information alone.

millimetres. This behaviour presumably results from points on the surface appearing more or less strongly dissimilar in the stereo images. The point cloud obtained by specular stereo (bottom row) is significantly less noisy, and the bent cross-section of the ring-shaped surface is clearly apparent. The shallow depression in the surface is also visible (it has already been shown above that its reconstructed depth is in good correspondence with tactile measurement).

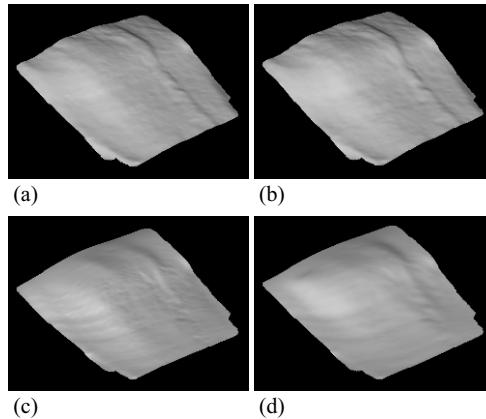


Fig. 5.31 Three-dimensional reconstruction result for poorly known reflectance parameters (connection rod example, cf. also Fig. 5.25). (a) Reflectance function with $\sigma_2 = 0$ (i.e. specular spike omitted) and correct parameters for the specular lobe. Polarisation angle data are used. (b) Reflectance function with $\sigma_2 = 0$ and σ_1 and m_1 decreased by a factor of 0.5, respectively. Polarisation angle data are used. (c), (d) Same reflectance parameters as in (a) and (b), but without taking into account polarisation angle data.

For the connection rod and the star pattern, the initially very small fraction of image pixels for which a stereo correspondence can be established increases by several orders of magnitude in the course of the iteration process (cf. left row of Fig. 5.30, top). For the ring-shaped flange, the number of (initially very noisy) three-dimensional points decreases by about a factor of 2, but the accuracy of the measured depth values significantly increases. At the end of the iteration process, the average disparity correction is smaller than 0.7 pixels for all three examples (cf. Fig. 5.30, left row, middle). A non-zero average value of Δd corresponds to a uniform offset of the surface profile in z direction. Another important self-consistency measure is the standard deviation $\sigma_{\Delta d}$ of the disparity correction, which directly quantifies how closely the rendered images derived from the reconstructed surface match the observed stereo images (cf. Fig. 5.30, left row, bottom). Initially, $\sigma_{\Delta d}$ is larger than 8 pixels for the connection rod, while it amounts to about 3 pixels for the star pattern and the flange. These fairly large values indicate large-scale discrepancies between the initial three-dimensional reconstruction and the stereo images. The specular stereo algorithm yields values for $\sigma_{\Delta d}$ of about 1 pixel for the connection rod and the flange and 0.4 pixels for the star pattern (for our stereo configuration, a disparity difference of 1 pixel corresponds to a depth difference of 0.135 mm in the regarded range of depth values). If the polarisation information is neglected, the self-consistency measures still show a similar behaviour (cf. right column in Fig. 5.30).

As a whole, the average values and the standard deviations of the disparity correction inferred for the three examples provide self-consistency measures for the accuracy of the three-dimensional reconstruction result of the specular stereo algorithm

that indicate residual errors which are comparable to the deviations between the three-dimensional reconstruction result and independently measured ground truth data.

5.3.4.4 Consequences of Poorly Known Reflectance Parameters

The proposed specular stereo algorithm requires knowledge about the parameters of the photometric and polarimetric reflectance functions. In this section we discuss the behaviour of the specular stereo algorithm for poorly known reflectance parameters, regarding the connection rod example. For the reflectance function according to Eq. (2.15) we determined the parameters $\sigma_1 = 3.85$ and $m_1 = 2.61$ for the specular lobe and $\sigma_2 = 9.61$ and $m_2 = 15.8$ for the specular spike based on a reference measurement. In the first experiment we omitted the specular spike by setting $\sigma_2 = 0$, which results in the three-dimensional profile shown in Fig. 5.31a. If we additionally decrease the intensity and increase the width of the specular lobe by setting σ_1 and m_1 to 0.5 times their true values, respectively, yields the three-dimensional profile shown in Fig. 5.31b. If furthermore the polarisation information is neglected, the three-dimensional profiles shown in Fig. 5.31c–d are obtained.

Compared to the reconstruction results shown in Fig. 5.25 obtained with the correct reflectance parameters, the three-dimensional profiles shown in Fig. 5.31 are tilted towards the light source. Hence, the SfPRD algorithm attempts to generate the large pixel brightness values observed in the images for specular reflections by decreasing the incidence angle between the normals of the corresponding surface parts and the direction of incident light. However, we observe a graceful degradation of the three-dimensional reconstruction result even when the assumed reflectance parameters strongly deviate from their true values.

5.3.5 Discussion

In the previous sections we have shown that the proposed combined framework for three-dimensional surface reconstruction based on photopolarimetric information and independently measured depth data can be favourably applied to the application scenario of industrial quality inspection. The accuracy of depth differences on the surface measured with the SfPRD method is about twice as high as the lateral pixel resolution of the utilised images. The duration of image acquisition for the SfPRD approach amounts to a few seconds. About one second is necessary to compute the three-dimensional reconstruction result on standard industrial hardware. To obtain depth from defocus information, the acquisition of two images at different apertures may be automated using a motor-iris lens. Large-scale deviations of the estimated three-dimensional surface shape from the true shape may occur if the available depth data are too sparse. Hence, if traditional stereo analysis techniques do not yield three-dimensional point data of satisfactory accuracy or density, the proposed spec-

ular stereo method establishes a number of stereo correspondences that may become several orders of magnitude higher than the number of correspondences obtained by classical blockmatching. The three-dimensional point cloud obtained with the specular stereo method is less noisy, contains a negligible number of outliers, and shows significantly more surface detail than the point cloud obtained by classical blockmatching. At a lateral pixel resolution of 86 µm, the absolute depth accuracy of the specular stereo approach amounts to 30–100 µm. For poorly known reflectance parameters, a graceful degradation of the accuracy of the three-dimensional reconstruction result is observed.

This section discusses the possible advantages and drawbacks of the examined framework in comparison to active triangulation-based devices in the context of industrial quality inspection. Active devices based on laser triangulation may be an alternative approach to the regarded close-range problems. Simple and inexpensive sensors of this kind only measure a single profile across the surface at a time, based on a laser line projected onto the surface and an image acquired by a camera calibrated with respect to the projector. Hence, either the sensor or the object has to be moved synchronously with image acquisition when an area measurement is performed, which may in turn introduce an intricate and expensive mechanical setup such as a laser probe combined with translation-rotation motors, articulated arms, or coordinate measurement machines (Beraldin, 2004). What is more, a large number of profiles are necessary to obtain a high lateral resolution, which results in long measurement cycles. For example, a three-dimensional reconstruction of the two metallic surfaces regarded in Sections 5.3.3 and 5.3.4 would require the acquisition of about 200 and 1000 line profiles, respectively, given the lateral resolution of about 0.1 mm per pixel in these experiments. Such three-dimensional reconstruction methods may well be suitable for the inspection of parts randomly selected from the production line but not for in-line inspection scenarios with a few seconds cycle time.

Area measurements can also be carried out by image-based evaluation of fringe or coded patterns projected on the surface (Batlle et al., 1998; Beraldin, 2004). However, as soon as it is desired to obtain a lateral and vertical resolution comparable to the one achieved in the experiments described in this chapter, the costs of such a measurement system are currently at least about an order of magnitude higher than those of the instrumental setup utilised in our experiments, i.e. a pair of standard industrial cameras, a rotating polarisation filter, and one or several LED illumination devices.

Furthermore, it is well known that projection-based systems relying on single lines or coded patterns suffer from strong difficulties in the presence of highly specular surfaces like those regarded in Sections 5.3.3 and 5.3.4. The reason for this behaviour is the fact that the intensity variations in the image of the projected pattern may be considerable because for such surfaces the amount of diffuse reflection is small while mirror-like reflection is dominant. As a consequence, for some surface parts the projected light is reflected specularly into the camera, leading to pixel saturation or blooming, while for other parts it is reflected past the camera, result-

ing in invisible parts of the pattern—the reconstructed profile then shows significant gaps for which no data are available.

According to this discussion, the proposed framework is an accurate, cost-efficient, and fast method for three-dimensional surface reconstruction in largely controlled environments as long as the material-specific reflectance properties are well-known. In uncontrolled settings, active triangulation-based devices are presumably more suitable. In many application scenarios, it may thus be favourable to combine depth data obtained by active triangulation-based devices with image data according to the framework outlined in Sections 5.3.3 and 5.3.4.

Chapter 6

Applications to Safe Human–Robot Interaction

In this chapter we address the scenario of safe human–robot interaction in the industrial production environment. For example, in car manufacturing, industrial production processes are characterised by either fully automatic production sequences carried out solely by industrial robots or fully manual assembly steps where only humans work together on the same task. Close collaboration between humans and industrial robots is very limited and usually not possible due to safety concerns. Industrial production processes may increase their efficiency by establishing a close collaboration of humans and machines exploiting their unique capabilities, which requires sophisticated techniques for human–robot interaction. In this context, the recognition of interactions between humans and industrial robots requires vision methods for three-dimensional pose estimation and tracking of the motion of human body parts based on three-dimensional scene analysis.

Section 6.1 illustrates the importance of gesture recognition in the general context of human–robot interaction. Most existing methods in this field merely achieve a two-dimensional evaluation of the scene. Furthermore, an introduction to systems for safe human–robot interaction in the industrial production environment is given. In Section 6.2 we evaluate the performance of the three-dimensional approach by Schmidt et al. (2007) to the detection and tracking of objects in point clouds (cf. Section 1.6.3) in a typical industrial production environment. In Sections 6.3 and 6.4, the methods introduced by Barrois and Wöhler (2008) and Hahn et al. (2007) for three-dimensional detection, recognition, pose estimation, and tracking of human body parts (cf. Section 1.6.2) are evaluated in similar scenarios.

6.1 Vision-based Human–Robot Interaction

The important role of gestures in the context of human–robot interaction is outlined in Section 6.1.1. Systems for safeguarding the cooperation between humans and industrial robots are described in Section 6.1.2, illustrating that three-dimensional scene reconstruction techniques are required to maintain an appropriate degree of

system flexibility. Section 6.1.3 provides an overview of methods for recognising gestures and body postures applied to human–robot interaction scenarios.

6.1.1 The Role of Gestures in Human–Robot Interaction

The classical human-computer interface, consisting of the elements keyboard, mouse, and monitor, tends to be insufficient for an efficient interaction between a human and a robot. Especially if a human-like communication with the robot is desired, which acts as a partner in the natural environment of the human, communication based on several channels through which information is exchanged has shown to be preferable. In the theory of communication, these channels are termed modalities (Hofemann, 2007). The most important modalities for human–robot interaction are writing (e.g. keyboard, mouse, touchscreen, but also handwritten text), speech, and gestures. A communicative channel always requires a device to record the corresponding human action, such as microphones for speech and cameras for gestures.

Generally, communication between humans involves various channels simultaneously. A typical behaviour is e.g. to talk about an object in the field of view (speech) and at the same time to reference it by pointing to it (gestures). Hence, many human communicative expressions are only understandable by simultaneous interpretation of speech and gestures. A multimodal robot system should thus perceive a human, understand and interpret verbal and nonverbal utterances, and react accordingly. The high relevance of gestures in the field of human–robot interaction is pointed out by Turk (2005). In the context of multimodal human-computer interaction, highly relevant tasks which can be solved by computer vision and pattern recognition methods are e.g. finding and recognising faces, analysing facial expressions, tracking the human body and hands, recognising gestures, and analysing actions.

A general distinction of gestures into deliberative and undeliberative gestures and a further subdivision of deliberative gestures is suggested by Ekman and Friesen (1969). More specifically, Pavlovic et al. (1997) define gestures for human–robot interaction as tasks performed by the human hand, including manipulations of objects and communicative actions. Nehaniv (2005) introduces five classes of gestures relevant for human–robot interaction, including irrelevant and manipulative gestures, e.g. arm motion while running, gripping a cup, side effects of expressive behaviour, e.g. motion of the hands, arms, and face during a conversation, symbolic gestures having a well-defined meaning which depends on the cultural context, interactional gestures (regulators) for initiating, synchronising, or terminating an interaction, and referencing (deictic) gestures such as pointing on objects.

Deictic gestures contain information about the direction in space in which the referenced object is situated. It is shown e.g. by Nickel et al. (2004) and by Groß et al. (2006) that knowledge about the head pose and the pointing direction of the arm allows to determine the referenced object. In some cases it is favourable to take into account the viewing direction of the person instead of the head pose (Nickel

and Stiefelhagen, 2004). Manipulative gestures may become an important part of the communication process as soon as a human, e.g. a parent, teaches to another human, e.g. a child, how to perform a certain task by demonstration rather than verbal explanation (Hofemann, 2007). According to Pavlovic et al. (1997), the technical representation of a gesture in the context of human–robot interaction is based on the time-dependent vector $\mathbf{T}(t)$ of pose parameters describing the orientation and internal degrees of freedom of the hand and the arm. The gesture is described by the trajectory in the space of pose parameters over a certain time interval.

It is shown in the following sections that the recognition of the movements and postures of humans as well as especially their hands and arms is highly relevant in the context of systems that safeguard human–robot interaction in the industrial production environment. Hence, the algorithms required for such systems are closely related to those used for gesture recognition.

6.1.2 Safe Human–Robot Interaction

Many systems for sensor-based avoidance of collisions between humans and robots use sensors which provide only local information. For example, capacitance sensors attached to the robot act as an artificial skin (Novak and Feddema, 1992; Feddema and Novak, 1994). Algorithms for whole-arm collision avoidance for robots with artificial skins are presented by Lumelsky and Cheung (1993). Furthermore, laser scanner techniques can be used to detect the human within the workspace of the robot. Som (2005) approximates the human by a three-dimensional cylinder based on the acquired distance data. The maximum allowed robot velocity depends on the distance between this cylinder and the robot.

An early vision-based approach to safeguarding the area around an industrial robot is described by Baerveldt (1992). Possible collisions between a human and the robot are detected by a two-dimensional difference image method. As long as no humans are present in the workspace of the robot, the robot performs its task at maximum velocity. If humans are present, the velocity is reduced, and the robot is stopped if a human approaches the robot too closely. A similar system relying on difference images is proposed by Ebert (2003). Only two velocity levels (normal velocity and stop) are implemented. The recognition of evasive movements of the human allows to reduce the number of stops. A system for safeguarding the assembly of small parts is described by Schulz (2003) and by Thiemermann (2005). Their work mainly concentrates on the ergonomic aspects of the robot movement. Cameras are used to monitor the workspace of the robot and to detect the human. The permitted velocity and acceleration of the robot are determined based on the distance to the human and the angle between the motion trajectories of the human and the robot.

The SIMERO system (Ebert and Henrich, 2003) for safe human–robot interaction provides an integrated framework for safeguarding the collaboration between a human and an industrial robot. An image processing module performs a two-

dimensional difference image analysis based on a reference image of the scene, resulting in the silhouette of the human. A robot modelling stage yields the silhouette of the robot relying on its joint angles. The extracted silhouettes of the human and the robot are used to predict collisions. Subsequently, the trajectory of the robot is planned. Gecks and Henrich (2005) regard the problem of pick-and-place operations which lead to discrete changes in the image. An update algorithm is introduced which adapts the reference image to the changed configuration of objects. Several images acquired from different viewpoints are used to confirm the presence or absence of an object. The multiple-camera setup, however, is merely used for checking the plausibility of an object detection but not for an actual three-dimensional reconstruction of the scene. Kuhn et al. (2006) present a method for safeguarding guided robot movement in the context of human–robot interaction. The presented approach is based on the fusion of data acquired by a camera and by a force-torque sensor, which measures the direction and amount of the force and moment applied to the robot. The workspace is monitored and distances between the robot and objects in the scene are estimated by applying an extended difference image method. The force-torque sensor provides guidance information. The maximum allowed velocity of the robot is limited according to the distance between the robot and the human or the closest obstacle, and the robot trajectory is determined based on the guidance information.

The vision-based collision avoidance module of the SIMERO system does not generate a three-dimensional representation of the workspace but relies exclusively on silhouette image information of the scene and the robot (Ebert and Henrich, 2003). For a safety system capable of monitoring an extended workspace around an industrial robot or a machine and precisely predicting possible collisions with a human, however, a three-dimensional reconstruction of the scene is indispensable. As a consequence, the SafetyEYE system has been created in cooperation between the Mercedes-Benz production facilities in Sindelfingen, Daimler Group Research and Advanced Engineering in Ulm, and the company Pilz GmbH & Co. KG, a manufacturer of safe automation systems (cf. Winkler (2006), who provides a detailed introduction to the SafetyEYE system, the cooperation in which it has been developed and commercialised, application-oriented engineering aspects concerning the integration of the system into automobile production facilities, and future extensions). The stereo vision and camera calibration algorithms for the SafetyEYE system have been developed in a research project led by Dr. Lars Krüger and the author. The SafetyEYE system consists of a calibrated trinocular camera sensor (cf. Fig. 6.1a), which monitors the protection area around the machine, and two industrial PCs. Two different algorithms for stereo image analysis determine a three-dimensional point cloud describing the structure of the scene being monitored. As soon as a certain number of three-dimensional points is detected inside the protection area, the system initiates the protective measures necessary to prevent an accident, either by slowing down or by stopping the machine. The system is installed by defining the three-dimensional virtual protection areas with a configuration software (cf. Fig. 6.1b). Setting up a traditional safety system consisting of several components such as metal

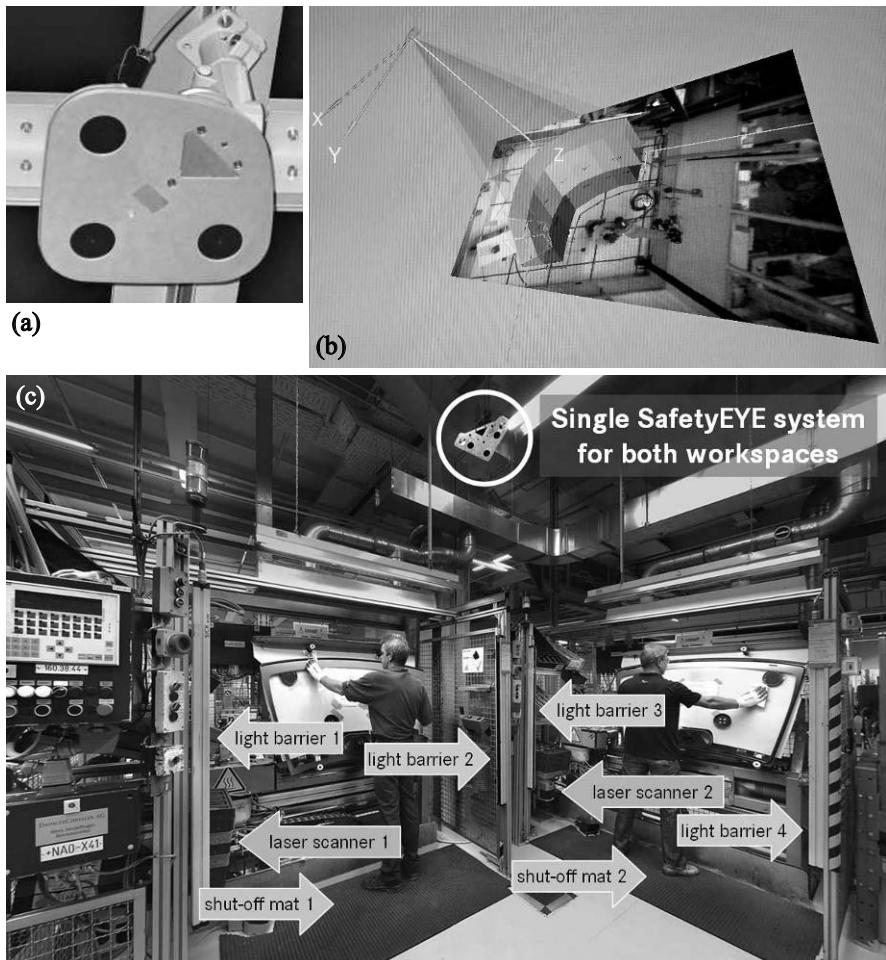


Fig. 6.1 The SafetyEYE system. (a) The trinocular camera sensor. (b) Configuration of virtual protection areas. (c) Typical application scenario, illustrating how different kinds of traditional safety systems can be replaced by SafetyEYE.

fences, light barriers, and laser scanners often takes as long as one day. The three-dimensional protection areas of the SafetyEYE system are usually configured within a few hours. An application scenario is shown in Fig. 6.1c. The system is designed to comply with the safety requirements according to the regulations EN 954-1 and SIL 2 EN 62061.

In its first version, the SafetyEYE system aims for preventing hazardous interferences between humans and robots. Hence, although it is more flexible than typical modern safety systems usually consisting of different components, its behaviour is still similar to traditional systems in that its main task is to strictly separate the human from the machine, slowing down and eventually stopping the machine if

a human enters the protection area. For future versions of the SafetyEYE system, it is intended to increase the system capabilities beyond the generation of a three-dimensional point cloud towards a distinction between persons and objects. This will be a step towards collaborative working environments in which persons and machines are able to work simultaneously on the same workpiece.

For such complex scenarios, it is necessary to perform a segmentation of the three-dimensional point cloud and a detection of objects in it, as it is achieved by the method introduced by Schmidt et al. (2007) outlined in Section 1.6.3 (cf. the evaluation in Section 6.2 regarding the industrial production environment). Furthermore, to precisely predict possible collisions between the human and the robot while they are working close to each other, it is necessary to detect parts of the human body, especially the arms and hands of a person working next to the robot, and to determine their three-dimensional trajectories in order to predict collisions with the robot. In Section 6.1.3 we give an overview of existing methods in the field of vision-based gesture recognition. These approaches are certainly useful for the purpose of human–robot interaction, but since most of them are based on two-dimensional image analysis and require an essentially uniform background, they are largely insufficient for safety systems. In contrast, the three-dimensional approaches introduced by Barrois and Wöhler (2008) and Hahn et al. (2007) outlined in Sections 1.6.2 and 1.6.3 yield reasonably accurate results for the pose of the hand–forearm limb independent of the individual person even in the presence of a fairly cluttered background (cf. Sections 6.3 and 6.4).

6.1.3 Pose Estimation of Articulated Objects in the Context of Human–Robot Interaction

The recognition of gestures for human–robot interaction is strongly related to pose estimation of articulated objects. An overview of the most important methods in this field is given in Section 1.6.2. Black and Jepson (1998) model human movements as temporal trajectories of a set of estimated pose parameters over time. These trajectories are characteristic for specific gestures. To match the trajectory models to the multi-variate input data, Black and Jepson (1998) use a random sampling technique based on the CONDENSATION algorithm (Blake and Isard, 1998) for incrementally matching models of the trajectories to the multi-variate input data, such that a gesture can be recognised before it is completed. Prior to the recognition step, the trajectories undergo a normalisation procedure. The trajectory-based approach by Black and Jepson (1998) has inspired many later works. As an example, joint angle trajectories inferred from the monocular body pose estimation system by Schmidt et al. (2006) are utilised by Hofemann (2007) for gesture recognition based on a kernel particle filter method. This approach, which is closely related to the CONDENSATION algorithm, approximates the probability density of the pose parameters as a superposition of Gaussian kernels. Nickel et al. (2004) assign pointing gestures to different classes relying on a hidden Markov model (HMM) approach. In the sys-

tem recently proposed by Hahn et al. (2008a), trajectories are generated with the MOCCD-based approach by Hahn et al. (2007) (cf. Sections 1.6.2 and 6.4). After a normalisation procedure with respect to the length of the trajectory and its position and orientation in space, recognition is performed with a nearest-neighbour approach.

In this section we do not go into detail concerning the problem of gesture recognition itself, but we concentrate on methods for tracking human body parts, especially the arms and hands, over time, which is a prerequisite for recognising gestures in the context of human–robot interaction. The survey by Erol et al. (2007) gives a broad overview of two-dimensional and three-dimensional methods for tracking the human hand. Model-based tracking approaches are initialised at each time step with the pose predicted based on the previously observed dynamic behaviour of the object, inherently assuming a temporal coherence of the motion across the image sequence. Over long image sequences, such techniques tend to become unstable due to imperfections resulting from occlusions and the complexity of the hand motion. This drawback can be alleviated by simultaneously tracking multiple hypotheses. In contrast, single-image pose estimation techniques do not assume temporal coherence of the object motion, leading to an increased computational effort but also to an increased stability of the pose estimation result especially in the presence of rapid motion of the hand and fingers.

6.1.3.1 Geometric Modelling of the Hand

An articulated hand model derived from human anatomy with 27 degrees of freedom is introduced by Lee and Kunii (1993). In their model, the fingers are modelled as planar kinematic chains consisting of rigid elements, which are attached to the palm at anchor points. In an alternative approach by Heap and Hogg (1996), the entire surface of the hand is represented by a deformable model. Hand pose or motion constraints are applied to reduce the search space for pose estimation. While motion models do not strongly depend on the individual human, in many systems the parameters of the geometric model need to be calibrated in a user-specific manner (Lee and Kunii, 1993).

In principle, the hand motion is characterised by many degrees of freedom, e.g. joint angles, but the biomechanically possible configurations of pose parameters lie in a lower-dimensional subspace. A learning-based method by Lin et al. (2000), relying on ground truth data acquired using data gloves, constructs such a subspace by applying a principal component analysis (PCA) to a large amount of joint angle data. Alternatively, the joint angle data can be used to guide the search in the high-dimensional space of pose parameters without dimensionality reduction (Lin et al., 2004). A different way to evaluate the glove data is to generate synthetic images of the hand, building up a set of templates which model the two-dimensional appearance of the hand under many possible poses (Shimada et al., 2001).

6.1.3.2 Kinematic Modelling of the Hand

Learning the dynamics of the hand motion may significantly increase the stability of tracking-based approaches. The eigendynamic analysis approach by Zhou and Huang (2003) models the hand motion in a reduced space of pose parameters determined by PCA. The individual motion patterns of the five fingers are combined to obtain a stochastic linear dynamic system for arbitrary motion of the fingers. The space of hand configurations is represented by Thayananthan et al. (2003a) as a tree constructed by hierarchical clustering techniques or regular partitioning of the eigenspace at multiple resolutions. The nodes of the tree correspond to clusters of hand poses acquired with a data glove, and the tree structure allows to search the pose space efficiently in a hierarchical manner. Relying on large amounts of training data, the dynamic model of the hand is modelled by a first order Markov process describing the state transitions between clusters.

To keep the computational effort moderate, it is favourable to describe the hand by simple geometric primitives (e.g. cylinders, spheres, ellipsoids) attached to each link or joint of the hand model. Stenger et al. (2004) use quadrics as shape primitives, for which fast algorithms are given for projection into the image and determination of visibility based on projective geometry properties. More realistic representations e.g. consist of a B-spline surface with control points attached to the links in the model (Kuch and Huang, 1994) or a deformable skin model (Bray et al., 2004). These methods make use of three-dimensional data acquired with depth sensors or stereo vision, such that no complex projection operations are necessary.

The problem of calibrating and fine-tuning the kinematic hand model with respect to an individual user is performed manually by most systems (Erol et al., 2007). A detailed three-dimensional reconstruction of the hand can be achieved e.g. using three-dimensional sensors (Bray et al., 2004) or three-dimensional reconstruction from multiple viewpoints (Kuch and Huang, 1994). For calibration of kinematic models of the full human body, some systems rely on magnetic sensors which provide data required for multiple rigid motion estimation algorithms (O'Brien et al., 2000; Taycher and Trevor, 2002). Alternatively, active or passive point markers can be used as shown by Kirk et al. (2005). Kuch and Huang (1994) employ a semi-automatic procedure for kinematic calibration of a hand model which relies on reference points on the hand manually extracted from images acquired from different viewpoints. A spline model is then adapted to the reference points. Alternatively, some calibration parameters may be estimated simultaneously with the pose parameters (Thayananthan et al., 2003a). Hu et al. (2004) acquire the three-dimensional locations of the fingertips based on colour LED markers used to calibrate a data glove and a hand model.

6.1.3.3 Two-dimensional vs. Three-dimensional Features

According to Erol et al. (2007), most gesture recognition systems localise the hand based on skin colour, static background subtraction, or adaptive background models,

involving simplifying assumptions such as the hand being the only skin-coloured object, uniform ambient lighting, or stationary background. In some systems, the hand is localised by applying a classifier to subregions of the image. A computationally efficient approach is achieved by employing the boosted cascade technique introduced by Viola and Jones (2001). The classifiers are trained on a large set of images displaying a broad range of hand poses, preferably from different viewpoints. In principle, classification techniques might also be used for hand pose estimation, but labelling a sufficient amount of data is not always feasible.

Many gesture recognition methods rely on the extraction of two-dimensional features, such as the fingertips and the intersection between palm and fingers, from the images. For this purpose, correlation techniques using a circular mask providing rotational invariance (Oka et al., 2002), local curvature maxima of the silhouette boundary (Malik and Laszlo, 2004), or the distance of the contour points to the hand position (Jo et al., 1998) can be used. In many systems, the two-dimensional orientation of the fingers and the hand is estimated by determining the principal axes of the silhouettes. A combination of edge orientation features and Chamfer matching is suggested by Thayanathan et al. (2003b) to increase the robustness in the presence of cluttered background. Models for the skin colour and the background colour may be additionally used to obtain measures for the likelihood of the segmentation.

A basic problem of monocular approaches which exclusively make use of two-dimensional techniques to obtain a low computational complexity is the fact that for a complex articulated object like the hand, a large number of strongly different configurations in the pose space may lead to very similar projections of the articulated object in a monocular image. Three-dimensional data contain valuable information which is able to resolve many problems arising from monocular approaches. A three-dimensional hand model composed of quadrics is used by Stenger et al. (2001) to segment and track the human hand by minimising the geometric error between the model projection and the edges of the hand in the image. An unscented Kalman filter is used for tracking. A method for three-dimensional pose estimation and tracking of the index finger based on range images, colour, and edge features extracted from multiple images of the hand is introduced by Jennings (1999). Three-dimensional tracking of multiple fingers based on cylindrical finger models is achieved by Davis and Shah (1999) for uniform background and in the absence of occlusions. In many systems, a volumetric model is projected into the image to obtain the occluding contours of the projection, and point correspondences between model contours and edges in the image are established e.g. based on the distance to the closest point in the direction perpendicular to the contour. It is straightforward to extend such an approach to several images in order to resolve ambiguities of the three-dimensional pose estimation as described by Gavrila and Davis (1996) in the context of full body pose estimation. Disparity information can be directly used for precisely locating the hand or parts of it (Malik and Laszlo, 2004). Depth data obtained by stereo vision (Delamarre and Faugeras, 2001) or projection of structured light (Bray et al., 2004) allows to segment the hand from the background as long as the hand is the closest object to the camera, where the depth cues are combined with skin colour. The distances between corresponding points on the reconstructed hand

surface and the model surface are used to compute a similarity measure. Alternative approaches to three-dimensional hand pose estimation make use of markers attached to distinct points of the hand such as the fingertips (Kim and Fellner, 2004). These techniques are fairly intrusive but yield higher pose estimation accuracies, thus allowing the distinction of a richer set of gestures.

6.1.3.4 Model-based Hand Tracking

For model-based tracking of an articulated object like the hand, a search in the space of pose parameters needs to be performed to find the configuration of pose parameters which minimises a matching error describing the similarity between model features and features extracted from the images (Erol et al., 2007). This optimisation procedure is initialised with the result of a prediction step which in turn relies on a model of the kinematic behaviour of the object. For the first image of the sequence or in case of failure of the tracking algorithm, an initialisation procedure is applied, performing a search in the space of pose parameters which is not constrained by a prediction step.

Single-hypothesis Tracking

Single-hypothesis tracking approaches are based on a local search in the space of pose parameters, keeping only the best estimate at each time step. However, these techniques tend to fail in the presence of background clutter, temporary occlusions or self-occlusions, and complex motion patterns. The most common approach to fitting the model to the extracted features consists of standard nonlinear optimisation techniques (Press et al., 1992) followed by a Kalman filter stage (Erol et al., 2007). An alternative approach are physical force models, where the matching error is used to compute forces which are exerted on the surface of the articulated model. The pose parameters are then determined by solving the resulting dynamic equations. As an example, Delamarre and Faugeras (2001) derive the force field using the ICP algorithm (cf. Section 1.6.1) to adapt a model to three-dimensional data obtained by stereo image analysis.

Multiple-hypothesis Tracking

Another common tracking approach for articulated objects is multiple-hypothesis tracking. For each time step, the system estimates several hypotheses for the pose parameters, such that the object can still be tracked when the best hypothesis fails. A general framework for multiple hypothesis tracking is Bayesian filtering (Arulampalam et al., 2002), where the a-posteriori probability density function of the pose parameters is computed based on the available observations of the object.

Shimada et al. (1998) apply the technique of multiple-hypothesis tracking with an extended Kalman filter to tracking of hand movements. Ambiguities of the hand pose which are due to the fact that a monocular camera is used appear as singularities of the Jacobian matrix. Upon detection of such a singularity, all possible resulting hypotheses for the hand pose are tracked separately. Another possible implementation of multiple-hypothesis tracking is tree-based filtering (Stenger et al., 2003), where representative templates generated from a large number of images based on an artificial hand model form the nodes of the tree. Traversing the tree yields the probability for the templates during the tracking process, where parts of the tree with small probability masses are skipped to increase processing speed. Shimada et al. (2001) utilise a synthetically generated template database for multiple-hypothesis tracking. From the silhouette contour of the hand, features which are invariant with respect to scale, position, and rotation are extracted. Several hypotheses are kept during tracking, and a search is performed for the neighbourhood of each hypothesis to find the best matching template and generate new hypotheses. Further refinement of the hypothesis is achieved by a local search algorithm. A similar tree-based technique that maximises the overall probability of observation, pose, and transition between subsequent pose configurations is utilised by Nickel et al. (2004).

Probably the most common approach to multiple-hypothesis tracking is the CONDENSATION technique (or particle filtering) introduced by Isard and Blake (1998), which has already been mentioned earlier in this section in the context of classification of temporal trajectories for gesture recognition. The basic concept of this Monte Carlo technique is to represent and maximise the probability density using weighted samples randomly drawn from a probability distribution which is easy to sample (the importance density). The samples are termed particles, and the probability of each sample corresponds to its weight. For tracking, the particles representing the hand pose are updated at each time step. The tracking accuracy increases with increasing number of particles. On the other hand, the computationally most expensive part of a tracking system is the determination of the error function, which is performed once for each particle. Hence, the most important problem of particle filtering is the fact that a large number of samples may be required to sample the probability distribution appropriately. Groß et al. (2006) detect persons in an omnidirectional image based on tracking skin-coloured blobs with the particle filter method by Isard and Blake (1998). A robust extraction of skin-coloured regions is achieved by Fritsch et al. (2004) based on Gaussian mixture models of the colour space. In the VAMPIRE system described by Bauckhage et al. (2005), the colour segmentation results are utilised to detect and track the hands of the user interacting with the system, relying on the method by Isard and Blake (1998). An important advantage of the particle filtering approach is that it provides a well-defined statistical framework for the integration of different cues, e.g. different modalities, into the error function, rendering it a valuable technique for cognitive vision systems. Such systems are regarded in more detail later in this section in the context of the interpretation of pointing gestures and interactions between humans and objects.

The recently introduced method by Tsechpenakis et al. (2008) for tracking the human hand is based on a combination of two pose hypotheses generated with a

discrete shape estimation method and a continuous tracker, respectively. The discrete tracker uses a database containing sequences of object shapes, thus introducing a temporal continuity constraint. The similarity between the input image and a database sample is determined based on the chamfer distance. The continuous tracker relies on two-dimensional edge-driven forces, optical flow, and shading. These features are converted into three-dimensional features using a pinhole camera model in order to compute the velocity, acceleration, and expected position of the hand. Finger movements between consecutive images are predicted based on a dynamic hand model. The discrete and the continuous tracker yield two independent pose hypotheses, where the model is reinitialised when continuous tracking fails. The two-dimensional tracking error corresponds to the chamfer distance between the boundary edges of the input image and those of the projection of the model into the image plane. A three-dimensional error measure is inferred from the two-dimensional error measure by learning the corresponding relation from examples using support vector regression. Tsechpenakis et al. (2008) apply their hand tracking approach to American sign language recognition.

6.1.3.5 Single-image Pose Estimation

The process of estimating the pose of an object based on a monocular image or a set of images acquired from different viewpoints without relying on the motion history of the object is termed single-image pose estimation. Essentially, this approach corresponds to a global optimisation in the complete space of pose parameters. In the context of gesture recognition, a single-image pose estimation stage is required for initialisation of the tracking algorithm – otherwise, the user would have to place the hand into a well-defined initial position and orientation, which is, depending on the application, not always feasible. Beyond initialisation, single-image pose estimation may help to avoid a complex tracking stage in a gesture recognition system. A motivation for the single-image approach is the ability of the hand to perform movements with high (positive and negative) accelerations, which are fairly often the reason for tracking failures. All methods for pose estimation of articulated objects described in Section 1.6.2 and also in Section 5.2 can in principle be used as well for gesture recognition in the context of human–robot interaction. A method especially suitable for hand pose estimation is inverse kinematics, where model constraints are exploited to regularise the problem of inferring the hand pose from the features extracted from the image, such as the fingertip locations. For example, Chua et al. (2000) derive closed-form solutions for the joint angles of a hand model, given the two-dimensional image positions of markers under orthographic projection. In the system suggested by Nölker and Ritter (1999), a parameterised self-organising map (PSOM) is used to learn from examples the mapping from a set of two-dimensional fingertip positions to a set of joint angles.

6.1.3.6 Including Context Information: Pointing Gestures and Interactions with Objects

Gesture recognition algorithms are an important part of cognitive systems, which communicate or interact with their environment and users rather than merely acquire information about their surrounding (Bauckhage et al., 2005; Wachsmuth et al., 2005). In this context, the accurate determination of the pointing gesture as well as interactions between humans and objects are essential issues.

To estimate the pointing direction, Nickel et al. (2004) propose to combine the direction of the hand-forearm line and the head orientation to accurately and robustly infer the object at which the user is pointing. The two-dimensional positions of the head and the hand are extracted based on skin colour analysis. They demonstrate that the highest rate of correctly identified objects is achieved relying on the direction of the line connecting the head and the pointing hand. Pointing gestures are assigned to different classes based on a dedicated hidden Markov model (HMM) architecture. In the mobile robotic system by Groß et al. (2006), a multi-target tracker based on the particle filter approach according to Isard and Blake (1998) is used in combination with the Adaboost face detector (Viola and Jones, 2001) and cylindric models fitted to depth data acquired with a laser range finder. The images are acquired with a fisheye camera and a standard camera. The positions of persons close to the robot are modelled by a mixture of Gaussians, leading to a Bayesian framework which allows to integrate all available sensor information in order to accept or reject object hypotheses. Upon detection of a person, the system attempts to recognise a deictic gesture and to eventually infer the pointing direction from the available monocular image data. Based on the head position, a region of interest containing the upper part of the body is divided into a subregion containing the head and another sub-region containing the pointing arm. The pointing direction of the arm is learned from a large set of example images based on a classifier cascade consisting of multilayer perceptrons with output neurons encoding the distance to the object at which the user points along with the direction in which it is situated. An average angular accuracy of typically about 10° is obtained for the estimated pointing direction.

In their cognitive vision system, Bauckhage et al. (2005) also use an extension of the particle filter approach by Isard and Blake (1998) to extract and classify hand trajectories. The situational context of the gesture, describing its preconditions and the effect it has on the scene, and the spatial context, i.e. the spatial relations between hand trajectories and objects in the scene, are incorporated into this framework according to the approach introduced by Fritsch et al. (2004). A context area in the image is given by a circle segment defining the part of the scene which is potentially relevant for a specific gesture. While the context area has an absolute orientation for objects characterised by a specific handling direction (e.g. a cup), it is oriented parallel to the motion direction of the hand otherwise. The situational context is taken into account by restricting the samples of the particle filter to those fulfilling the preconditions of the current situation. The spatial context influences the update step of the particle filter due to its influence on the weights of samples

that match the observations, depending on how well the observed scene corresponds to the expected symbolic context.

6.1.3.7 Discussion in the Context of Industrial Safety Systems

The previously described hand gesture recognition systems are able to detect and track the moving hand largely in laboratory environments. Especially the more recent approaches are able to cope with cluttered background. Methods have been presented to recognise the meaning of observed gestures, and it has been shown that it is possible to incorporate contextual information such as relations between the hand and objects in the scene.

Many of the systems described above rely on a detection of the hand based on skin colour. An important drawback of such approaches is the fact that colour-based segmentation tends to be unstable in the presence of variable illumination, especially changing spectral characteristics of the incident light. Although adaptively modelling the colour distribution in the image may partially solve this problem, skin colour cues are not suitable for detecting humans in the industrial production environment since in many domains of industrial production dedicated work clothes such as gloves are obligatory. An industrial safety system thus needs to be able to recognise the human body or parts of it based on shape cues alone. Furthermore, the recognition performance should not decrease in the presence of a cluttered background displaying arbitrary colours. It is thus favourable to abandon the assumption of a specific colour distribution separating relevant from background objects.

Most gesture recognition systems described above require a user-dependent configuration and calibration stage. An industrial safety system, however, needs to be able to detect and localise arbitrary humans without relying on any kind of a priori knowledge about them as a large number of persons may have access to the workspace to be safeguarded by the system. As a consequence, we will see in the following sections that it is favourable to utilise person-independent models of the human body and its parts rather than attempting to adapt a highly accurate person-specific model to the observations.

In the systems described above, multiple-hypothesis tracking is a fairly common technique, but most existing systems merely evaluate two-dimensional cues. In many cases this is sufficient for interpreting gestures for enabling an interaction between a human and a robotic system, but for an industrial system for safeguarding large workspaces it is required to acquire three-dimensional data of the scene in order to be able to reliably avoid collisions or other hazardous interferences between the human and the machine.

In the following sections of this chapter, the issue of user-independent three-dimensional detection and tracking of persons and human body parts not involving skin colour cues is addressed in the context of safe human–robot interaction in the industrial production environment.

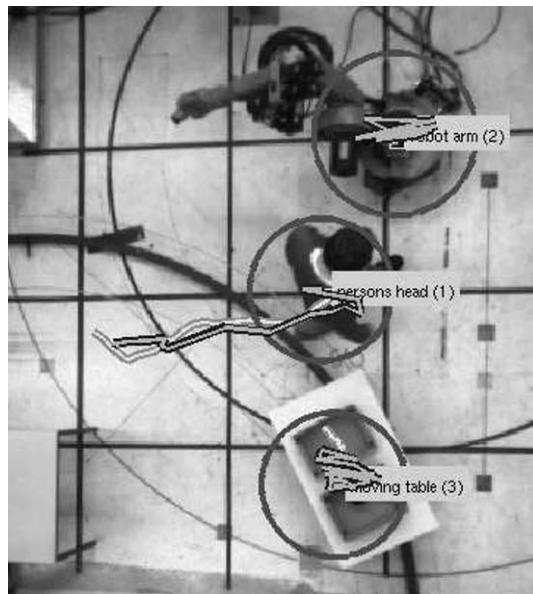


Fig. 6.2 Trajectories of tracked objects (dark grey) with annotated ground truth (light grey) for a typical industrial scene.

6.2 Object Detection and Tracking in Three-dimensional Point Clouds

This section describes the evaluation of the method for three-dimensional detection and tracking of persons in three-dimensional point clouds introduced by Schmidt et al. (2007) described in Section 1.6.3 in an industrial production scenario, involving a workspace with a human worker, a robot, and a moving platform. We utilised stereo image sequences recorded with a Digiclops CCD camera system with an image size of 640×480 pixels, a pixel size of $7.4 \mu\text{m}$, and a focal length of 4 mm. The stereo baseline corresponds to 100 mm. The average distance to the scene amounts to 5.65 m.

A three-dimensional point cloud of the scene is generated by combining the correlation-based stereo technique by Franke and Joos (2000) and the spacetime stereo approach outlined in Section 1.5.2.4 as described in Section 1.6.3. We empirically found for the correlation matrix element Σ_z in Eq. (1.153) the value $\Sigma_z = 0.292$, regarding a set of three-dimensional points obtained with the spacetime stereo algorithm and belonging to a plane scene part, while $\Sigma_x = \Sigma_y = 1$ are equally scaled. The velocity scaling factor is set to $\rho = 380$ s, where the velocity is expressed in metres per second and the spatial coordinates in metres. The kernel widths for Eq. (1.154) are chosen as $H_{r,\max} = 1.88$ m, $H_{r,\min} = 0.135$ m, $H_d = 0.113$ m, and $H_v = 0.188$ m for the industrial scenes. While the object sizes as well as the overall size of the

Table 6.1 Tracking results compared to ground truth. The RMSE values are given in mm.

sequence	# images	object	with velocity		without velocity	
			RMSE	tracked fraction (%)	RMSE	tracked fraction (%)
industrial 1	69	person	265	100.0	383	84.8
		table	603	100.0	218	69.7
		robot	878	95.5	1118	98.5
industrial 2	79	person	427	100.0	318	94.8
		table	435	100.0	275	100.0
		robot	121	98.7	177	96.1
industrial 3	24	person	196	100.0	147	100.0
		table	249	100.0	225	90.9
		robot	171	100.0	293	100.0
industrial 4	39	person	247	75.7	352	89.2
		table	270	100.0	245	97.3
		robot	91	100.0	200	97.3
industrial 5	24	person	208	90.9	254	81.8
		table	219	100.0	329	100.0
		robot	86	77.3	331	100.0

scenes are far different, the kernel widths merely need to be scaled by an empirical uniform factor, such that the relative parameter values remain constant. The value of ρ depends on the typical velocities encountered in the scene. Hence, we set for the tabletop scene $H_{r,\max} = 4.14$ m, $H_{r,\min} = 0.297$ m, $H_d = 0.249$ m, $H_v = 0.414$ m, and $\rho = 3200$ s. For each sequence, ground truth was generated manually by marking the center of the objects of interest in each frame, e.g. the head of the person or the center of the car, and transforming them into three-dimensional coordinates using the known geometry of the scene and the objects, e.g. the height of the person and the position of the ground plane. The trajectories of the tracked objects are compared to the ground truth based on the corresponding value of the RMSE in world coordinates. Fig. 6.2 shows a typical tracking result achieved by our system. The results in Table 6.1 illustrate that objects can be tracked in a stable manner at reasonable accuracy. Using epipolar velocity as an additional feature yields a more accurate localisation result for 10 of 16 detected objects, and detection is usually possible in a larger fraction of the frames. For four other objects the RMSE but at the same time also the detection rate is lower when the velocity information is neglected. The system is designed to segment the point cloud into clusters of differing velocity. As a consequence, the proposed system works best for objects with homogeneous velocity. For example, we observed that for a walking person moving the arms backwards the object hypothesis does not include the arms due to their velocity.

The experimental evaluation shows that the proposed method is a promising technique for applications in the field of human–robot interaction in industrial production scenarios. It does not rely on colour cues, and the objects in the scene are detected and tracked based on a weak model (a cylinder) which does not require any a-priori knowledge about the objects.

6.3 Detection and Spatio-temporal Pose Estimation of Human Body Parts

The method for model-based object detection and spatio-temporal three-dimensional pose estimation described in Section 1.6.2 (Barrois and Wöhler, 2008) is evaluated by analysing three realistic image sequences displaying a hand-forearm limb moving at non-uniform speed in front of complex cluttered background (cf. Figs. 1.40 and 6.3). The distance of the hand-forearm limb to the camera amounts to 0.85–1.75 m, the image resolution to 2–3 mm per pixel. For acquisition of stereo image pairs a Digiclops CCD camera system is used. The time step between subsequent image pairs amounts to $\Delta t = 50$ ms. Spacetime stereo information according to Section 1.5.2.4 is determined based on triples of subsequent stereo image pairs, where we have set $\delta_{\max} = 2$. Ground truth information has been determined based on three markers attached to points located on the upper forearm (${}^W\mathbf{p}_1$), the wrist (${}^W\mathbf{p}_2$), and the front of the hand (${}^W\mathbf{p}_3$) as depicted in Fig. 6.5. The three-dimensional coordinates of the markers were determined by bundle adjustment.

For each of the two parts of the hand-forearm model, the corresponding 5 translational and rotational pose parameters (denoted by the vector \mathbf{T}) are determined independently. For the evaluation, our method is employed as a “tracking by detection” system, i.e. the pose $\mathbf{T}(t)$ and the pose derivative $\dot{\mathbf{T}}(t)$ are used to compute a pose $\mathbf{T}_{\text{init}}(t + n \Delta t) = \mathbf{T}(t) + \mathbf{T}(t) \cdot (n \Delta t)$ for the next time step ($t + n \Delta t$) at which a model adaptation is performed. The pose $\mathbf{T}_{\text{init}}(t + n \Delta t)$ is used as an initialisation for the model adaption procedure. The temporal derivatives of the pose parameters are determined independently for each model part, where the translational motion is constrained such that the two model parts are connected with each other at the point ${}^W\mathbf{p}_2$.

For each sequence, the evaluation is performed for different values of n . The highest regarded value is $n = 16$, corresponding to an effective rate of only 1.25 frames per second. As a measure for the accuracy of the estimated pose, the mean Euclidean distances in the scene between the estimated and the ground truth positions are shown for ${}^W\mathbf{p}_1$ (circles), ${}^W\mathbf{p}_2$ (squares), and ${}^W\mathbf{p}_3$ (diamonds) column-wise for the three test sequences in the first row of Fig. 6.4. The second row shows the mean errors and standard deviations of the translational motion components U_{obj} (circles), V_{obj} (squares), and W_{obj} (diamonds) per time step. For each value of n , the left triple of points denotes the forearm and the right triple the hand. The third row displays the mean errors and standard deviations of the rotational motion components ω_p (circles) and ω_o (squares). For each value of n , the left pair of points denotes the forearm and the right pair the hand.

The Euclidean distances between the estimated and true reference points typically amount to 40–80 mm and become as large as 150 mm in the third sequence, which displays a pointing gesture (cf. Fig. 6.3). Being independent of n , the values are comparable to those reported by Hahn et al. (2007). The deviations measured for our system are comparable to the translational accuracy of about 70 mm achieved by the stereo-based upper body tracking system proposed by Ziegler et al. (2006).

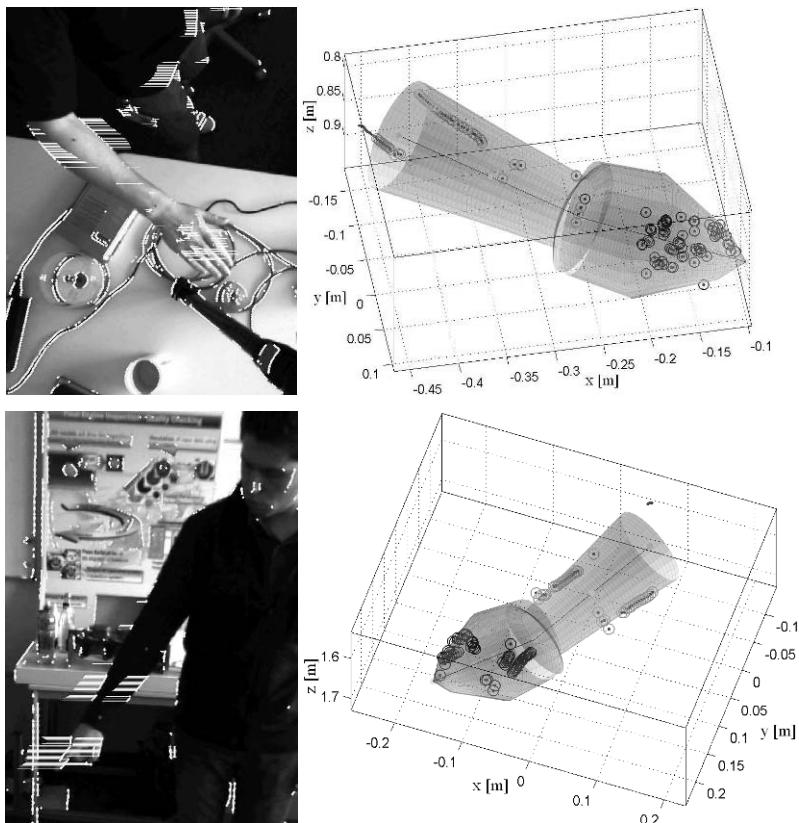


Fig. 6.3 Spacetime stereo and model adaptation results for example images from the second (top) and the third (bottom) test sequence. Three-dimensional points are indicated by bright dots, epipolar velocities by white lines.

Furthermore, the average joint angle error of 25.7° determined in that work for the lower arm is equivalent to a translational error of about 100 mm, using our assumed forearm length of 220 mm, which is slightly higher than but comparable to the deviations obtained for our system.

The discrepancies observed for our system are to some extent caused by a shift of the model along the longitudinal object axis but also by the fact that the lengths of the partial models of the forearm and the hand are fixed to 220 and 180 mm, respectively. For the three sequences, the true lengths correspond to 190, 190, and 217 mm for the forearm and to 203, 193, and 128 mm for the hand. Especially in the third sequence the hand posture is poorly represented by the model, as the hand forms a fist with the index finger pointing. However, this does not affect the robustness of the system. Our evaluation furthermore shows that the motion between two subsequent images is estimated at a typical translational accuracy of 1–3 mm, which is comparable to the pixel resolution of the images, and a typical rotational accuracy

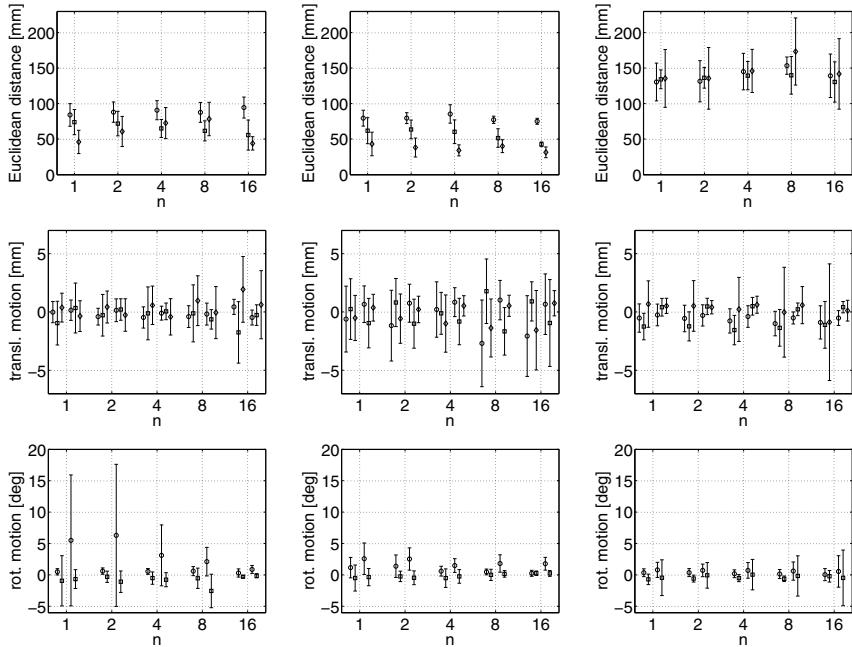


Fig. 6.4 Evaluation results, columnwise displayed for the three test sequences. “Motion” refers to the pose variation between subsequent images.

of 1–3 degrees. Due to its roundish shape, the rotational motion of the hand is estimated less accurately than that of the more elongated forearm (the very large errors observed for ω_p in the first sequence are due to sporadically occurring outliers). A robust detection and pose estimation of the hand-forearm limb is achieved for time intervals between subsequent model adaptations as long as 800 ms ($n = 16$). The accuracy of the estimated pose and its temporal derivative is largely independent of n .

The proposed spatio-temporal pose estimation method relies on a new extended constraint line approach introduced to directly infer the translational and rotational motion components of the objects based on the low-level spacetime information. Our evaluation of this approach in a “tracking by detection” framework has demonstrated that a robust and accurate estimation of the three-dimensional object pose and its temporal derivative is achieved in the presence of cluttered background without requiring an initial pose.



Fig. 6.5 Determination of the ground truth.

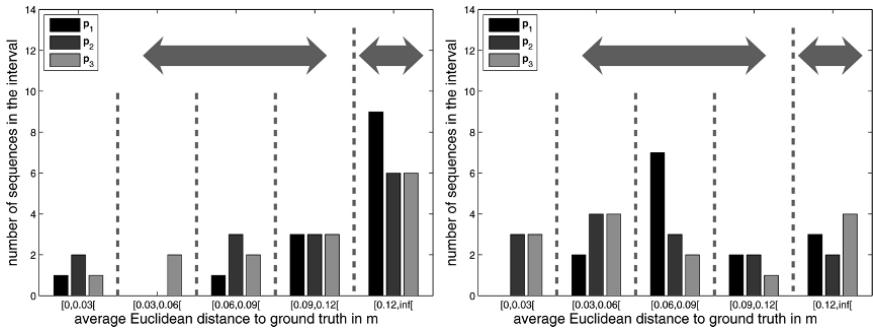


Fig. 6.6 Left: Results of a single Kalman filter with a constant-velocity model. Right: Results of two Kalman filters with constant-acceleration and constant-velocity models.

6.4 Three-dimensional Tracking of Human Body Parts

This section describes a quantitative evaluation of the system introduced by Hahn et al. (2007) for three-dimensional tracking of the human hand-forearm limb. The evaluation is performed on 14 real-world test sequences. These sequences contain movements of different hand-forearm configurations of four different test persons in front of complex cluttered background. Each sequence contains at least 100 image triples. For acquisition of image triples, we utilised a Digiclops CCD camera system with an image size of 640×480 pixels, a pixel size of $7.4 \mu\text{m}$, and a focal length of 4 mm. The stereo baseline corresponds to 100 mm. The mean distance of the test persons to the camera system varies from 0.85 m to 1.75 m. The ground truth was obtained based on three points in the world coordinate system (Fig. 6.5). These three points correspond to the points wP_1 , wP_2 , and wP_3 of the three-dimensional model of the human hand-forearm limb according to Eq. (1.146). To compute the ground truth, three markers were fixed to the tracked limb. These markers were man-

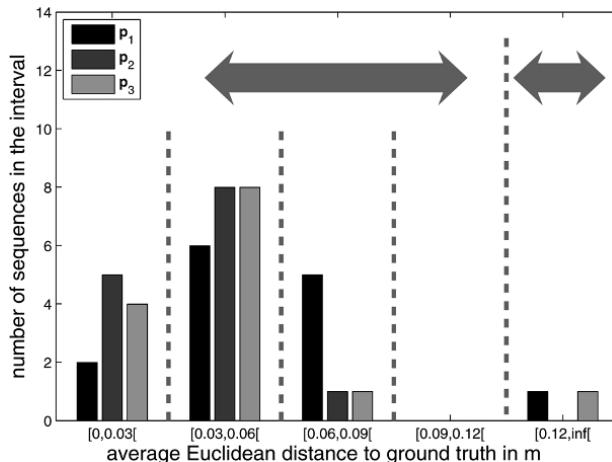


Fig. 6.7 Results of three Kalman filters representing constant-acceleration, constant-velocity, and constant-position models.

ually labelled in the three images of the camera system and the three-dimensional coordinates were computed by bundle adjustment.

The results of a system variant are depicted as a histogram (Fig. 6.6). We computed the Euclidean distance to the ground truth at every time step for the estimated points ${}^W\mathbf{p}_1$, ${}^W\mathbf{p}_2$, and ${}^W\mathbf{p}_3$ of the three-dimensional model. The mean Euclidean distance of the three points to the ground truth for a single sequence is computed, and the results of all sequences are arranged as a histogram. The abscissa of the histogram is divided from the left to the right into five intervals:

1. very good: $[0, 0.03[\text{ m}$
2. good: $[0.03, 0.06[\text{ m}$
3. satisfactory: $[0.06, 0.09[\text{ m}$
4. sufficient: $[0.09, 0.12[\text{ m}$
5. insufficient: $[0.12, \infty[\text{ m.}$

The choice of the intervals is justified by the context of safe human–robot interaction. A distance of 0.12 m is the upper limit, as this range can be covered by sensors on the robot triggering an emergency stop. On the ordinate the number of sequences whose mean Euclidean distance to the ground truth for the respective point falls into the respective histogram bin is depicted.

The histogram of the system variant using a single Kalman filter with a constant-velocity model is depicted in Fig. 6.6 (left). Many sequences fall into the “insufficient” histogram bin. This behaviour is due to false predictions of the Kalman filter during motion reversal, e.g. when the test person tightens a screw. The false predictions result from the inadequate motion model, leading to poor convergence of the MOCCD algorithm during the subsequent time steps. If the misprediction exceeds the convergence radius of the MOCCD, the MOCCD cannot compensate the too

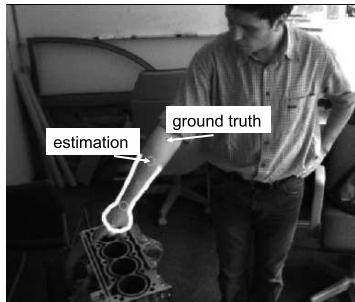


Fig. 6.8 Image from a sequence in which the hand–forearm limb is tracked in a stable but somewhat inaccurate manner.

large distance between the depicted object and its assumed position. The use of a multi-hypothesis tracker consisting of two Kalman filters with constant-acceleration and constant-velocity kinematic models improves the system performance (cf. right diagram of Fig. 6.6), as only a fifth of all sequences falls into the “insufficient” interval. The measurement selection component is able to recognise false predictions and therefore the tracking is more stable than with only one Kalman filter.

The use of three Kalman filters with the different kinematic models constant-acceleration, constant-velocity, and constant-position leads to a further improvement (cf. Fig. 6.7) and is the best system variant obtained in this investigation. Only two sequences fall into the “insufficient” interval and most of the sequences fall into the “good” histogram bin. This yields an error rate of about 5 percent. The remaining 95 percent of the sequence have a maximal error of less than 90 mm. This is one fourth better than required in our system concept and will lead to a better stability of the overall system as the close range sensors will rarely have to stop the robot. Fig. 6.10 shows some short scenes taken from the sequences and the corresponding result of the best system variant.

In the result histogram of the best system variant shown in Fig. 6.7 it is apparent that two sequences still fall into the “insufficient” histogram bin. The reason is the coarse symmetric model of the hand–forearm limb. The arising silhouette of the pointing hand can not be represented by the present hand–forearm model. The system estimates the elbow point ${}^W\mathbf{p}_1$ too far towards the hand (cf. Fig. 6.8) as the overall scaling of the arm depends on the scaling of the hand. The estimated model likelihood is nearly constant when shifting the outline of the forearm model along the depicted forearm. Thus it does not stop the MOCCD algorithm from shrinking the model. However, even in presence of such input images violating the assumed geometry of the model the tracking does not fail for the whole sequence.

With another experiment we examined the importance of using three cameras. Although the system is still capable of performing triangulation, the performance decreases as about 50 percent of the sequences fall into the “insufficient” interval. The reason is that the MOCCD is strongly affected by the aperture problem in one direction due to the elongated shape of the hand–forearm limb. With the use of the

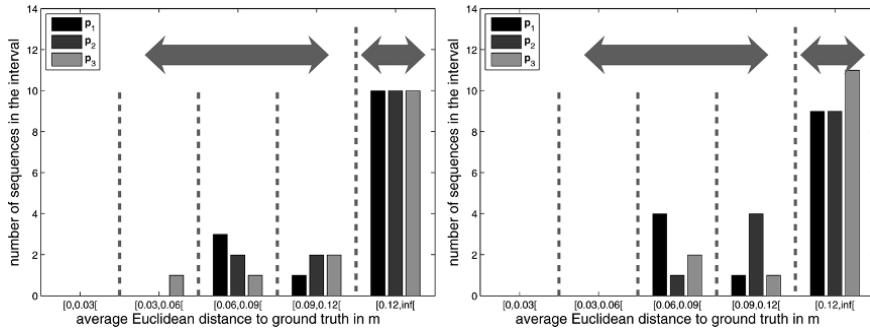


Fig. 6.9 Left: Results of the particle filter based curve model approach (Isard and Blake, 1998). Right: Results of the three-dimensional active contour approach (d'Angelo et al., 2004).

third camera in Fig. 6.7 the ambiguities are avoided, resulting in a good triangulation in both directions.

To compare our tracking system we applied two available approaches – a curve model tracked by a particle filter (Isard and Blake, 1998) and the three-dimensional active contour approach described in Section 1.6.2 (d'Angelo et al., 2004) – which both yielded good results in former applications. The particle filter based approach by Isard and Blake (1998) was extended to use three-dimensional measurements by projecting a three-dimensional model curve to the three images. Furthermore we extended the three-dimensional active contours by d'Angelo et al. (2004) by a tracking system using the same three Kalman filter framework as for the MOCCD. Processed with the particle filter based approach (cf. Fig. 6.9), about 70 percent of the sequences fall into the “insufficient” interval. Although we use 10000 particles at nine degrees of freedom, the algorithm gets stuck at edges due to the cluttered background and shading-related edges. This result illustrates the higher robustness of the MOCCD due to the fact that it adapts to the spatial scale of an edge. Analysing the sequences using three-dimensional active contours (cf. Fig. 6.9) yields a similar error rate of about 70 percent of the sequences in the “insufficient” histogram bin. The algorithm gets stuck in background and shading-related edges that have a sufficient strength. Again the adaptive behaviour of the MOCCD proves to be superior compared to edge extraction methods on fixed scales, regardless of the subsequent recognition method.

Our experiments with the 14 real-world test sequences have shown that the use of three MOCCD algorithms with three different kinematic models within the tracking system leads to a reasonably accurate and temporally stable system. Only coarse information about the test person is necessary: the lengths of forearm and hand as well as the initial position. The test sequences include the following hand-forearm configurations: an outstretched hand, a fist, a hand holding a cup, a hand holding a wrench, and a pointing hand. It is possible to track the motion of different hand-forearm configurations such as reversing motion, e.g. tightening of a screw, movement of the hand in depth with a constant forearm position, lateral movement of



Fig. 6.10 Results of different test sequences computed with three Kalman filters and the different kinematic models constant acceleration, constant velocity, and constant position. The three-dimensional models corresponding to the estimated pose parameters are shown in the second, fourth, and sixth column.

the complete hand–forearm limb, and pointing gestures. The system is able to track these motion patterns in a temporally stable manner at an accuracy high enough to make the method suitable for human–robot interaction in the industrial environment.

A first evaluation of the spatio-temporal shape flow algorithm by Hahn et al. (2008b) on five test sequences of trinocular colour images indicates that the pose estimation accuracy of the shape flow method amounts to 40–100 mm for the average Euclidean distance between the reference points of the hand–forearm model and the corresponding ground truth points, thus being comparable to the MOCCD algorithm and the spatio-temporal ICP-based approach by Barrois and Wöhler (2008) (cf. Sections 1.6.2 and 6.3). Typical RMSE values of the estimated velocities of the reference points are 4–6 mm per time step for the lateral velocity components and 4–9 mm per time step for the velocity component parallel to the depth axis. These values are about 2–3 times higher than the effective lateral resolution of 1.5–3 mm per pixel on the object.

For comparison, state-of-the-art multiple-view approaches such as the method by Brox et al. (2008) achieve somewhat higher metric accuracies than the MOCCD and

the shape flow algorithm, but they require person-specific detailed modelling, more cameras, wider baselines, and less complex backgrounds.

Chapter 7

Applications to Lunar Remote Sensing

This chapter addresses applications in the domain of remote sensing, especially the generation of elevation maps of the lunar surface. In contrast to e.g. the planet Mars, which has been topographically mapped at high resolution by laser altimetry by the Mars Global Surveyor spacecraft in the 1990s and more recently by stereophotogrammetry based on high-resolution imagery of the European spacecraft Mars Express, for only about a quarter of the lunar surface fairly accurate topographic data are available, which date back to the Apollo era of the late 1960s and early 1970s. On large scales, the lunar surface has been topographically mapped by the Clementine spacecraft in 1994 based on laser altimetre (LIDAR) measurements, but both the lateral resolution and the vertical accuracy of this data set is fairly low (cf. Section 7.1). While the LIDAR measurements clearly show large-scale structures such as the previously unknown south pole Aitken basin, they do not reveal subtle small-scale details on the floors of lunar craters or structures of non-impact (endogenic) origin such as ridges, tectonic faults, or volcanic domes. Hence, for large parts of the lunar surface no small-scale topographic data are available, with the exception of crater depths and mountain heights derived from image-based shadow length measurements.

The application of three-dimensional computer vision methods to the domain of lunar remote sensing addressed in this chapter concentrates on the derivation of digital elevation maps of small parts of the lunar surface at high lateral resolution and high vertical accuracy. Section 7.1 provides a general overview of existing methods used for constructing elevation maps of planetary bodies and the corresponding data sets. The three-dimensional reconstruction of lunar craters at high resolution, i.e. beyond a determination of their depth and rim height, is regarded in Section 7.2, while Section 7.3 discusses the three-dimensional reconstruction of lunar wrinkle ridges and tectonic faults. Section 7.4 describes the generation of digital elevation maps of subtle volcanic features on the Moon, especially lunar domes. It provides the first collection to date of reasonably accurate topographic data for a representative set of lunar domes. Hence, a novel classification scheme for these volcanic objects is established, which is based on the topographic measurements as well as spectrophotometric data. To place these activities into the context of planetary science, a brief

outline of geologic insights that can be gained from topographic measurements of lunar domes is given.

7.1 Three-dimensional Surface Reconstruction Methods for Planetary Remote Sensing

A general overview of activities in the field of topographic mapping of planetary bodies in the inner and the outer solar system is provided in Section 7.1.1.1. The utilised methods can be divided into active approaches, shadow length measurements, classical photogrammetric techniques, and photoclinometric approaches. Section 7.1.2 describes how the reflectance behaviour of planetary regolith surfaces encountered for bodies without an atmosphere in the inner solar system, i.e. Mercury, the Moon, and the asteroids, is modelled in the context of three-dimensional reconstruction of such surfaces based on photometric methods. These models are a basis for the work about topographic mapping of the lunar surface described later in this chapter in Sections 7.2–7.4.

7.1.1 *Topographic Mapping of Solar System Bodies*

7.1.1.1 Active Methods

Three-dimensional reconstruction of planetary surfaces can either be performed by active or by passive methods. Active methods mainly involve radar or laser altimetry. Ground-based radar observations of Mercury, Venus, and Mars were performed for the first time in the early 1970s, providing surface cross-sections with vertical accuracies of 100–150 m (Goldstein et al., 1970; Downs et al., 1971, 1975; Harmon and Campbell, 1988; Wu and Doyle, 1990). From the orbit, the three-dimensional surface profile of Venus has been explored by radar measurements of the Pioneer Venus Orbiter (Pettengill et al., 1980). The laser altimetres on the Apollo 15–17 spacecrafts provided lunar topographic data for surface points at a high accuracy of 2 m (Wu and Doyle, 1990). The lunar surface has been globally mapped at a lateral resolution of 0.25° in longitude and latitude, i.e. better than 7.5 km, by laser altimetry from the Clementine spacecraft (Neumann, 2001; Bussey and Spudis, 2004). The surface of Mars was mapped almost entirely at a lateral resolution of 230 m by the MOLA (Mars Orbiter Laser Altimetre) instrument on the Mars Global Surveyor Orbiter spacecraft (Abshire et al., 2000). It is, however, impossible to reveal by means of active sensing methods small-scale structures that can easily be resolved from the same observing position by traditional two-dimensional imaging methods.

7.1.1.2 Shadow Length Measurements

The classical image-based passive approach to determine height differences on planetary surfaces is the measurement of shadow lengths. This method dates back to the year 1609, when Galileo Galilei performed his first telescopic observations of the Moon (Whitaker, 1999). Modern shadow length measurements have been systematically performed using spacecraft imagery of the Moon (Wood, 1973; Wood and Andersson, 1978), Mercury (Pike, 1988), and Mars (Cintala et al., 1976) to determine the depths and rim heights of craters and the heights of their central peaks.

7.1.1.3 Stereo and Multi-Image Photogrammetry

Elevation maps of about a quarter of the lunar surface are provided by the lunar topographic orthophotomap (LTO) series generated based on orbital imagery acquired by the Apollo 15–17 command modules with modified aerial cameras, where the spacecraft motion allowed the acquisition of images under multiple views of the surface. These topographic maps were computed based on stereophotogrammetry and represent the highest-resolution lunar topographic data currently available with elevation standard errors of 30 m (Wu and Doyle, 1990). A three-dimensional reconstruction of lunar surface regions not covered by the LTO series is performed by Cook et al. (1999) based on correlation-based stereoscopic analysis of images acquired by the Lunar Orbiter and Clementine spacecraft from the orbit around the Moon. Since large correlation masks have to be used to overcome ambiguities of the correspondence problem, the lateral resolution of the obtained surface profiles is not better than 1 km, and the accuracy of the measured elevation values is of the order of 100 m.

Topographic maps of the surface of Mars derived from Viking imagery in a similar manner have vertical accuracies of 30–60 m (Wu et al., 1982). Local elevation maps obtained by stereoscopic evaluation of Viking lander imagery reach accuracies between 0.1 and 0.7 m (Wu and Doyle, 1990). More recent efforts to generate a global topographic map of Mars are based on images of the High Resolution Stereo Camera (HRSC), a line scan camera installed on the Mars Express spacecraft. A software system to generate a global topographic map of scale 1 : 200000 is presented by Gehrke et al. (2006), where also larger scales of 1 : 100000 and 1 : 50000 can be obtained if images of sufficiently high resolution are available. The mapping system is demonstrated based on the region Iani Chaos, for which images of a lateral resolution of 12 metres per pixel are available. An in-depth discussion of the HRSC experiment is given by Jaumann et al. (2007). The dense stereo algorithm based on semi-global matching introduced by Hirschmüller (2006), which has originally been developed for the three-dimensional reconstruction of surfaces with pronounced depth discontinuities from satellite images, such as urban areas with many buildings (cf. Section 1.5), is applied to HRSC images by Hirschmüller et al. (2007). A comparison between the inferred topographic maps and MOLA data yields a mean difference of 9 m and a root mean square deviation of 51 m at a lat-

eral image resolution of 15 m. A lower image quality leads to deviations which are about five times as high. According to Hirschmüller et al. (2007), the main advantage of the semi-global matching algorithm compared to previous techniques for establishing stereo correspondences between HRSC images is its high computational efficiency.

Regarding bodies of the outer solar system, images acquired by the Voyager 2 spacecraft during its Uranus flyby were utilised by Wu et al. (1987) to generate a topographic map of Miranda, a moon of Uranus with a diameter of 470 km. A topographic map of the large volcano Tvashtar Patera on Io, the innermost Galilean moon of Jupiter, is inferred by Turtle et al. (2007) from high-resolution imagery taken by the Galileo spacecraft.

7.1.1.4 Photoclinometry and Shape from Shading

While the lateral resolution of a topographic map generated by stereophotogrammetric techniques is generally significantly lower than the pixel resolution of the images from which it has been produced, the technique of photoclinometry described in Section 2.2.2.1 can be used to generate cross-sectional profiles of planetary surfaces at a resolution comparable to that of the available images. It has been pointed out in Section 2.2.2.1 that photoclinometric methods yield a fairly accurate representation of depth differences on small scales but tend to produce systematic depth errors on scales much larger than the pixel resolution. Early work by Wilhelms (1964) describes the photoclinometric determination of the statistical distribution of small-scale slopes on the lunar surface, using telescopic photographic lunar images. Similarly, photoclinometry is used by Mousginis-Mark and Wilson (1979) to measure crater depths and hill heights on Mercury, relying on Mariner 10 images. Wildey (1975) applies a photoclinometry technique to Mariner 9 images of the surface of Mars that takes into account not only the reflectance behaviour of the surface but also the scattering behaviour of the Martian atmosphere.

More recent work by Fenton and Herkenhoff (2000) describes the determination of topography and stratigraphy of the northern Martian polar layered deposits by a combined approach that takes into account photoclinometric, stereophotogrammetric, and laser altimetry data. Beyer and McEwen (2002) use a photoclinometric technique to examine the suitability of selected regions on the Martian surface as landing sites for rover missions by determining surface gradients on metre scales. An estimation of the topographic surface roughness of the north polar residual ice cap of Mars in order to examine its suitability as a landing site is performed by Herkenhoff et al. (2002) by determining the statistical distribution of slope angles inferred by photoclinometry. A multiple-image method termed multi-image shape from shading is used by Lohse et al. (2006) to construct topographic maps of the lunar surface based on Clementine images. Their method directly relates image grey values to the heights of the topographic map and the parameters of the photometric model which describes the surface reflectance behaviour. The height values, which are determined directly in object space, are estimated along with the photometric

parameters from the image grey values in a least-mean-squares adjustment, where a uniform surface albedo is assumed. Their method works best for images acquired under an oblique viewing geometry. A similar approach is used by Gaskell et al. (2007) to construct a full three-dimensional geometric model of the asteroid Eros, relying on images acquired by the Near Earth Asteroid Rendezvous (NEAR) space-craft.

A three-dimensional reconstruction of ridges on the icy surface of Europa, one of the Galilean moons of Jupiter, is performed by Cordero-Tercero and Mendoza-Ortega (2001) using a photoclinometric technique. Schenk and Pappalardo (2002) generate topographic maps for regions of Europa's surface based on a combined approach. Topography on kilometre scales is determined using standard stereophotogrammetry, while pixel-scale topography is obtained with a photoclinometric technique. Their method allows to take into account a non-uniform surface albedo, relying on images of the same regions acquired at low phase angles. The stereophotogrammetric data are used to control the large-scale topography of the highly resolved elevation map obtained by photoclinometry.

7.1.2 Reflectance Behaviour of Planetary Regolith Surfaces

When the inner planets of the solar system formed, all of them underwent internal differentiation processes which led to the formation of a core mainly consisting of iron and nickel, a mantle consisting of dense silicate rock, and a silicic crust. In the absence of an atmosphere, the rocky surfaces were shattered by meteorite impacts over billions of years. At the same time, they were eroded in a less spectacular but also fairly efficient manner by the particles of the solar wind, e.g. highly energetic protons, which cause slow chemical changes in the material directly exposed to them. As a consequence, the uppermost few metres below the surface of an atmosphere-less, silicate-rich planetary body consist of a finely grained, porous material termed regolith (McKay et al., 1991). This has been shown directly by the drilling experiments performed during some of the Apollo missions to the Moon, and there is general consensus that the surfaces of other atmosphere-less planetary bodies in the inner solar system such as Mercury have similar properties. In the outer solar system, most of the moons of the large gaseous planets have icy surfaces (with the famous exception of Jupiter's innermost Galilean moon Io showing strong sulfuric and silicic volcanism) with completely different physical properties.

For a regolith surface, the observed reflectance behaviour is essentially governed by the soil composition, grain size, porosity, and roughness. Most theoretical work in this field has been directed towards the derivation of semiempirical photometric functions from physical laws (Goguen, 1981; Lumme and Bowell, 1981; Hapke, 1981, 1984, 1986, 1993, 2002). The Hapke model is most widely used since it has been proven accurate in laboratory and planetary measurements (Clark and Roush, 1984; Warell, 2004).

Most parameters of the Hapke model are directly related to physical properties of the surface material. The reflectance function by Hapke (1993) depends on several parameters: the particle single-scattering albedo w , the width h and the amplitude B_0 of the opposition effect, the subresolution-scale mean surface slope angle $\bar{\theta}$, and one or two parameters describing the shape of the angular particle scattering function $p(\alpha)$ of the average particle (single-particle phase function), where α denotes the phase angle. The term opposition effect describes a strong increase in the intensity of the light reflected from the surface for phase angles smaller than about a few degrees. According to Hapke (1986), the opposition effect has two major sources. The shadow-hiding opposition effect is due to the fact that regolith surfaces are porous, and under moderate and large phase angles the holes between the grains are filled by shadows. These shadows disappear for small phase angles, leading to an increased intensity of the light reflected from the surface. The coherent backscatter opposition effect is due to coherent reflection of light at the surface particles under low phase angles.

The reflectance function by Hapke (1981), which does not yet include the macroscopic surface roughness, is given by

$$R_{\text{H81}}(\theta_i, \theta_e, \alpha) = \frac{w \cos \theta_i}{4\pi(\cos \theta_i + \cos \theta_e)} \left([1 + B(\alpha)] p(\alpha) - 1 + H(\cos \theta_i) H(\cos \theta_e) \right). \quad (7.1)$$

The function $B(\alpha)$ describes the shadow-hiding opposition effect. With the parameters B_0 and h describing the width and the strength of the opposition effect, $B(\alpha)$ can be expressed as

$$B(\alpha) = \frac{B_0}{1 + \tan(\alpha/2)/h} \quad (7.2)$$

(Hapke, 1986). The half width of the peak is approximately given by $\Delta\alpha = 2h$. If the opposition effect were exclusively due to shadow hiding, one would expect a value of B_0 between 0 and 1. In several studies, however, values of B_0 larger than 1 are allowed to additionally take into account a coherent backscatter contribution to the opposition effect (Helfenstein et al., 1997; Warell, 2004), as it is difficult to separate the two contributions from each other.

For the particle angular scattering function $p(\alpha)$, McEwen (1991) uses the phase function introduced by Henyey and Greenstein (1941) involving a single asymmetry parameter. However, it is often more favourable to adopt the double Henyey-Greenstein formulation

$$p_{\text{2HG}}(\alpha) = \frac{1+c}{2} \frac{1-b^2}{(1-2b\cos\alpha+b^2)^{3/2}} + \frac{1-c}{2} \frac{1-b^2}{(1+2b\cos\alpha+b^2)^{3/2}} \quad (7.3)$$

introduced by McGuire and Hapke (1995). In Eq. (7.3), the parameter $b \in [0, 1]$ describes the angular width of the forward and backward scattering lobes, where $b=0$ for isotropically scattering particles. The parameter c denotes the amplitude of the backscattered lobe relative to the forward scattered lobe. In principle, the value of c is unconstrained but has to be chosen such that $p_{\text{2HG}}(\alpha)$ is positive for all phase

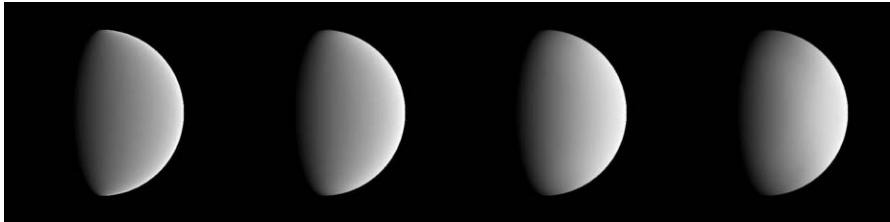


Fig. 7.1 Intensity across a planetary disk of uniform albedo for a phase angle of $\alpha = 70^\circ$ for macroscopic roughness parameters $\bar{\theta}$ of (from the left) 0° , 10° , 20° , and 30° , computed based on the reflectance function introduced by Hapke (1984). The other parameters of the Hapke model have been chosen according to solution 1 obtained for the Moon by Warell (2004).

angles. For backscattering particles, the value of c is positive. Eq. (7.3) provides a good representation of the double-lobed scattering properties of real particles and at the same time involves only a small number of parameters.

The function $H(x)$ in Eq. (7.1) takes into account multiple scattering processes. It is tabulated by Chandrasekhar (1950), and a first order approximation is given by Hapke (1984) according to

$$H(x) \approx \frac{1+2x}{1+2x\sqrt{1-w}}. \quad (7.4)$$

An improved, second-order approximation of $H(x)$ is introduced by Hapke (2002), which is relevant essentially for high-albedo surfaces. Only very small differences, however, are observed compared to the formulation in Eq. (7.4) for low-albedo surfaces like that of the Moon.

Eq. (7.1) provides a good representation of the reflectance behaviour of smooth particulate laboratory samples across a broad range of values for the single-scattering albedo w . However, planetary surfaces are not smooth like laboratory samples but are rough on scales ranging from some tenths of a millimetre to the size of topographic features not resolved by the imaging device (Hapke, 1984). It has been shown by Helfenstein (1988) based on the analysis of close-up stereo images of the lunar surface acquired during the Apollo missions that the strongest influence on the reflectance behaviour comes from surface roughness on submillimetre and millimetre scales. The reflectance function including the macroscopic roughness $\bar{\theta}$ according to Hapke (1984) is left out for brevity. For illustration, intensity distributions across a planetary disk are shown in Fig. 7.1 for values of $\bar{\theta}$ between 0° and 30° . The other parameters of the reflectance functions are set to $w = 0.168$, $b = 0.21$, $c = 0.7$, $h = 0.11$, and $B_0 = 3.1$ according to Warell (2004) (cf. Table 1 in that study, Moon solution 1 therein). For the angular scattering function, the formulation according to Eq. (7.3) has been used. Fig. 7.1 illustrates that the Hapke model yields an unrealistically bright rim around the limb of the planetary disk for $\bar{\theta} = 0^\circ$, i.e. when the macroscopic surface roughness is not taken into account—cf. Fig. 7.2 for a



Fig. 7.2 Image of the lunar disk, acquired under a phase angle of $\alpha = 59.6^\circ$. No bright rim at the limb of the disk is apparent.

real image of the lunar disk, in which no bright rim is visible at the limb¹. The bright rim disappears more and more for increasing values of $\bar{\theta}$, and for the phase angle configuration shown in Fig. 7.1, the intensity distribution increasingly resembles that of a Lambertian surface. Most planetary surfaces display macroscopic roughness values between 10° and 30° (Hapke, 1984; Helfenstein, 1988; Everka et al., 1988; Warell, 2004).

It is not straightforward, however, to utilise the reflectance function introduced by Hapke (1981, 1984, 1986) for the purpose of three-dimensional surface reconstruction based on photometric methods. According to McEwen (1991), a reflectance function that fairly well describes the scattering of light from dark porous materials such as the lunar regolith at low phase angles is of the form

$$R_L(\theta_i, \theta_e, \alpha) = f(\alpha) \frac{\cos \theta_i}{\cos \theta_i + \cos \theta_e}, \quad (7.5)$$

where $f(\alpha)$ is a function of the phase angle. For constant $f(\alpha)$, Eq. (7.5) corresponds to the Lommel-Seeliger reflectance law (Lohse and Heipke, 2004). For large phase angles approaching 180° , the reflectance of the lunar surface becomes increasingly Lambertian. McEwen (1991) proposes the Lunar-Lambert function, which is a combination of the Lommel-Seeliger and the Lambertian reflectance function according to

$$R_{LL}(\theta_i, \theta_e, \alpha) = \rho \left(2L(\alpha) \frac{\cos \theta_i}{\cos \theta_i + \cos \theta_e} + (1 - L(\alpha)) \cos \theta_i \right) \quad (7.6)$$

with ρ as an albedo parameter that also absorbs quantities specific to the image acquisition process and $L(\alpha)$ as an empirical phase angle dependent parameter (cf. Section 2.2.1 for the corresponding BRDF). The formulation in Eq. (7.6) has the

¹ In astrogeology, the term “planet” is often also used for the Earth’s Moon due to its similarity to the four inner (terrestrial) planets of the solar system.

advantage that it can be used directly for ratio-based photoclinometry and shape from shading methods as described in Section 2.3.2.1, as the surface albedo ρ then cancels out for each image pixel.

To obtain a function $L(\alpha)$ that provides a realistic description of the reflectance behaviour of the lunar surface, McEwen (1991) computes intensity profiles along the photometric equator and the mirror meridian based on the photometric function by Hapke (1986). The photometric equator is the great circle through the subsolar and the subspacecraft points, and the mirror meridian is the perpendicular great circle with $\theta_i = \theta_e$. The intensity values are evenly sampled in photometric longitude, the angle measured along the photometric equator, and photometric latitude, the angle measured along the mirror meridian. The opposition effect parameters B_0 and h have a significant influence on the intensity distribution only for phase angles below about 5° , where it is not favourable to apply photoclinometric techniques also with respect to illumination and viewing geometry. As it tends to be difficult to fit the Lunar-Lambert function to the Hapke function near the limb of the planetary disk (and such oblique viewing angles are not desirable when applying photoclinometry or shape from shading), emission angles larger than $70^\circ + \alpha/9$ are excluded from the fit.

The behaviour of $L(\alpha)$ is illustrated by McEwen (1991) as a sequence of reference diagrams for different values of the single-scattering albedo w , different parameters of the particle angular scattering function, and macroscopic roughness values between 0° and 50° . For low-albedo surfaces with $w \approx 0.1$ and macroscopic roughness values $\bar{\theta}$ larger than about 10° , the shape of the particle angular scattering function has only a minor influence on $L(\alpha)$. For small phase angles, $L(\alpha)$ is always close to 1, while it approaches zero for phase angles larger than about 140° . For intermediate phase angles, the behaviour of $L(\alpha)$ strongly depends on the value of $\bar{\theta}$, where in the practically relevant range for values of $\bar{\theta}$ between 0° and 30° , $L(\alpha)$ monotonously decreases with increasing $\bar{\theta}$. As a consequence, the reflectance behaviour becomes increasingly Lambertian for increasing phase angles and $\bar{\theta}$ values (cf. Fig. 7.1). For the lunar surface, the behaviour of $L(\alpha)$ can also be expressed as a third order polynomial in α (McEwen, 1996). We utilise the Lunar-Lambert reflectance function with $L(\alpha)$ as established by McEwen (1991) in Sections 7.2 and 7.4 for the three-dimensional reconstruction of lunar craters and lunar volcanic features.

7.2 Three-dimensional Reconstruction of Lunar Impact Craters

7.2.1 Shadow-based Measurement of Crater Depth

Determining the horizontal and vertical dimensions of impact craters from spacecraft imagery requires knowledge about the relative geometry of the Sun, the planet, and the spacecraft. Depths of craters with bowl-shaped interiors can be estimated

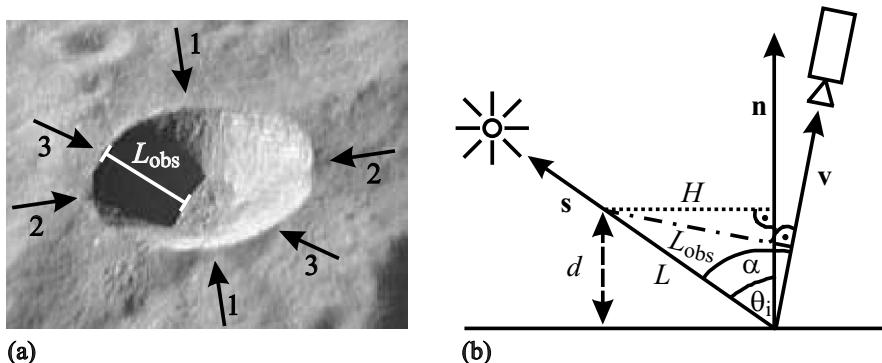


Fig. 7.3 Determination of crater depth from oblique orbital images based on shadow length measurement. (a) Section of Apollo image AS15-97-13190, showing a slightly modified bowl-shaped crater. Maximum perspectival foreshortening is observed in direction 1, no foreshortening occurs in direction 2. The crater is illuminated by the Sun in direction 3. The observed shadow length L_{obs} is indicated by the white line. (b) Geometric configuration of the camera, light source, surface normal, and shadow. The crater depth is denoted by d and the true shadow length by L .

from the length of the shadow cast by the crater rim crest, provided that the shadow tip falls into the centre of the crater. A generally applicable method is described by Pike (1988), who utilises it for crater measurements in Mariner 10 images of Mercury. Fig. 7.3a illustrates the corresponding geometric relations based on the orbital Apollo 15 image AS15-97-13190, displaying a slightly modified bowl-shaped lunar crater. Maximum perspectival foreshortening is observed in direction 1, no foreshortening occurs in direction 2, while the azimuthal direction of incident sunlight corresponds to direction 3. For computing the crater depth, the incidence angle θ_i and the phase angle α need to be known. The observed apparent shadow length viewed obliquely from the spacecraft as shown in Fig. 7.3a is denoted by L_{obs} , the crater depth by d , and the true shadow length by L . Accordingly, Fig. 7.3b yields the relations $d/L = \cos \theta_i$ and $L_{\text{obs}}/L = \sin \alpha$, leading to

$$d = \frac{L_{\text{obs}} \cos \theta_i}{\sin \alpha}. \quad (7.7)$$

The value of d only represents the true crater depth if the shadow tip falls near the lowest point on the crater floor, which is generally close to its centre. Hence, the horizontal extent H of the shadow, which corresponds to $H = L_{\text{obs}} \sin \theta_i / \sin \alpha$ (cf. Fig. 7.3b), should be close to half the crater diameter. To establish a depth-diameter relation for Mercurian craters, Pike (1988) use only craters with shadow tips closer to the crater centre than 5 percent of the crater diameter as their analysis specifically refers to simple bowl-shaped craters. The shadow length in Fig. 7.3a represents the full crater depth despite the shadow being shorter than half the crater diameter, since although the crater is bowl-shaped it is slightly modified, and a part of its floor is flat.

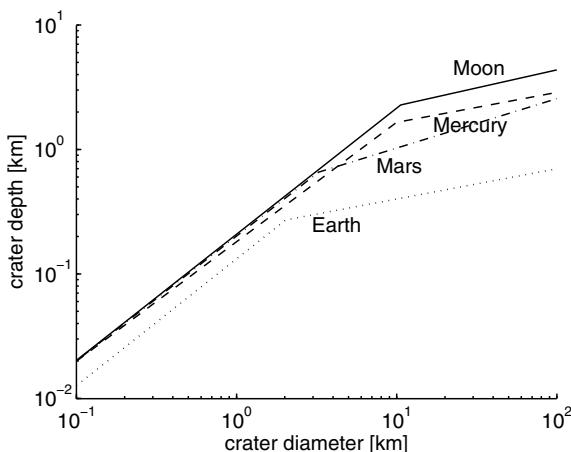


Fig. 7.4 Depth-diameter relations for Mercury, the Moon, the Earth, and Mars according to Pike (1980).

Statistical analyses of the relations between crater depths and diameters, based on the described shadow length measurement method, provide important information about crater formation and modification on specific planetary bodies. Furthermore, a comparison of depth-diameter characteristics between different planets may provide insights into the influencing quantities important for cratering and subsequent modification processes, such as surface gravity, atmospheric effects, or substrate variations (Cintala et al., 1976). A sketch of depth-diameter relations for impact craters on Mercury, the Moon, and Mars established by Pike (1980) is shown in Fig. 7.4. The lines are regression curves obtained based on several hundreds of craters, respectively. For all three planets, the slope of the curve in the double-logarithmic diagram displays a sudden change at a critical diameter which marks the transition from simple bowl-shaped craters to more complex structures displaying (with increasing diameter) flat floors, scalloped rims, central peaks, and terraced walls (Wood, 1973; Wood and Andersson, 1978; Pike, 1980, 1988). This transition occurs for crater diameters of about 10–16 km for Mercury and the Moon, 3–6 km for Mars, and 3 km for the Earth. For simple craters, the behaviour of depth vs. diameter is similar for the four planets with a largely constant depth-diameter ratio of about 1/5. Complex structures tend to be shallower on Mars than on Mercury and the Moon and still shallower on the Earth. There is a trend that complex craters become shallower with increasing gravitational acceleration (corresponding to 1.63, 3.63, 3.73, and 9.81 m s^{-2} on the Moon, Mercury, Mars, and the Earth, respectively). For example, differences in crater shape between the Earth and the Moon appear to depend largely on gravity (Pike, 1980). On the other hand, the distinct difference between the relations established for Mercury and Mars, whose gravitational accelerations differ by only a few percent, indicates that both gravitational acceleration and target characteristics have affected the morphology of impact craters in the solar system.

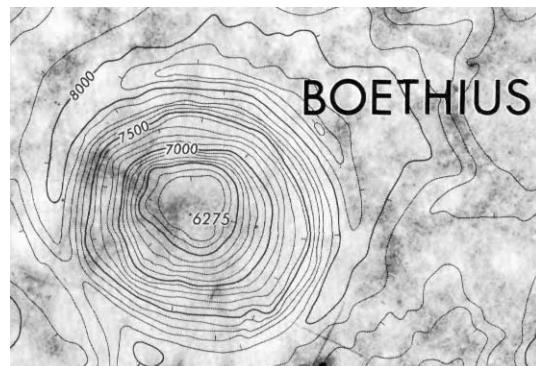


Fig. 7.5 Section of lunar topographic orthophotomap LTO 63D1, showing the crater Boethius. Contour interval is 100 m.

The relative importance of gravity and target type differs from planet to planet. An in-depth discussion of these issues is provided by Pike (1980, 1988).

7.2.2 *Three-dimensional Reconstruction of Lunar Impact Craters at High Resolution*

A reliable morphologic classification of impact craters is favourably based on a topographic map of high resolution. As an example, a section of lunar topographic orthophotomap LTO 63D1 with an elevation contour representation of the small crater Boethius (also referred to as Dubiago U in older lunar maps) is shown in Fig. 7.5. The crater diameter corresponds to 10 km, the contour interval is 100 m, the indicated elevation values refer to a reference lunar radius of 1738 km. However, no such accurate high-resolution topographic information is available for most of the lunar surface, as the LTO series covers only about one fifth of it. Hence, this section describes results obtained for the three-dimensional reconstruction of lunar impact craters at high resolution based on the methods discussed in Chapters 2 and 4.

The utilised telescopic images were acquired with ground-based telescopes of 125 mm and 200 mm aperture equipped with a CCD camera (Philips ToUCam). Each image was generated by stacking several hundreds of video frames. For this purpose we made use of the Registax and Giotto software packages, employing a cross-correlation technique similar to the one described by Baumgardner et al. (2000). In that work, however, digitised analog video tapes were processed, while we directly acquired digital video frames. The scale of our images is between 500 and 800 m per pixel on the lunar surface. Due to atmospheric seeing, however, the effective resolution (denoted by the diameter of the point spread function rather than the spatial resolution per pixel) of our best images approximately corresponds to

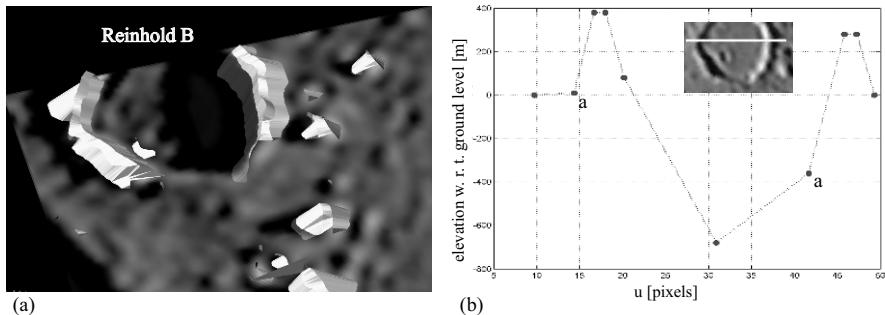


Fig. 7.6 (a) Shadow-based three-dimensional reconstruction of the crater Reinhold B according to Section 2.1.1. (b) Cross-section through the crater floor, obtained based on shadow length measurements in three images acquired at different illumination conditions.

1 km on the lunar surface. All other telescopic lunar images regarded in this section were acquired with the same technique.

As an example of a complex crater shape not immediately revealed by the image data, Fig. 7.6a displays the lunar crater Reinhold B. Due to the fact that the time intervals between the images used for extraction of shadow information typically amount to several weeks, the relative position of camera and object cannot be kept constant during acquisition of all necessary images of the scene, such that an image registration procedure (Gottesfeld Brown, 1992) becomes inevitable to obtain pixel-synchronous images.

Although the floor of the crater Reinhold B looks flat, shadow analysis based on two images according to Section 2.1.1 reveals that the elevation difference between the ridge of the eastern (right) crater rim and the centre of the crater floor amounts to 1000 m and is thus 700 m larger than the elevation difference between the ridge of the western (left) rim, while the ridges of both rims are 300–400 m above the level of the environment. This means that the western part of the crater floor is on about the same level as the environment, while its central part is 700 m deeper. Hence, the western half of the crater floor is inclined by an angle of approximately 4° , given the crater diameter of 19 km. Fig. 7.6b shows a cross-section of Reinhold B which additionally contains the results of shadow analysis obtained with a third image obtained at again different illumination conditions (marked by “a”). Here, an identical ground level east and west of the crater has been assumed (Hafezi and Wöhler, 2004).

High-resolution digital elevation maps (DEMs) of lunar impact craters have been generated with the integrated method based on shadow and shading described in Section 4.2 (Wöhler and Hafezi, 2005). It is not possible, however, to obtain the required images from existing image archives of lunar spacecraft missions, as these generally do not contain several images of a surface region under sufficiently different illumination conditions. Hence, we again used telescopic lunar CCD images. Selenographic positions were obtained from Rükl (1999). The three-dimensional

reconstruction approach described in Section 4.2 assumes that the grey value I of a pixel is proportional to the incident flux F . However, for many cameras it is possible to adjust the gamma value γ manually from within the camera control software, causing the grey value I to be proportional to F^γ . Hence, we performed a calibration of the gamma scale in the camera control software by evaluating flatfield frames of different intensities acquired through different neutral density filters with known transmission coefficients, then fitting a characteristic curve of the form $I = aF^\gamma$ to the measured flatfield intensities.

In the scenario of three-dimensional reconstruction of small regions of the lunar surface under roughly perpendicular view and oblique illumination, $\cos \theta_e$ in the Lunar-Lambert reflectance function (7.6) hardly deviates from its average value $\langle \cos \theta_e \rangle \approx 1$. Since $\cos \theta_i \ll 1$, changes in θ_i do not significantly influence the value of the denominator $(\cos \theta_e + \cos \theta_i)$, such that it can be approximated by the constant value $(\langle \cos \theta_e \rangle + \langle \cos \theta_i \rangle)$ as long as the surface gradients are small. This results in an effective albedo $\rho_{\text{eff}} = \rho \left[\frac{2L(\alpha)}{\langle \cos \theta_e \rangle + \langle \cos \theta_i \rangle} + 1 - L(\alpha) \right]$. This approximation is acceptable for the lunar impact craters regarded here because the regions examined in this section are not close to the limb of the lunar disk, and the observed surface slopes are usually smaller than about 10° . Furthermore, the average depth deviation between the surface profile computed with a Lunar-Lambert reflectance function from the Lambertian solution is always well within the error range of shadow analysis in the example scenes. In the examples of Fig. 7.7a, the root mean square deviation is smaller than 40 m for $L(\alpha)$ in the range $-0.1, \dots, 1$. If, however, the surface is imaged under a more oblique viewing angle, as it is the case for some of the lunar domes regarded in Section 7.4, it is important to fully take into account the Lunar-Lambert reflectance function.

For each examined crater, the obtained surface profile is checked for consistency with the image data by means of a raytracing software. The appearance of the surface profile under solar elevation angles μ (shading image) and μ_{shadow} (shadow image) is simulated assuming a Lambertian surface. For all examples, both simulated images are satisfactorily similar to the corresponding real images. A good cross-check is provided by the rendered shadow image due to the fact that the shape of the simulated shadow is highly sensitive with respect to small deviations of the surface profile from the correct shape. Additionally, we have shown (see beginning of Section 4.2) that the quality of reconstruction is hardly affected by inaccurate knowledge of the reflectance function.

Fig. 7.7a shows the three-dimensional surface reconstruction of the outer western rim of the lunar crater Copernicus, obtained by employing the technique outlined in Sections 4.2.1 and 4.2.2 that consists of selecting a shape from shading solution consistent with shadow analysis in a first step and taking into account the detailed shadow structure in a second step. The shading image alone does not reveal the large-scale eastward slope of the profile, which can only be derived from the shadow image. Effectively, the shadow allows to adjust the albedo ρ such that the shape from shading algorithm yields a surface profile consistent with both the small-scale depth variations evident in the shading image and the large-scale slope derived from the shadow image. Fig. 7.7b shows the inner eastern rim of the lunar

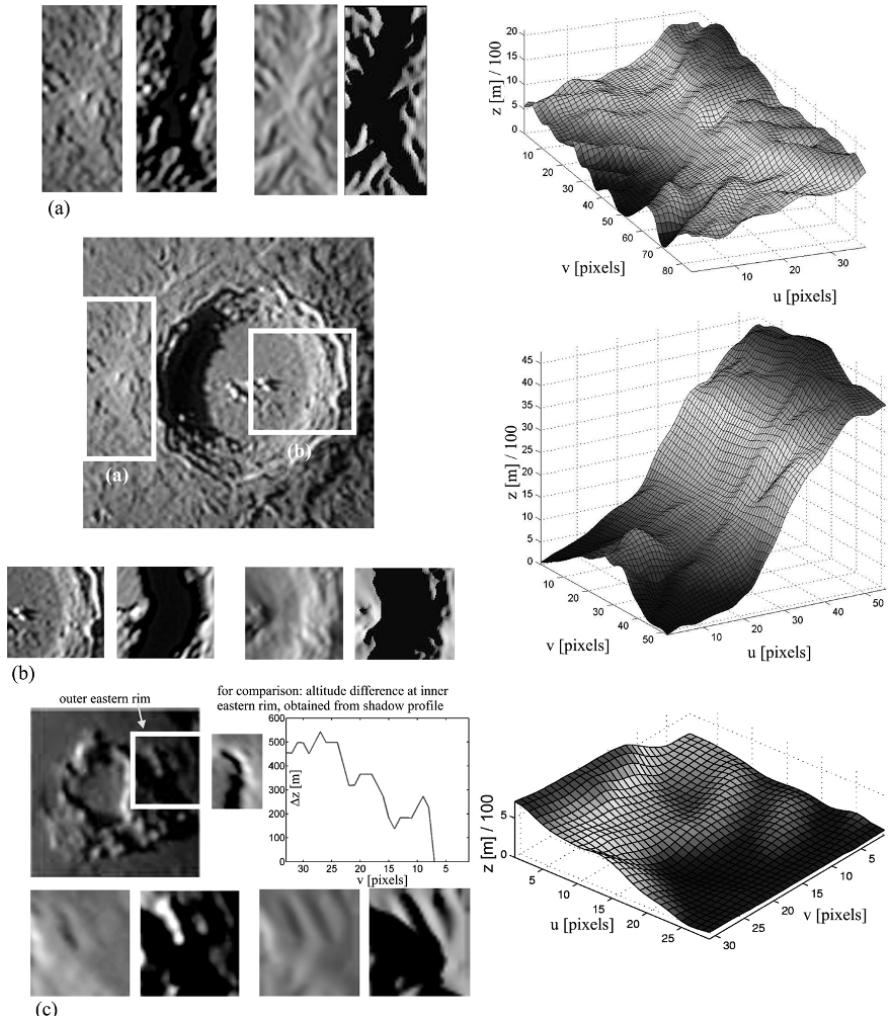


Fig. 7.7 Three-dimensional reconstruction of lunar craters by combined shading and shadow analysis according to Section 4.2. The original shading and shadow images are shown on the left, respectively, along with the obtained three-dimensional surface profile and the corresponding rendered images, given the same illumination conditions as used for shading and shadow analysis. (a) Outer western rim of lunar crater Copernicus. (b) Inner eastern rim of lunar crater Copernicus. (c) Outer eastern rim of lunar crater Wolf.

crater Copernicus. The surface profile reveals terraces and small craters in the crater wall. The performance of the reconstruction algorithm is slightly decreased by small shadows in the shading image cast by the central peaks. The correspondence of the simulated shadow contour with its real counterpart is reasonable. In Fig. 7.7c, the outer eastern rim of the lunar crater Wolf is shown along with the depth differences at the corresponding inner rim obtained by shadow analysis alone. A comparison re-

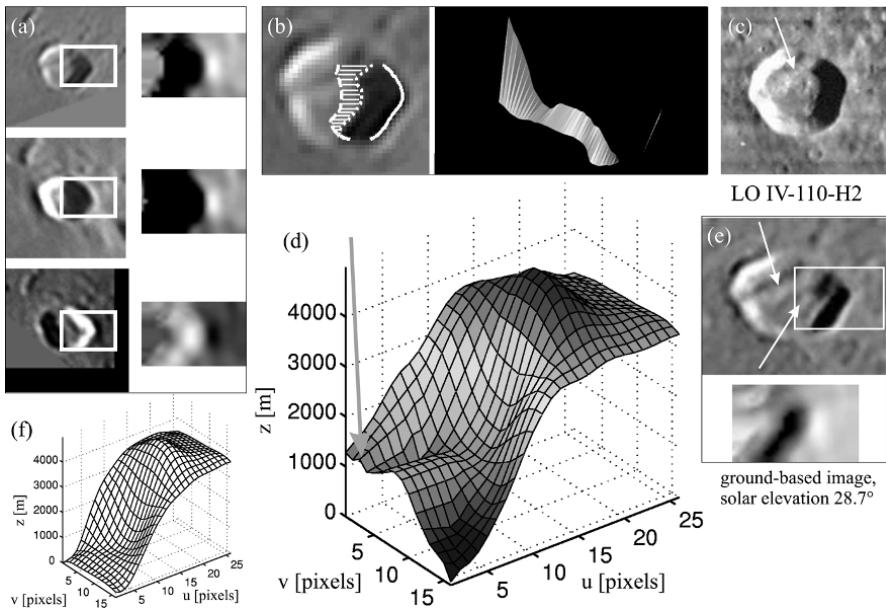


Fig. 7.8 Three-dimensional reconstruction of the eastern half of lunar crater Theaetetus based on the initialisation of the shape from shading algorithm by shadow analysis according to Section 4.2.3. (a) The two upper images are evaluated with respect to shadow, the third one with respect to shading. The reconstructed surface part is marked by a white rectangle, respectively, the corresponding simulated images are shown to the right. (b) Surface patch between the two shadow lines (hatched) along with the initial surface profile derived from shadow analysis. (c) Lunar Orbiter photograph IV-110-H2, shown for comparison. (d) Reconstructed surface profile. Although the marked ridge is hardly visible in the three ground-based images used for surface reconstruction, it clearly appears in the reconstructed surface profile due to its effect on the shadows. (e) Further ground based image of Theaetetus shown for comparison, not used for reconstruction, along with the simulated image derived from the reconstruction result. The ridge crossing the crater and the low central elevation are marked by arrows. (f) Three-dimensional reconstruction result obtained with traditional shape from shading, selecting the solution consistent with the first shadow image. None of the previously mentioned details on the crater floor is visible.

veals that the crater floor is lying on the same level as the surrounding mare surface. The simulated shadow shows all features displayed by the real shadow image.

Fig. 7.8 shows the reconstructed surface profile of the floor of the lunar crater Theaetetus, which has a diameter of 25 km. It was generated by the technique described in Section 4.2.3 that relies on an initialisation of the shape from shading algorithm by surface gradients obtained by the analysis of several shadows observed under different illumination conditions. Both the simulated shading image and the contours of the simulated shadows correspond well with their real counterparts. Even the ridge crossing the crater floor, which is visible in the upper left corner of the region of interest in Fig. 7.8a and in the Lunar Orbiter photograph shown in Fig. 7.8c for comparison, is apparent in the reconstructed surface profile (arrow). Hence, the reconstruction technique reliably indicates even small-scale structures

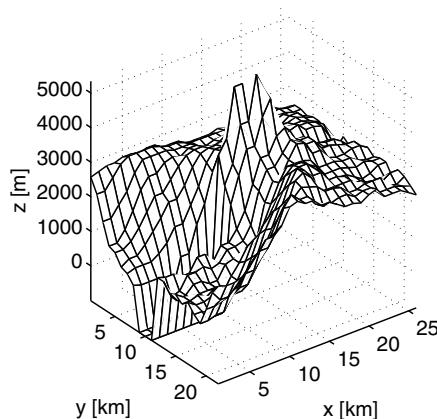


Fig. 7.9 Elevation map of the eastern half of Theaetetus, extracted from the topographic data by Cook (2007).

on the surface that cover only a few pixels. Furthermore, it turns out that the crater floor is inclined from the north to the south, and a low central elevation (rising to about 250 m above the floor level) becomes apparent in the reconstructed surface profile. Such features, which are debris mounds in the crater interior, are important for a geological interpretation of the crater, as they essentially mark the difference between simple bowl shaped craters and incipient complex craters such as Theaetetus (Spudis, 1993). This central elevation does not appear in the images in Fig. 7.8a used for reconstruction but is clearly visible in the telescopic image of Theaetetus acquired at a solar elevation angle of $\mu = 28.7^\circ$ shown in Fig. 7.8e (lower arrow). The corresponding simulated image (lower part of Fig. 7.8e) is very similar to the real image although that image has not been used for reconstruction. This kind of comparison is suggested by Horn (1989) as an independent test of reconstruction quality. For comparison, traditional shape from shading as outlined in Section 2.2.2.2 yields an essentially flat crater floor and no ridge (Fig. 7.8f). This shows that the traditional approach is obviously not able to extract reliable information about surface gradients perpendicular to the azimuthal direction of illumination under the given illumination conditions.

As an independent comparison, Fig. 7.9 shows a three-dimensional reconstruction of the eastern half of Theaetetus extracted from the elevation maps provided by Cook (2007). These topographic data have a lateral resolution of 1 km and were computed with the stereophotogrammetric approach described by Cook et al. (1999), relying on Clementine orbital imagery. The elevation map in Fig. 7.9 is fairly noisy and contains some spike artifacts on the crater rim. Furthermore, for some points on the crater floor no data are available. No small-scale structures appear on the crater floor. However, the crater depth according to Fig. 7.8d obtained with the combined method based on shadow and shading features is in good correspondence

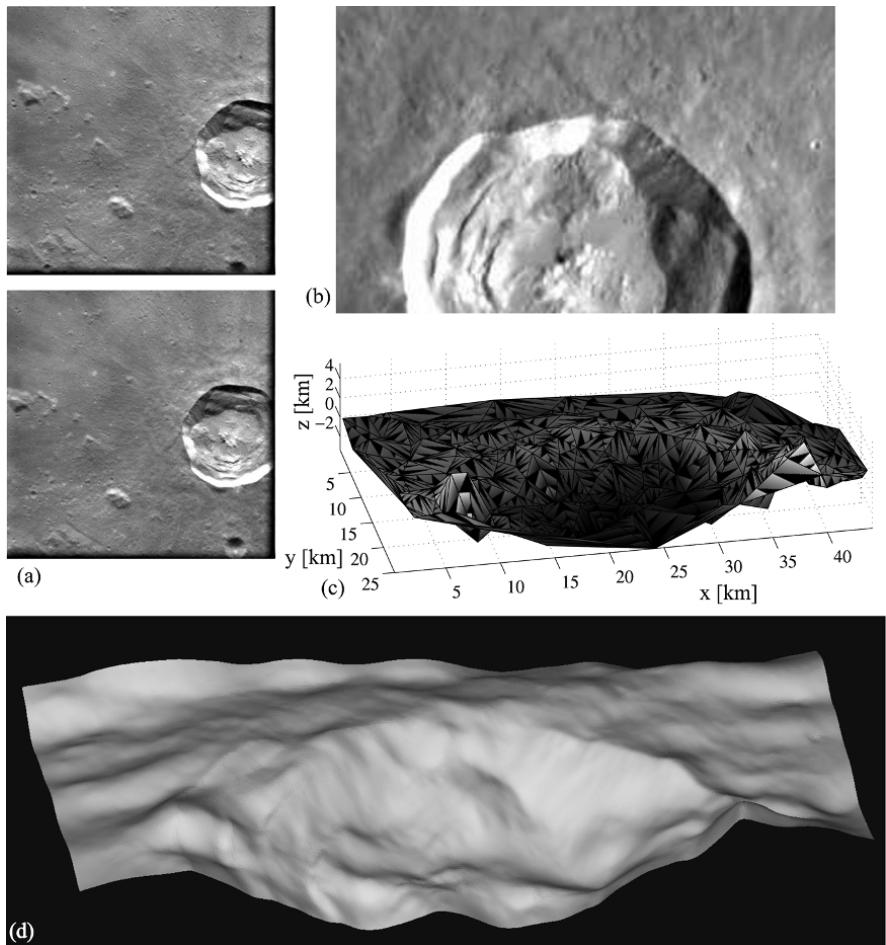


Fig. 7.10 Digital elevation map of the lunar crater Kepler. (a) First and last image of the five-image sequence acquired by the SMART-1 spacecraft (European Space Agency, 2007). (b) Region selected for three-dimensional reconstruction, oriented such that north is to the top and west to the left. (c) Three-dimensional reconstruction obtained by structure from motion (triangulated three-dimensional point cloud). (d) Dense three-dimensional reconstruction using the combination of structure from motion with shape from shading.

with the depth indicated in Fig. 7.9 by the result of the stereophotogrammetric approach.

As an example of the combination of shape from shading with sparse depth data, a sequence of five images of the lunar crater Kepler acquired by the SMART-1 spacecraft on January 13, 2006, from heights above the lunar surface between 1613 and 1702 km (European Space Agency, 2007) has been analysed according to the method described in Section 4.3.3 (d'Angelo and Wöhler, 2008). The crater diameter amounts to 32 km. During image acquisition the spacecraft flew over the crater

and at the same time rotated around its axis, such that the crater remained in the field of view over a considerable period of time. The first and the last image of the sequence are shown in Fig. 7.10a. Image size is 512×512 pixels. Fig. 7.10b shows the reconstructed part of the surface, which is smaller than the complete field of view as the surface albedo becomes non-uniform at larger distances from the crater. The image is rotated such that north is to the top and west to the left. Relying on a structure from motion analysis based on bundle adjustment (cf. Section 1.2), a three-dimensional point cloud is extracted from the image sequence as shown in Fig. 7.10c after Delaunay triangulation. Since no lens calibration data are available it is assumed that the lens can be described by the pinhole model with the principal point in the image centre. The image scale amounts to 146 m per pixel (European Space Agency, 2007), such that the scaling constant can be readily determined for the structure from motion result.

Since no polarisation information is available, the shape from shading method has been combined with the result of structure from motion (cf. Section 4.3.3), making use of the Lunar-Lambert reflectance function according to Eq. (7.6). At the time of image acquisition, the phase angle was $\alpha = 51^\circ$ for the spacecraft, corresponding to $L(\alpha) = 0.8$ according to McEwen (1991), and the solar elevation angle was $\mu = 37^\circ$. The viewing direction was determined according to the normal vector of a plane fitted to the three-dimensional point cloud extracted by structure from motion analysis. For this non-specular surface, the albedo ρ was estimated based on all image pixels in the course of the iteration process according to Eq. (4.24) as explained in Section 4.3.2. Saturated (white) pixels were excluded from the shape from shading analysis.

The three-dimensional reconstruction result shown in Fig. 7.10d distinctly reveals the uneven crater floor of Kepler as well as the material that has slumped down the inner crater wall at several places, especially at the northern rim. The reconstructed surface obtained with the combined structure from motion and shape from shading approach reveals much finer detail than the structure from motion data alone. The typical depth difference between crater floor and rim amounts to about 2850 m. No ground truth is available for this crater since it is not covered by the existing lunar topographic maps. A crater depth of 2750 m is reported in the lunar atlas by Rükl (1999). This is an average value since most crater depths given in lunar atlases were determined by shadow length measurements based on telescopic or spacecraft observations. The crater depth extracted from our three-dimensional reconstruction result is in reasonable agreement with the value given by Rükl (1999).

This example demonstrates the usefulness of the combination of intensity data and sparse depth data obtained from a camera moving in an uncontrolled manner, regarding a surface with well-defined reflectance properties under accurately known illumination conditions. The self-consistent solution for the three-dimensional surface profile obtained according to Section 4.3.3 yields a crater rim of largely uniform height, indicating that the estimated surface gradients in the horizontal and in the vertical image direction are essentially correct. In contrast, surface reconstruction by shape from shading alone based on images acquired under identical illumination conditions is not able to simultaneously estimate both surface gradients for each

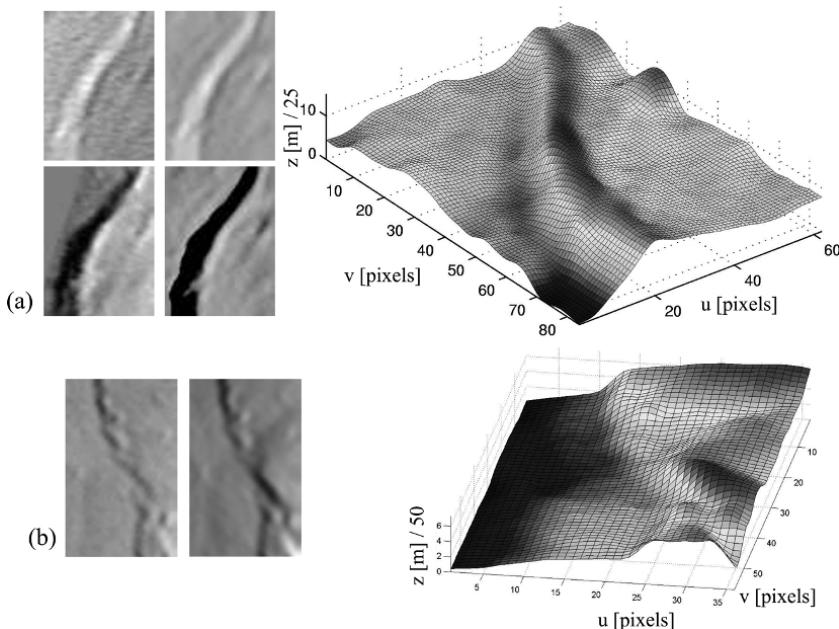


Fig. 7.11 Telescopic CCD images, rendered counterparts, and three-dimensional reconstruction result for (a) mare ridge or fault south-west of the crater Aristarchus and (b) the end of a lunar lava flow at the border between Mare Imbrium and Mare Serenitatis.

pixel as long as no boundary values are known for the surface to be reconstructed. What is more, in contrast to previous methods (Samaras et al., 2000; Fassold et al., 2004; Horovitz and Kiryati, 2004), the sparse depth points do not introduce spurious artifacts into the reconstructed surface profile despite the considerable noise in the three-dimensional point cloud (cf. Fig. 7.10c) extracted by structure from motion.

7.3 Three-dimensional Reconstruction of Lunar Wrinkle Ridges and Faults

The mare regions of the Moon display numerous low wrinkle ridges (dorsa) and a small number of tectonic faults (rupes). For many of these features no elevation information is available. Wrinkle ridges were formed when the mare lava contracted while cooling down. Tectonic faults are linear structures at which the surface elevation suddenly changes. They represent a surface expression of crustal fractures which formed as a consequence of the large basin impacts on the Moon (Wilhelms, 1987). This section presents three-dimensional reconstruction results according to Wöhler and Hafezi (2005) and Wöhler et al. (2006a) for some typical wrinkle ridges and tectonic faults.

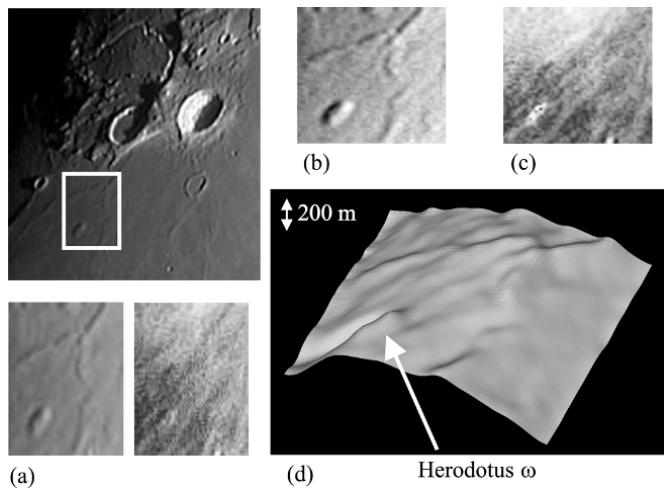


Fig. 7.12 Reconstruction of the region around lunar dome Herodotus ω based on two images acquired at different illumination conditions ($\mu_1 = 5.0^\circ$, $\mu_2 = 15.5^\circ$) and ratio-based intensity error term (2.50). (a) Scene at lower (left image) and higher (right image) solar elevation. Contrast has been enhanced. (b) Ratio image $I_{uv}^{(1)} / I_{uv}^{(2)}$. Due to the distance of the surface region from the centre of the Moon's apparent disk, the image has been correspondingly scaled. (c) Albedo map obtained according to Eq. (2.49). (d) Reconstructed surface profile, shown as shaded relief to accentuate subtle surface structures. Along with the actual surface features, the bending of the surface due to the spherical shape of the Moon is visible in horizontal image direction.

Fig. 7.11a shows a structure situated south-west of the crater Aristarchus which appears like a wrinkle ridge. Three-dimensional surface reconstruction was performed based on the approach described in Section 4.2 combining shadow and shading information. The initial adaptation of the surface profile according to Section 4.2.1 was skipped and the surface was reconstructed by directly applying the algorithm described in Section 4.2.2 that takes into account the detailed shadow structure, making use of integrability constraint (2.26). The surface profile reveals a step-like structure of this mare ridge, elevated by 250 m and 180 m above the terrain to its west and east, respectively. The elevation difference indicates that this structure is probably a fault rather than a typical wrinkle ridge.

It is possible to employ the reconstruction technique described in Section 4.2 even when no shadow image is available, by setting the value of $(\Delta z)_{\text{shadow}}^{\text{ave}}$ (cf. Eq. (4.5) in Section 4.2.1) according to a-priori knowledge about the large-scale behaviour of the surface profile. Such large-scale depth information may also be obtained from space-based radar or laser altimetry measurements. The surface profile of a step-like structure between Mare Serenitatis and Mare Imbrium shown in Fig. 7.11b has been derived by setting $(\Delta z)_{\text{shadow}}^{\text{ave}} = 0.2$ pixels ≈ 160 m, corresponding to the average height of the step. This fault-like formation is not of tectonic origin but the end of a lunar lava flow.

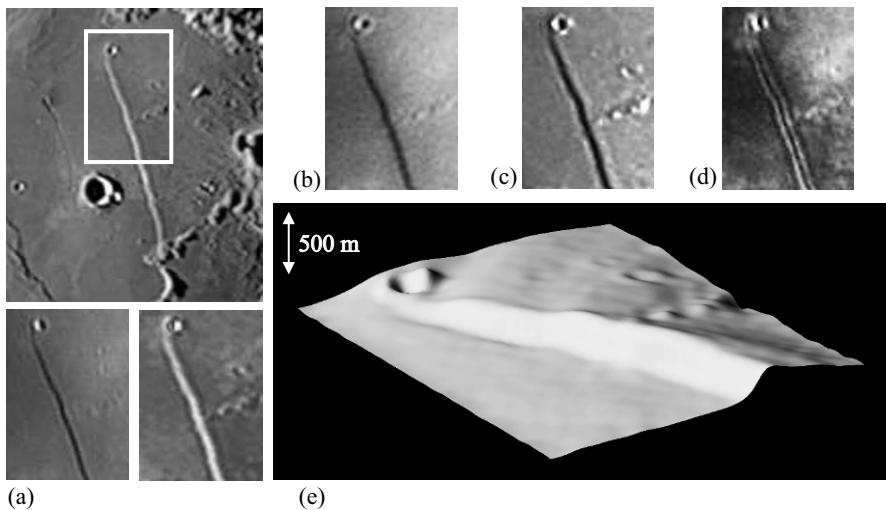


Fig. 7.13 Digital elevation map of the northern part of the tectonic fault Rupes Recta based on two images acquired at different illumination conditions ($\mu_1 = 13.7^\circ$, $\mu_2 = 175.6^\circ$) by first applying the ratio-based intensity error term (2.50) with smoothness and shadow constraints (2.21) and (4.10), then using the resulting surface profile and albedo map as an initialisation to the shape from shading scheme that involves integrability constraint (2.26). (a) Scene illuminated from the east (left image) and from the west (right image). (b) Shadow image. (c) Ratio image. (d) Albedo map obtained according to Eq. (2.49). (e) Reconstructed surface profile, shown as shaded relief to accentuate subtle surface structures. Along with the actual surface features, the bending of the surface due to the Moon's spherical shape is visible in horizontal image direction.

Fig. 7.12 displays the three-dimensional reconstruction result for a section of the lunar surface south-west of the crater Aristarchus with its bright ray system. This surface part displays a strongly non-uniform albedo. In this example, surface reconstruction is performed based on two images taken at different solar elevation angles $\mu_1 = 5.0^\circ$ and $\mu_2 = 15.5^\circ$ but virtually identical solar azimuth angles, using error term (2.50). The average pixel intensity $\langle I_{uv}^{(1)} \rangle$ of the first image is scaled such that $\langle I_{uv}^{(1)} \rangle / \langle I_{uv}^{(2)} \rangle = \sin \mu_1 / \sin \mu_2$, which means that on the average, the surface section is assumed to be flat. A large-scale surface slope of angle δ in the direction of incident light might be imposed by setting $\langle I_{uv}^{(1)} \rangle / \langle I_{uv}^{(2)} \rangle = \sin(\mu_1 + \delta) / \sin(\mu_2 - \delta)$ —with an absolutely calibrated CCD sensor one might think of deriving such a large-scale slope directly from the surface reconstruction procedure. The reconstructed surface profile contains several low ridges with altitudes of roughly 50 m along with the lunar dome Herodotus ω . The albedo map obtained according to Eq. (2.49) displays a gradient in surface brightness from the lower right to the upper left corner along with several ray structures running radial to the crater Aristarchus.

Fig. 7.13 shows the northern part of the lunar tectonic fault Rupes Recta. The images shown in Fig. 7.13a were acquired at solar elevation angles $\mu_1 = 13.7^\circ$ and $\mu_2 = 175.6^\circ$, which means that the scene is illuminated from opposite directions.

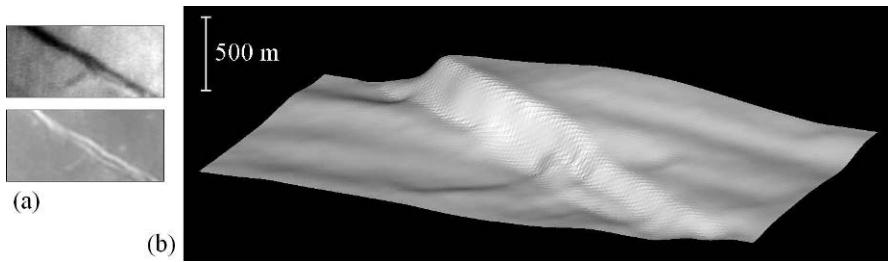


Fig. 7.14 (a) Pixel-synchronous pair of images of the central part of Rupes Cauchy, acquired at local sunrise (top) and at local sunset (bottom). Image courtesy P. Lazzarotti. (b) DEM obtained based on combined shadow and shading analysis according to Section 4.2. The vertical axis is 10 times exaggerated. The curvature of the lunar surface has been subtracted.

The scene was reconstructed by first applying the ratio-based intensity error (2.50) with smoothness and shadow constraints (2.21) and (4.10), then using the resulting surface profile and albedo map obtained by Eq. (2.49) as an initialisation to the shape from shading scheme that involves integrability constraint (2.26). A third image (cf. Fig. 7.13b) provides the required shadow information. The height of the reconstructed part of Rupes Recta ranges from 200 m to about 400 m, which is well consistent with the shadow length in Fig. 7.13b. The albedo map ρ_{uv} shows bright spots in the lower part and especially in the upper right part of the image. The linear structures in the albedo map along Rupes Recta are artifacts; presumably, ρ_{uv} is not very accurate at places where the surface strongly bends and the residual intensity error \tilde{e}_i according to Eq. (2.50) is comparably large due to the regularisation constraints.

A three-dimensional reconstruction of the central part of the lunar tectonic fault Rupes Cauchy obtained based on a combined shadow and shading analysis according to Section 4.2 is shown in Fig. 7.14 (Wöhler et al., 2006a). The upper image in Fig. 7.14a acquired at local sunrise was used for extracting shadow information. The resulting DEM reveals that the elevation of the fault is not constant, obtaining values of more than 350 m in its highest parts. In the southern part of the DEM small structures appear on the slope of the fault, reaching a height of about 100 m relative to the surface level to the west.

7.4 Three-dimensional Reconstruction of Lunar Domes

7.4.1 General Overview of Lunar Mare Domes

Mare domes are smooth low features with convex profiles gently bent upwards. They are circular to elliptical in shape. Most lunar domes were formed by outpouring of magma from a central vent, leading to a shield volcano (Wilhelms and McCauley, 1971; Head and Gifford, 1980). Some low domes are possibly due to

magmatic intrusion, i.e. subsurface accumulation of magma causing an up-doming of the bedrock layers, creating a smooth, gently sloping positive relief (Spurr, 1945; Baldwin, 1963; Wilhelms and McCauley, 1971). According to the literature (Head and Gifford, 1980), however, suchlike features may alternatively be interpreted as low effusive edifices due to lava mantling of highland terrain, or kipukas ("islands" of highland surface surrounded by mare lava), or structural features. Domes representing volcanic sources are smooth-surfaced and usually have summit pits or elongated vents, fissures, or pit chains (Wilhelms, 1987). Most vents related to domes appear to be associated with surrounding lava plains of known volcanic origin or in association with pyroclastic deposits (Head and Gifford, 1980; Jackson et al., 1997). Isolated domes may be found in almost all lunar maria, but they are concentrated on the lunar nearside and show significant abundance in the Hortensius region, Mare Insularum, Oceanus Procellarum, and in Mare Tranquillitatis.

Effusive lunar domes probably formed during the terminal phase of a volcanic eruption. Initially, lunar lavas were very fluid due to their high temperature. Thus, they were able to form extended basaltic mare plains. Over time, the temperature of the erupting lavas became lower, flow rate decreased, and crystallisation occurred. This changed the characteristics of the lava such that it began to "pile up" around the effusion vent and formed a dome (Cattermole, 1996; Mursky, 1996). Weitz and Head (1999) show that steeper domes represent the result of cooler, more viscous lavas with high crystalline content, possibly at the final stages of the eruption. Factors governing the morphological development of volcanic edifices are interrelated, including the viscosity of the erupted material, its temperature, its composition, the duration of the eruption process, the eruption rate, and the number of repeated eruptions from the vent. The viscosity of the magma depends on its temperature and composition, where the amount of crystalline material will depend upon how it is transported from the reservoir to the surface and on the crystallisation temperature of its component phases.

Mare volcanic eruptions are fed from source regions at the base of the crust or deeper in the lunar mantle. In this scenario, pressurised magma produces narrow, elongated fractures in the crust, the so-called dikes, through which it ascends towards the surface. According to Wilson and Head (1996) and Wilson and Head (2002), some dikes intruded into the lower crust while others penetrated to the surface, being the sources of extensive outpourings of lava. Thus the surface manifestation of dike emplacement in the crust is depending on the depth below the surface to which the dike penetrates. Wilson and Head (1996) state that if a dike does not propagate near the surface but stalls at greater depth, the strain will be insufficient to cause any dislocation near the surface. If a dike propagates at intermediate depths the strain will cause extensional deformation, eventually leading to graben formation. On the contrary, if a dike propagates to shallow depth and gains surface access at some points, a subsequent lava effusion will occur and the surface manifestation of the dike will be a fracture at which a dome may form (Jackson et al., 1997). Depending on the magma density relative to the density of the crust and the mantle (Wieczorek et al., 2001), and also on the stress state of the lithosphere, some dikes

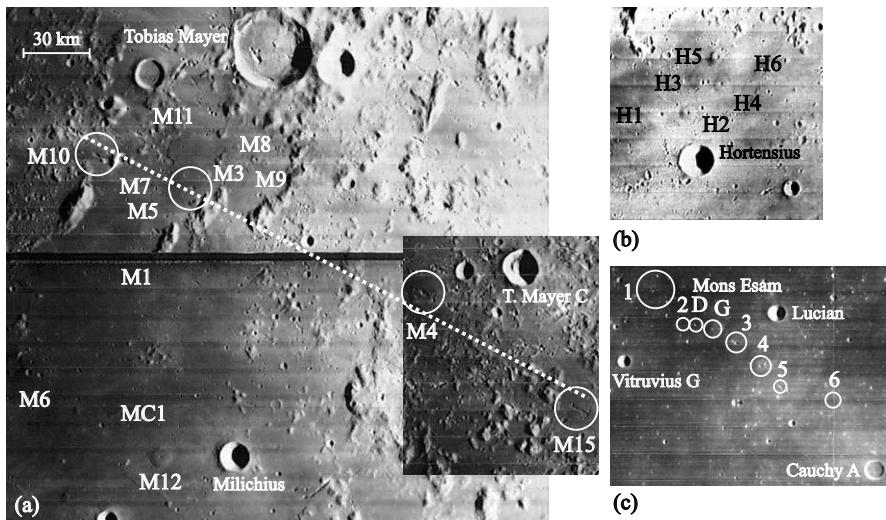


Fig. 7.15 (a) Lunar Orbiter images IV-133-H2 and IV-126-H2, showing the mare dome field between Milichius and Tobias Mayer. The dashed line connects four aligned large domes. (b) Lunar Orbiter image IV-133-H1, showing the domes near Hortensius. (c) Lunar Orbiter image IV-073-H2, showing the domes Diana (D) and Grace (G) and the domes of the Northern Tranquillitatis Alignment (NTA). Circles indicate dome diameters. The images are reproduced at uniform scale, north is to the top and west to the left.

erupt at the surface while others penetrate to depths shallow enough to produce linear graben.

A dike propagating to the surface erupted lavas that produced extensive mare units at high effusion rates. At the terminal stage of the eruption the mass flux decreased, resulting in the formation of domes by increased crystallisation in the magmas and decreasing temperatures. As discussed qualitatively by Weitz and Head (1999) and shown based on rheologic modelling by Wilson and Head (2003) and Wöhler et al. (2006b), the flatter mare domes were formed by lavas of low viscosity erupting at high effusion rates, favouring a low shield to develop, while steeper domes are favoured by lower mass fluxes and temperatures, resulting in higher viscosities and a high crystalline fraction. By comparing the time scale of magma ascent through a dike with the time scale on which heat is conducted from the magma into the host rock, Wöhler et al. (2007b) find evidence that the importance of magma evolution processes during ascent such as cooling and crystallisation increases with increasing lava viscosity. Accordingly, different degrees of evolution of initially fluid basaltic magma are able to explain the broad range of lava viscosities of five orders of magnitude found for lunar mare domes.

Head and Gifford (1980) provide a qualitative morphological classification scheme for lunar domes. Their classes 1–3 refer to largely symmetric volcanic features resembling terrestrial shield volcanoes, displaying comparably steep flanks (class 1), pancake-like cross-sectional shapes (class 2), and very low flank slopes

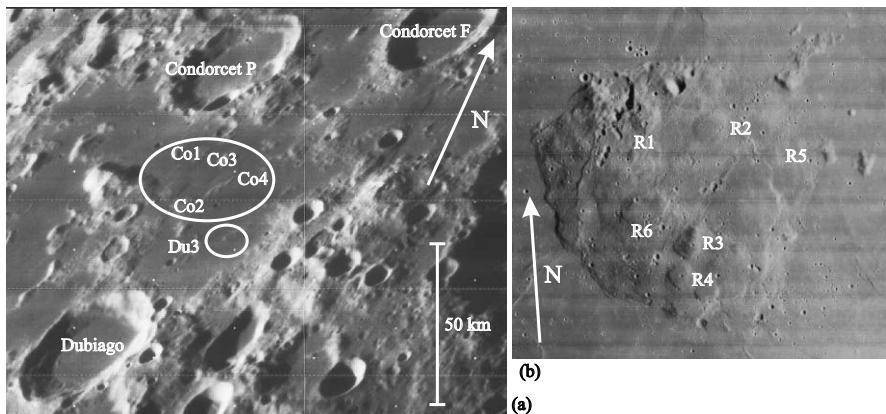


Fig. 7.16 (a) Lunar Orbiter image IV-178-H1, showing an oblique view on Mare Undarum with its dome field. North is indicated by the arrow in the upper right of the image, the undistorted image scale in the lower right. (b) Lunar Orbiter image IV-163-H2, showing the lunar dome complex Mons Rümker. The labels R1–R6 indicate individual volcanic edifices.

(class 3). Domes of class 4 are associated with mare ridges and may be related to shallow intrusions or structural warping and low-angle overthrusts associated with mare subsidence, while class 5 domes are assumed to be formed by lava mantling of pre-existing highland terrain. Classes 6 and 7 describe small elevated patches of highland surface surrounded by mare basalt and complex edifices of irregular outline, respectively. Wöhler et al. (2006b) introduce an extension of the definitions of classes 1–3 of the scheme by Head and Gifford (1980). They base the distinction between these shield-like volcanoes on their associated spectral and morphometric quantities (cf. Section 7.4.4).

7.4.2 Observations of Lunar Mare Domes

7.4.2.1 Spacecraft Observations of Lunar Mare Domes

Most lunar mare domes can only be observed under oblique illumination due to the low slopes of their flanks. Consequently, the Lunar Orbiter images mostly acquired at solar elevation angles between 20° and 30° display steeper effusive mare domes like e.g. some (but not all) of the domes near Milichius (cf. Fig. 7.15a), the Hortensius dome field (cf. Fig. 7.15b), the Marius Hills, the domes in Mare Undarum (cf. Fig. 7.16a), and the Mons Rümker complex (cf. Fig. 7.16b). The lower domes in Mare Tranquillitatis, however, are invisible in the Lunar Orbiter images, only some of their summit pits are apparent (cf. Fig. 7.15c). What is more, these images are not suitable for photogrammetric analysis aiming for generating topographic data due to the lack of geometric and photometric calibration; the relation between incident



Fig. 7.17 Apollo 15 orbital image AS15-97-13252, showing an oblique view of the lunar dome complex Mons Rümker from southern direction.

flux and pixel grey value is nonlinear and unknown because the images were acquired on photographic film scanned on board the spacecraft. An orbital Apollo 15 photograph of the large dome complex Mons Rümker situated in northern Oceanus Procellarum is shown in Fig. 7.17. It was acquired on film with a hand-held Hasselblad camera from the Apollo 15 command module and is therefore also unsuitable for a photometric evaluation. However, this image qualitatively reveals the large-scale shape of Mons Rümker, indicating that the western part of the plateau is more elevated than its eastern part (cf. Section 7.4.3.1).

While the Apollo images used for preparing the Lunar Topographic Orthophotomaps are virtually distortion-free, they cannot be used for photoclinometric analysis, again due to an unknown nonlinear relation between incident flux and film density, and many of them were acquired at high solar elevation angles. The morphometric properties of a limited number of lunar volcanic edifices have been derived from the Lunar Topographic Orthophotomaps by Pike and Clow (1981).

The fact that nearly all Clementine images were acquired at local lunar noon implies high illumination angles for the equatorial regions of the Moon, where the mare domes are situated. Consequently, for the important lunar dome fields these images are neither suitable for stereophotogrammetry nor photoclinometry or shape from shading.

7.4.2.2 Telescopic CCD Imagery

Due to the lack of photometrically calibrated spacecraft imagery acquired under oblique illumination we utilise telescopic CCD images for the determination of the morphometric properties of lunar domes. To acquire images of lunar domes, telescopes with apertures between 200 and 400 mm were utilised in combination with digital CCD video cameras of different types (Atik, ToUCam, Lumenera), relying on the acquisition technique described in Section 7.2.2.

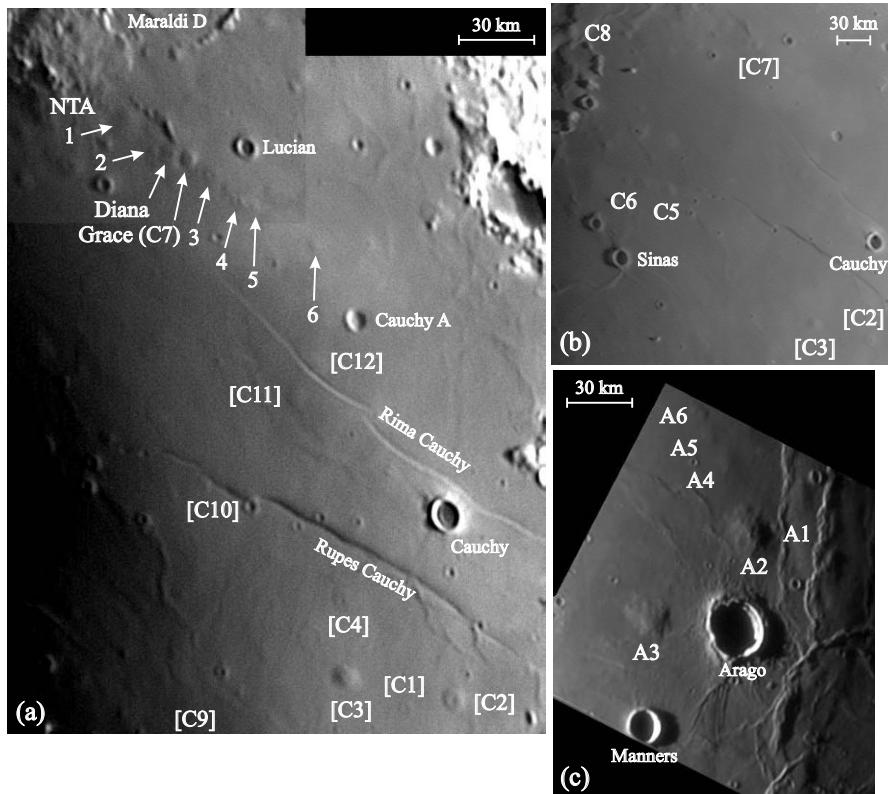


Fig. 7.18 (a) Telescopic CCD image of the northern part of Mare Tranquillitatis, showing the dome field around the crater Cauchy and the NTA domes. (b) Telescopic CCD image of the north-western part of Mare Tranquillitatis, showing further domes in the Cauchy region. (c) Telescopic CCD image of the region around Arago. Image courtesy P. Lazzarotti. North is to the top and west to the left.

The images shown in Figs. 7.18b and 7.19d were taken in the Johnson I band, a bandpass filter transmitting near-infrared wavelengths between 700 and 1100 nm, while the other telescopic images were acquired in integral visible light through a UV+IR block filter which transmits wavelengths between 400 and 700 nm. The telescopic CCD images in Figs. 7.18–7.22 are not geometrically rectified, which implies a non-uniform direction-dependent pixel scale. The scale bars in these figures therefore indicate the average pixel scale for each image. Labels without brackets denote that a three-dimensional reconstruction of the corresponding dome has been performed based on the respective image data, while labels in brackets are merely shown for comparison. For our set of lunar domes, the image sections used for three-dimensional reconstruction were extracted from the telescopic CCD images rectified to perpendicular view. A correction of the gamma value of the camera has been performed as described in Section 7.2.2.

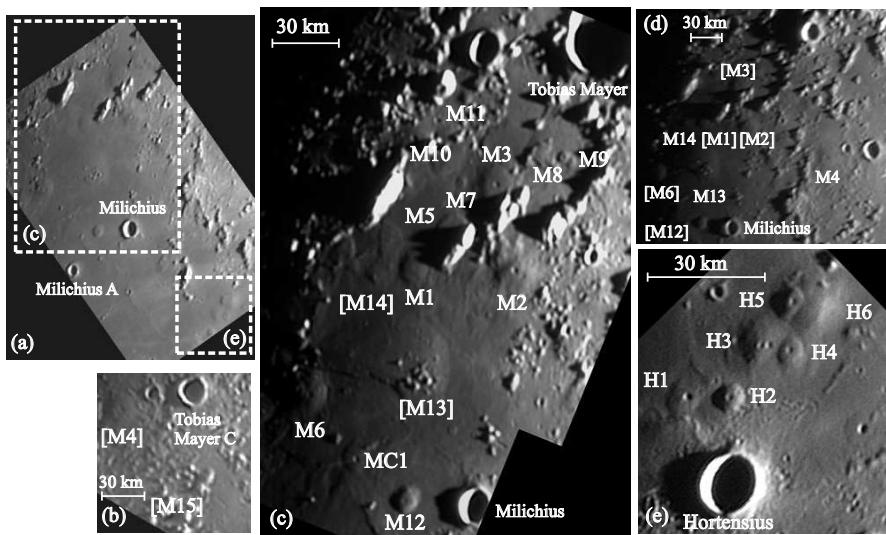


Fig. 7.19 Telescopic CCD images of the Hortensius and Milichius/Tobias Mayer dome fields. (a) Low-resolution image for orientation. (b) Region northeast of Milichius, showing the dome M4 and the elusive dome M15. Image courtesy R. Lena. (c) Dome field between Milichius and Tobias Mayer. Image courtesy J. Phillips. (d) Region north of Milichius. (e) Dome field north of Hortensius. Image courtesy M. Wirths. North is to the top and west to the left.

Fig. 7.18 shows the dome fields near Arago in western Mare Tranquillitatis and around Cauchy in central Mare Tranquillitatis as well as the dome chain at the northern border of Mare Tranquillitatis termed “Northern Tranquillitatis Alignment” (NTA) by Wöhler et al. (2007b). The dome fields situated in Mare Insularum near the craters Hortensius, Milichius, and Tobias Mayer are shown in Fig. 7.19, while Fig. 7.20 displays the dome field in Mare Undarum (Lena et al., 2008) and the large volcanic complex Mons Rümker in northern Oceanus Procellarum (Wöhler et al., 2007a). For comparison, several isolated domes situated outside the large dome fields are shown in Fig. 7.21, while Fig. 7.22 shows a dome at the southern rim of the crater Petavius which is associated with a pyroclastic deposit composed of dark material distributed across the surface by a violent volcanic eruption.

7.4.3 Image-based Determination of Morphometric Data

7.4.3.1 Preparation of DEMs

A robust method for estimating elevation differences on the lunar surface is the analysis of shadow lengths as described in Section 2.1. In the CCD images the diameter D of the dome and the length l of its shadow were measured in pixels. The corresponding image scale in kilometres per pixel was obtained by measuring the

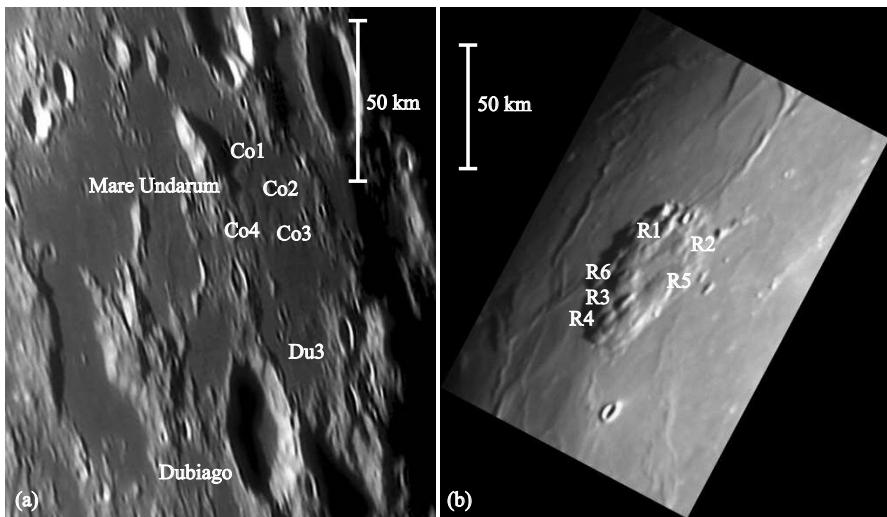


Fig. 7.20 (a) Telescopically obtained CCD image of the dome field in Mare Undarum. Image courtesy P. Lazzarotti. (b) Dome complex Mons Rümker in northern Oceanus Procellarum. Image courtesy K. C. Pau. North is to the top and west to the left.

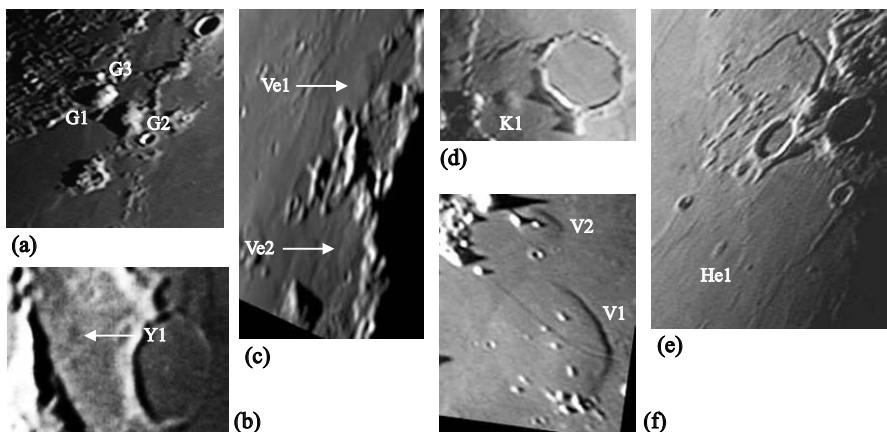


Fig. 7.21 Isolated mare domes in various regions of the Moon, examined for comparison (cf. also Table 7.3). (a) Highland domes Gruithuisen γ (G1), δ (G2), and NW (G3). (b) Mare dome Yerkes 1 near crater Yerkes in Mare Crisium. (c) Mare domes Vendelinus 1 and 2 near the western rim of Vendelinus. Image courtesy J. Phillips. (d) Mare dome Kies π (K1). (e) Mare dome Herodotus ω (He1). (f) Valentine dome V1 and its northern neighbour V2. Image courtesy K. C. Pau.

diameters of craters of known size in the field of view. The possible effect of sloping terrain has to be carefully considered, as the measured shadow length is longer for downward slope and shorter for upward slope, compared to horizontal terrain. A further problem arising especially for low domes is that even under strongly oblique illumination the shadow does not begin at the dome summit but at some point on

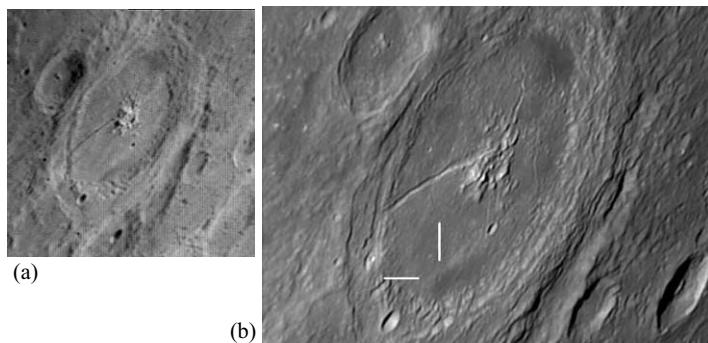


Fig. 7.22 (a) Image of the crater Petavius under moderately high solar illumination, showing two dark pyroclastic deposits located at its southern and northern rim, respectively. (b) Image of Petavius, showing the dome associated with the southern pyroclastic deposit (marked by white lines). Image courtesy P. Lazzarotti. In both images, north is to the top and west to the left.

Table 7.1 Flank slope and height values for several mare domes, determined using the shadow-based method by Ashbrook (1961).

Dome	Slope ($^{\circ}$)	Height (m)
C11	0.6	60
A2	1.5	310
H7	1.5	100
M11	3.0	150
M12	2.7	230

its flank, such that the determined height difference value merely represents a lower limit to the true dome height.

Ashbrook (1961) shows that the average slope of the dome flank equals the solar elevation angle when the shadow covers one quarter of the dome diameter, assuming a hemispherical shape of the dome. The observer determines the moment in time, corresponding to a solar elevation angle $\tilde{\mu}$, for which $l = D/4$ is given, leading to a dome height $h = (D/2)\tan\tilde{\mu}$. The Ashbrook method has primarily been devised for visual observations. The assumption of hemispherical dome shape, however, represents a significant restriction to its applicability. In this study, it was used to determine the heights of domes C11, A2, H7, M11, and M12 (cf. Table 7.1).

Photometric three-dimensional surface reconstruction methods are more generally applicable to the determination of the morphometric properties of lunar domes. In a first step we follow the photoclinometric approach according to Section 2.2.2.1, which consists of computing depth profiles along image rows. For all domes in our data set, the terrain is gently sloping, i.e. $|p|, |q| \ll 1$, the illumination is highly oblique, and the scene is illuminated nearly exactly from the east or the west, corresponding to $q_s = 0$. The Lunar-Lambert reflectance (7.6) thus depends much stronger on the surface gradient p in east-western direction than on the gradient q in north-southern direction, such that we may initially set $q = 0$. Note that this approximation is exact for cross-sections in east-west direction through the summit of

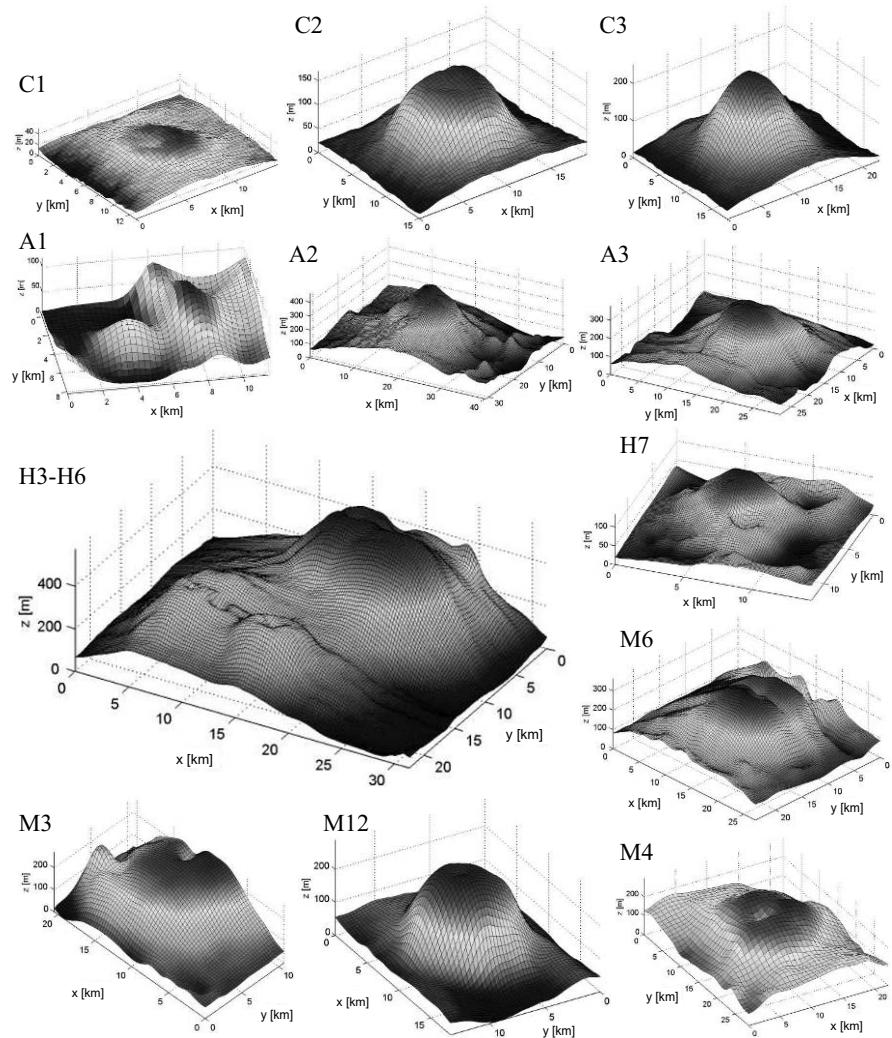


Fig. 7.23 Typical DEMs of domes in the Cauchy, Arago, Hortensius, and Milichius/Tobias Mayer dome fields. The dome identification according to Table 7.3 is indicated in the upper left of each plot, respectively. For C1, C2, and C3, the vertical axis is exaggerated by a factor of 50, for the other domes by a factor of 30.

a feature, while it is otherwise a reasonable approximation. A constant albedo ρ is assumed, which is justified for all examined domes as they are virtually indistinguishable from their surroundings in Clementine 750 nm images. The value of ρ is chosen such that the average surface slope over the region of interest is zero. Under these assumptions, Eq. (2.17) is solved for the surface gradient p_{uv} for each pixel with intensity I_{uv} .

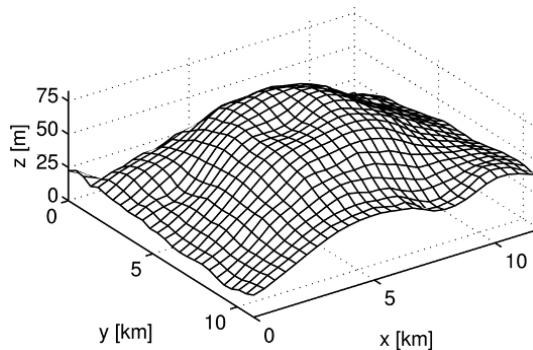


Fig. 7.24 DEM of the very low dome NTA3, viewed from south-western direction. The vertical axis is 50 times exaggerated.

The result of photoclinometry is used as an initialisation to the shape from shading scheme described in Section 2.2.2.2. The ground-based CCD images are affected by a slight blur due to atmospheric seeing. Hence, the observed image I_{uv} is assumed to be a convolution of the true image with a Gaussian point spread function (PSF) $G(\sigma)$ (cf. Chapter 3). The PSF radius σ is determined from the intensity profile of shadows cast by steep mountains, e.g. crater rims, where an abrupt transition from illuminated surface to darkness is expected. Convolving a synthetically generated abrupt change in intensity with a Gaussian PSF and comparing the result with the intensity profile observed in the image allows for an estimation of the PSF radius σ . A similar method is used by Baumgardner et al. (2000) to estimate the PSF for ground-based Mercury images, using the limb of the planetary disk as a reference. The intensity error term is then given by

$$e_i = \sum_{u,v} [I_{uv} - G(\sigma) * R(\rho, p_{uv}, q_{uv})]^2 \quad (7.8)$$

(Joshi and Chaudhuri, 2004). The uniform surface albedo ρ and approximate values for the surface gradients p_{uv} in east-west direction are known from Section 2.2.2.1, obtained there under the assumption of zero values of the surface gradients q_{uv} in north-south direction. The surface profile z_{uv} is computed such that the departure from integrability error term (2.26) is minimised. The surface gradients p_{uv} and q_{uv} are always initialised with zero values.

To obtain a DEM of the dome in the southern part of the crater Petavius shown in Fig. 7.22, the ratio-based photoclinometry approach described in Section 2.3.2 was employed, relying on the image shown in Fig. 7.22b and a further image acquired under lower solar illumination (Lena et al., 2006). The albedo map ρ_{uv} determined according to Eq. (2.49) was then inserted into the single-image shape from shading scheme described in Section 2.2.2, making use of the integrability constraint (2.26). The surface gradients p_{uv} determined by ratio-based photoclinometry based on Eq. (2.48) were used as initial values for the iterative update rule (2.31).

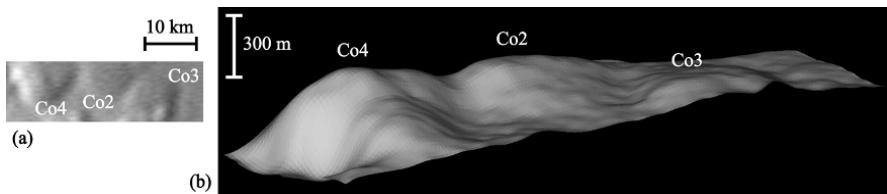


Fig. 7.25 (a) Telescopic CCD image of the domes Condorcet 2–4, situated in Mare Undarum. The image is rectified to perpendicular view, north is to the top and west to the left. Image courtesy P. Lazzarotti. (b) DEM of Condorcet 2–4, viewed from southwestern direction. The vertical axis is 10 times exaggerated, the curvature of the lunar surface has been subtracted.

Typical DEMs obtained for the regarded lunar domes are shown in Fig. 7.23, illustrating the rich variety of three-dimensional shapes occurring in the examined four lunar dome fields. For example, dome C1 near Cauchy is remarkably low but clearly effusive due to its summit pit. The nearby domes C2 and C3, traditionally known as Cauchy ω and τ , are steeper but quite different in shape, since C2 clearly displays a flattened top with a summit pit while C3 is of more conical shape. The large edifices A2 and A3, traditionally designated Arago α and β , are somewhat irregularly shaped, while dome A1 next to A2, situated near a mare ridge, is of regular conical shape. The domes near Hortensius have large and deep summit pits. This is true even for the comparably low dome H7 displaying a summit pit with a depth corresponding to about at least half the height of the dome. The domes near Milichius show remarkably manifold three-dimensional shapes. Examples are M3 with its pancake-like shape, M4 and M6 with their rougher and more strongly textured surfaces, and M12 showing a very regular, flattened shape which is quite similar to that of C2, with a central summit pit. The chained domes belonging to the Northern Tranquillitatis Alignment (NTA) are very low, except for Diana and Grace. As an example, a DEM of the dome NTA3 is shown in Fig. 7.24, illustrating its low profile. A summit pit is clearly visible on top of NTA3.

The domes Condorcet 1–4 and Dubiago 3 are situated in Mare Undarum, near the eastern limb of the apparent lunar disk (Lena et al., 2008). A DEM of a part of this dome field is shown in Fig. 7.25, revealing that the flank slope of Condorcet 4 is quite steep, while Condorcet 2 and 3 are lower and have a more pancake-like cross-sectional profile. Under the oblique viewing and illumination angles and low phase angles encountered for these domes, the reflectance behaviour of the lunar surface largely obeys the Lommel-Seeliger law—for the lunar macroscopic surface roughness estimated by Warell (2004), a value of $L(\alpha) = 0.9$ in the Lunar-Lambert reflectance function (7.6) is indicated by McEwen (1991). Hence, the dependence of the surface reflectance on the incidence angle and in turn on the surface slope is much less pronounced than for Lambertian reflectance. As a consequence, the heights of the domes in Mare Undarum obtained assuming Lambertian reflectance are too low by about a factor of two, compared to the values derived with the Lunar-Lambert reflectance law.

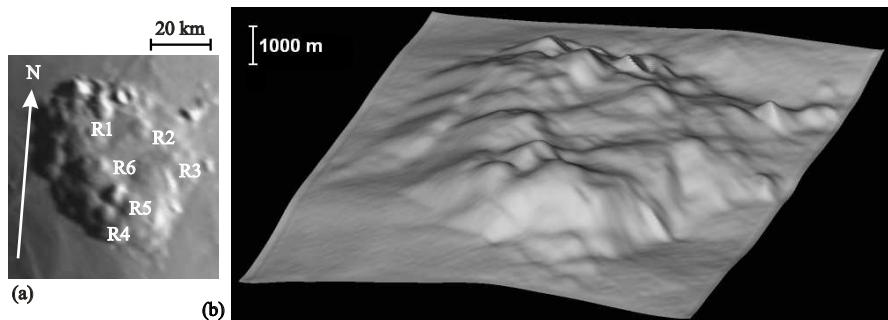


Fig. 7.26 (a) Telescopic CCD image of the dome complex Mons Rümker, situated in northwest Oceanus Procellarum. The individual edifices R1–R6 are indicated. The image is rectified to perpendicular view, north is indicated by the white arrow. Image courtesy K. C. Pau. (b) DEM of Mons Rümker, viewed from southeastern direction. The vertical axis is 10 times exaggerated, the curvature of the lunar surface has been subtracted.

The DEM of the large dome complex Mons Rümker shown in Fig. 7.26 was generated by employing the described combined photoclinometry and shape from shading technique as a multi-resolution approach to stabilise the convergence behaviour and facilitate at the same time the reconstruction of small-scale surface features. The DEM shows that the height of the plateau amounts to about 900 m in its western and northwestern part, 1100 m in its southern part, and 650 m in its eastern and northeastern part. The overall volume of erupted lava corresponds to about 1800 km^3 (cf. Section 7.4.3.2). About 30 individual domes on the Rümker plateau are reported by Smith (1974), six of which are sufficiently well resolved in the telescopic CCD image for morphometric evaluation. The DEM derived for Mons Rümker is qualitatively consistent with the Apollo 15 orbital photograph shown in Fig. 7.17.

The DEM of the dome at the southern rim of Petavius, situated in a region of non-uniform albedo characterised by a dark pyroclastic deposit, is shown in Fig. 7.27a. The cross-section through the summit pit (cf. Fig. 7.27b) reveals that the dome summit is elevated by 240 m above the surrounding surface. To the south of the pit, the terrain rises further up to a height of 530 m. The albedo map of the region covered by the DEM is shown in Fig. 7.27c, clearly indicating the dark pyroclastic material. The average flank slope of the higher summit south of the dome amounts to 3.1° , and parts of the flank are even steeper than 4° . Hence, it is probably too high and too steep to be of volcanic origin. The interpretation by Lena et al. (2006) is that the dome is placed adjacent to a hummocky, impact-related deposit, which is supported by the close proximity of this region to the rugged inner crater rim of Petavius.

A DEM of the Valentine dome is shown in Fig. 7.28. This large dome is of possibly intrusive origin. Its shape is asymmetric as its eastern edge is fairly pronounced and steep, while the western edge merges smoothly with the surrounding mare surface. The dome surface displays several protrusions and is traversed by a curved rille.

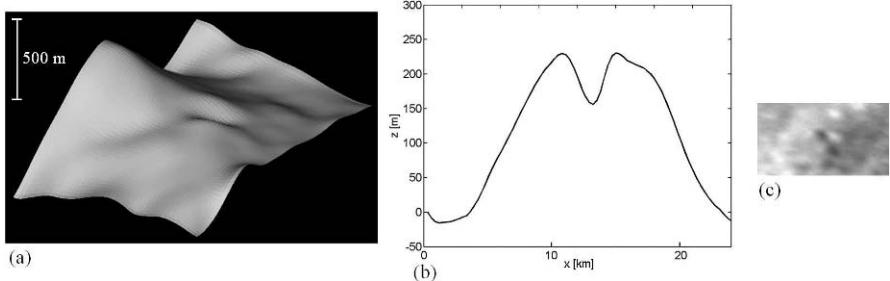


Fig. 7.27 (a) DEM of the dome at the southern rim of Petavius, viewed from northwestern direction. The vertical axis is ten times exaggerated. (b) Cross-section in east-western direction through the summit pit of the dome. (c) Albedo map obtained by Eq. (2.49).

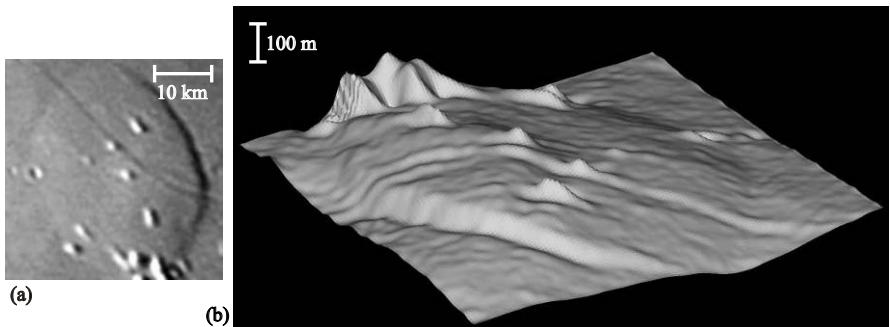


Fig. 7.28 (a) Telescopic CCD image of the Valentine dome, situated at the western border of Mare Serenitatis and possibly being of intrusive origin. The image is rectified to perpendicular view, north is to the top and west to the left. Image courtesy K. C. Pau. (b) DEM of the Valentine dome, viewed from northeastern direction. The vertical axis is 30 times exaggerated, the curvature of the lunar surface has been subtracted.

7.4.3.2 Features Derived from the DEMs

The image scale (in kilometres per pixel) is obtained for each image relying on craters of known diameters. The dome diameter D is measured in pixels; for non-circular domes an appropriate average value is used. The height h of a dome is obtained by measuring the depth difference in the reconstructed three-dimensional profile between the dome summit and the surrounding surface, taking into account the curvature of the lunar surface. The average flank slope ζ is determined according to

$$\zeta = \arctan \frac{2h}{D}. \quad (7.9)$$

The dome volume V is computed by integrating the reconstructed three-dimensional profile over an area corresponding to a circular region of diameter D around the dome centre. If only a part of the dome can be reconstructed, as it is the case for a few domes in the Milichius field due to shadows cast on the dome surfaces by

nearby hills, the volume is estimated based on a cross-section through the centre of the dome, assuming rotational symmetry. A rough quantitative measure for the shape of the dome is given by the form factor

$$f = \frac{V}{\pi h(D/2)^2}. \quad (7.10)$$

It is $f = 1/3$ for domes of conical shape, $f = 1/2$ for parabolic shape, $f = 1$ for cylindrical shape, and intermediate values occur for hemispherical shape. For effusive domes, it is possible in some cases to measure the diameter D_c of the summit pit in the telescopic images. However, it should be noted that for several domes with elongated summit pits or fissures (Wöhler et al., 2007b), these diameters are average values. The reconstructed three-dimensional profiles also yield values for the depth d_c of the summit pit, but due to insufficient resolution, the obtained values must be regarded as lower limits to the true depth rather than accurate measurements. For the domes H3 and M6 having two summit pits, respectively, only the larger pit is considered in the subsequent analysis.

The spectral properties derived from Clementine UVVIS data and the morphometric quantities derived from the telescopic CCD images and reconstructed three-dimensional profiles of the lunar domes examined in this study are listed in Tables 7.3 and 7.4. If possible, the examined domes are identified with preliminary names assigned by Head and Gifford (1980), consisting of the name of the nearest crater and a number, as well as traditional, mostly unofficial designations for some of the well-known domes. A similar identification scheme is used in the lunar dome catalogue published by Kapral and Garfinkle (2005).

7.4.3.3 Error Estimation

In the context of image-based three-dimensional reconstruction of lunar domes, three main sources of error can be identified. The parameter $L(\alpha)$ of the reflectance function is not exactly known and may show variations over the surface for different terrain types. We assume a standard error of $L(\alpha)$ of 0.15, an error range that also includes larger values of the macroscopic surface roughness $\bar{\theta}$ of up to 20° (McEwen (1991), Fig. 16 therein). The radius σ of the Gaussian PSF for our telescopic CCD images could be estimated at an accuracy of about 0.5 pixels. We have found that the uncertainties in $L(\alpha)$ and σ affect the measured height values of the lunar domes by no more than a few metres and can therefore be regarded as irrelevant. We expect the influence of the PSF to become more important for strongly wrinkled surfaces.

A more important issue is the slight nonlinearity of the CCD sensor, which is compensated by the gamma calibration procedure described in Section 7.2.2. The uncertainty of the determined gamma values approximately amounts to 0.05 for the Atik and the ToUCam CCD cameras. The Lumenera CCD camera has a linear ($\gamma = 1.0$) characteristic curve.

The uncertainty in γ results in a relative standard error of the dome height h of 10 percent, which is independent of the height value itself. The dome diameter D can be measured at an accuracy of 5 percent, since the outer rim is not well-defined for most domes. Based on experiments, we found that these uncertainties in h and D lead to a typical standard error of the edifice volume V , computed by integration of the three-dimensional profile, of about 20 percent.

Furthermore, we have examined possible systematic deviations between the true values of the dome height and volume and the results obtained with the described photoclinometry and shape from shading method. No sufficiently accurate morphometric lunar dome properties are available as ground truth values, and our method cannot be directly applied to images of terrestrial volcanoes (for which such data are available) since it is usually not possible to describe the reflectance behaviour of the terrestrial surface by a relatively simple relationship like the Lunar-Lambert law. We therefore performed experiments on synthetic ground truth data. We regarded rendered images of a dome of parabolic shape with a diameter of 12 km and a height of 200 m, yielding an average flank slope of 1.91° , a form factor of 0.5 (Wilson and Head, 2003), and an edifice volume of 11.31 km^3 . Image scale was set to 300 m per pixel. We assumed the synthetic dome to be located on the lunar equator at 10 different equally spaced selenographic longitudes λ_s between -45° and $+45^\circ$. We rendered images of the dome under sunrise illumination at solar elevation angles μ of 4° and 6° for each value of λ_s . Assuming zero libration, the phase angle amounts to $\alpha = 90^\circ + \lambda_s - \mu$, and the Lunar-Lambert parameter $L(\alpha)$ was chosen according to McEwen (1991) for each rendered image. For all examined configurations of λ_s and μ , the root mean square deviation between the reconstructed and the true three-dimensional profile is smaller than 5 m. The deviations between the dome height and volume determined from the synthetic images based on our photoclinometry and shape from shading approach and the respective ground truth values are always smaller than 1 percent. Hence, we may conclude that our reconstruction results are not perceptibly affected by systematic errors.

The dome height values independently obtained based on shadow length measurement according to Ashbrook (1961) may also be used for comparison, yielding a good consistency with the corresponding values obtained by photoclinometry and shape from shading (cf. Table 7.1).

When no ground truth is available, the consistency of height measurements across several images is a further good indicator of their reliability. We have determined the height values of the domes M3 and M12 based on the images in Figs. 7.19a and 7.19b, which were acquired with different telescopes and CCD cameras from different locations. The obtained dome heights are 190 m and 230 m in Fig. 7.19a, compared to 170 m and 210 m in Fig. 7.19b, respectively, which is a reasonable correspondence.

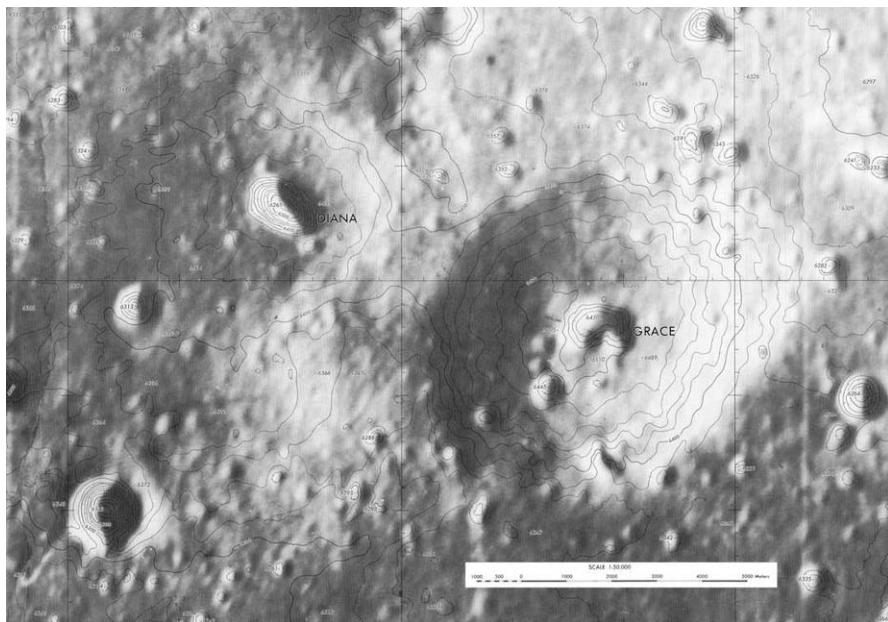


Fig. 7.29 High-resolution Lunar Topophotomap LT 61A2S1 (50), showing the dome pair Diana and Grace situated at the northern border of Mare Tranquillitatis. The contour interval corresponds to 20 m.

7.4.3.4 Comparison to Previous Height Measurements

Not too much topographic data about lunar domes have been published prior to this study. The most significant contribution to this field has been provided by Brungart (1964), who compiled a catalogue of 261 domes reporting their coordinates, diameters, heights, flank slopes, and morphological characteristics, utilising the Orthographic Lunar Atlas (Kuiper, 1961) that consists of telescopic photographs. Brungart (1964) determines values for the dome heights and flank slopes based on shadow length measurement but at the same time characterises the obtained results as merely representing order of magnitude estimates. As an example, for Arago α and β (entries no. 3 and 4 in the Brungart catalogue) a height of 700 m and 800 m with an average flank slope of 5.5° and 6.0° is reported, respectively. Our results indicate lower heights of 330 m and 270 m along with flank slopes of 1.5° and 1.3° , respectively. For Milichius π (entry no. 190), Brungart (1964) states a height of 742 m with an average flank slope of 9° . We estimated the height of this dome with the Ashbrook method, yielding an average slope angle of 2.7° and a height of 230 m, which is found to be in excellent agreement with the photoclinometry and shape from shading analysis. If the height estimates by Brungart (1964) for these three domes were correct, the domes would have to display shadows of length 13.5 km, 9.9 km, and 15.7 km in Figs. 7.18c and 7.18a, respectively, which is clearly not the case. Similarly, the height estimates for the domes H1–H6 near Hortensius are sys-

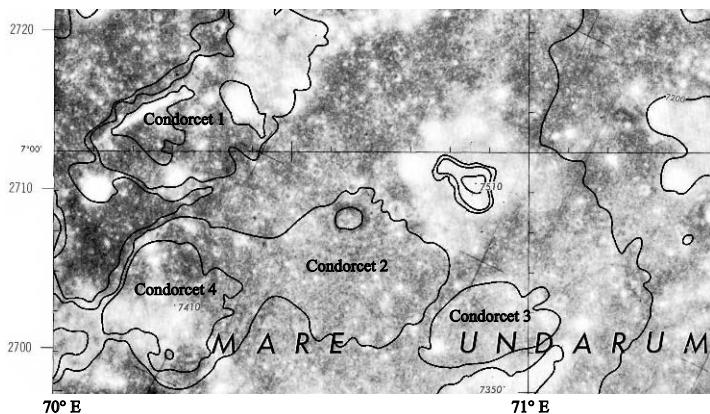


Fig. 7.30 Section of Lunar Topographic Orthophotomap LTO 63D1. The circumferences of the domes Co2, Co3, and Co4 as well as the southern and the northwestern border of Co1 appear as contour lines (cf. Fig. 7.25). The contour interval corresponds to 100 m.

tematically higher than ours by a factor of two and more, and for the flank slopes values of up to 20° are stated. The height and flank slope values given by Brungart (1964) for these domes would imply large shadows of a length of up to 25 km for the domes in Fig. 7.19c, which do not exist. From these results we conclude that the height estimates by Brungart (1964) are systematically too high by a significant amount. This finding clearly shows how difficult it is to measure accurately shadow lengths in high-contrasted photographic reproductions, since shading effects are easily confused with black shadows, leading to a systematic over-estimation of height and flank slope.

Accurate height measurements of lunar volcanic edifices are presented by Pike and Clow (1981). Their values are based on the Lunar Topographic Orthophotomaps with a standard elevation error of 30 m. Their data set primarily contains lunar cones, but they have determined the height of one mare dome of our data set, Cauchy ω , stating a value of 116 m. The height of 125 m we obtained by means of photoclinometry and shape from shading analysis is in good agreement. For comparison, the corresponding height value by Brungart (1964) of 306 m (entry no. 36) is again too large.

The domes Diana and Grace appear in the high-resolution Lunar Topophotomap LT 61A2S1 (50) (cf. Fig. 7.29). For Diana, this elevation map yields an elevation difference with respect to the surrounding mare plain of 80 m on the western and 50 m on the eastern flank. The corresponding values for Grace are 100 m on the western and 140 m on the eastern flank. Based on the telescopic CCD image shown in Fig. 7.18a, we obtained height values of 70 m and 140 m for Diana and Grace, respectively, which is in good accordance with the result derived from orbital imagery.

The domes in Mare Undarum are incorporated in Lunar Topographic Orthophotomap LTO 63D1 (cf. Fig. 7.30). The fairly complex structure of the dome Con-

dorcer 1 with an elongated non-volcanic mountain at its north-western rim is clearly revealed by this map. According to the elevation contours, the height of the volcanic edifice amounts to 100–200 m. The domes Condorcet 2 and 3 appear as single contour lines, thus indicating approximate heights of 100 m. For the dome Condorcet 4, the topographic map yields an elevation difference of roughly 200–300 m between its summit and its bottom. These height values, however, are very approximate since the resolution of the topographic map is too low to obtain well-defined values for the elevation of the mare plains surrounding the domes. The lateral resolution of LTO 63D1 is not sufficiently high to display the three-dimensional shapes of the domes in some detail, while the vertical resolution is too low to yield accurate dome heights. Based on our combined photoclinometry and shape from shading technique we obtained height values of 150, 130, 110, and 270 m for Condorcet 1–4, respectively, which are consistent with the very approximate elevation differences derived from LTO 63D1 for these domes.

7.4.4 Geophysical Insights Gained from Topographic Data

In this closing section we go somewhat beyond the domain of determining morphometric data of lunar domes and give a concise discussion of the geophysical conclusions that can be drawn from the determined morphometric dome properties. An in-depth analysis of these aspects and an extensive literature overview are provided by Wöhler et al. (2006b) and Wöhler et al. (2007b). This summarising section is to exemplify that topographic mapping of planetary bodies is no end in itself but—especially for the Moon—is a precondition for a deeper understanding of the geophysical processes that occurred in their early geologic history and formed their surfaces.

7.4.4.1 Classification of Lunar Domes

Previous classification schemes for lunar domes are based on a qualitative description of dome shape and geologic setting rather than morphometric quantities (Head and Gifford, 1980). The assignment of a dome to a certain class thus remains ambiguous in many cases. Hence, a classification scheme has been developed by Wöhler et al. (2006b) and Wöhler et al. (2007b) which is complementary to the scheme introduced by Head and Gifford (1980) in that it subdivides lunar mare domes according to their morphometric properties in a quantitative way.

Additionally, the spectral properties of the dome surfaces are taken into account, regarding the absolute reflectance R_{750} at 750 nm wavelength and two spectral ratios derived from multispectral imagery acquired by the Clementine spacecraft (Eliason et al., 1999). The first spectral ratio R_{415}/R_{750} of the reflectances measured at wavelengths of 415 nm and 750 nm is highly correlated with the titanium dioxide (TiO_2) content of the soil, which in turn has an influence on lava viscosity. The second

Table 7.2 Approximate spectral and morphometric properties characterising the dome classes as defined by Wöhler et al. (2006b) and Wöhler et al. (2007b).

Class	R_{415}/R_{750}	ζ ($^{\circ}$)	D (km)	V (km^3)
A	> 0.64	0.3–1.0	5–13	< 3
B ₁	0.55–0.64	2.0–5.4	6–15	5–32
B ₂	0.55–0.64	1.3–1.9	8–15	2–21
C ₁	0.55–0.60	0.6–1.8	13–20	7–50
C ₂	0.60–0.64	1.0–2.5	8–17	4–17
D	> 0.64	1.3–1.5	≈ 25	40–67
E ₁	0.58–0.62	2.0–4.0	< 6	< 1.2
E ₂	0.58–0.62	< 2.0	< 6	< 1.2
G	0.55–0.60	> 6.0	7–30	20–390

spectral ratio R_{950}/R_{750} is related to the strength of the so-called mafic absorption band, representing a measure for the iron oxide (FeO) content of the soil and being also sensitive to the optical maturity of mare and highland materials—the optical maturity of the lunar regolith increases with the period of time the soil has been exposed to the solar wind. These spectral characteristics allow for the mutual distinction between different types of basaltic mare soils and highland soils. In the field of geophysical research, a large body of literature about the spectral characterisation of lunar soils is available (Adams and McCord, 1970; McCord et al., 1972; McCord and Adams, 1973; Charette et al., 1974; Burns et al., 1976; McCord et al., 1976; Melendrez et al., 1994; Staid et al., 1996; Lucey et al., 1998; Gillis and Lucey, 2005).

The new dome classification scheme introduced by Wöhler et al. (2006b) and refined by Wöhler et al. (2007b) is based on the following features: Reflectance R_{750} at 750 nm wavelength, reflectance ratios R_{415}/R_{750} and R_{950}/R_{750} , flank slope ζ , diameter D , height h , volume V , and form factor f . A grouping of the domes based on a principal component analysis in the eight-dimensional feature space, inspired by Pike (1978), yields five classes A–E. Classes A, B, C, and E denote morphologically simple and likely monogenetic edifices which presumably formed during a single eruption event and thus do not show individual lava flows or other traces of several subsequent eruption events.

Class A domes display small to moderate diameters between 5 and 13 km with very low flank slopes and volumes and were formed by spectrally strongly blue lavas of high R_{415}/R_{750} spectral ratio. Typical representatives of class A are the domes A4–A6 and C1, C4, and C6, all situated in Mare Tranquillitatis.

Class B domes have small to moderate diameters between 6 and 15 km and were formed from lavas of low to moderate R_{415}/R_{750} spectral ratio. Steep and voluminous class B domes with flank slopes larger than 2° are assigned to subclass B₁ while the lower edifices with flank slopes below 2° make up subclass B₂. The domes H2–H6 near Hortensius, M6, M11, and M12 near Milichius, Herodotus ω in Oceanus Procellarum, Co4 in Mare Undarum, and R1, R3, and R4 on the Rümker plateau belong to class B₁. Typical representatives of class B₂ are H1 and H7 near Hortensius, M4 near Milichius, Co2 in Mare Undarum, and R2 and R6 on the Rümker plateau.

Class C domes are larger (diameter between 8 and 20 km) with relatively low flank slopes typically below 2°. Edifices formed from spectrally red lavas of low to moderate R_{415}/R_{750} ratio, having large diameters between 13 and 20 km and large edifice volumes of several tens of km^3 , are assigned to subclass C₁, while spectrally bluer domes of moderate to high R_{415}/R_{750} ratio, smaller diameters between 8 and 13 km, and lower edifice volumes of less than 17 km^3 are assigned to subclass C₂. The domes M1–M3, M5, M10, and M15 near Milichius, Ve1 and Ve2 in eastern Mare Fecunditatis, Kies π in Mare Nubium, Co3 in Mare Undarum, and Pe1 in the southern part of the floor of Petavius belong to class C₁. Dome class C₂ is represented by the domes C2, C3, C5, C8, and Grace in Mare Tranquillitatis.

A further group made up by domes with exceptionally small diameters below 6 km and very low edifice volumes below 1.2 km^3 represents intermediate objects between lunar domes and lunar cones, small volcanic edifices formed by explosive volcanic eruptions. These small domes are denoted by class E according to Wöhler et al. (2007b). In analogy to class B, class E is further subdivided into subclasses E₁ and E₂, denoting the steep-sided (flank slope larger than 2°) and the low edifices of this class, respectively. The domes M8 and M9 near Milichius belong to class E₁, while the domes M7 near Milichius and R5 on the Rümker plateau belong to class E₂.

Not for all examined lunar domes an unambiguous class assignment can be made. In northern Mare Tranquillitatis, some domes are classified as B₂–C₁ (NTA3 and NTA4, also Co1 in Mare Undarum) or A–E₂ (NTA5 and NTA6). The dome Diana is intermediate between the classes A, C₂, and E₂, and NTA1 is assigned to classes A–C₁–C₂.

Class D represents edifices of more complex morphology, displaying large diameters, low flank slopes, and very high edifice volumes, such as the domes Arago α and β (cf. Fig. 7.18) which probably formed during several effusion phases.

Table 7.2 summarises the spectral and morphometric dome properties associated with the different mare dome classes (Wöhler et al., 2007b). For completeness, we have added a further class, G, to Table 7.2, representing the large and steep highland domes found in the Gruithuisen and Mairan region. To give an impression of the horizontal and vertical extent of the mare dome classes A–C and E denoting monogenetic edifices, Fig. 7.31 shows images and cross-sections of typical representatives of these classes, also illustrating the intermediate domains between them. The class assignments for all examined lunar domes are given in Tables 7.3 and 7.4².

² We adopt the class assignments by Wöhler et al. (2006b) and Wöhler et al. (2007b), except for the domes C12 in Mare Tranquillitatis and M14 in the Milichius/Tobias Mayer region, which were originally classified as domes of possibly intrusive origin. Here, we revise the classification of C12 and assign it to class E₂ based on its morphological similarity (small diameter, circular outline) to the dome M7, which is clearly effusive due to the presence of a summit vent. For the dome M14, close inspection of Lunar Orbiter image IV-133-H2 reveals a smooth-rimmed pit on its surface which might be interpreted as an effusive vent. According to low-sun telescopic images, M14 has a somewhat irregular outline and surface texture, which is in contrast to the domes of possibly intrusive origin. However, we cannot provide an unambiguous class assignment for M14 based on the available data.

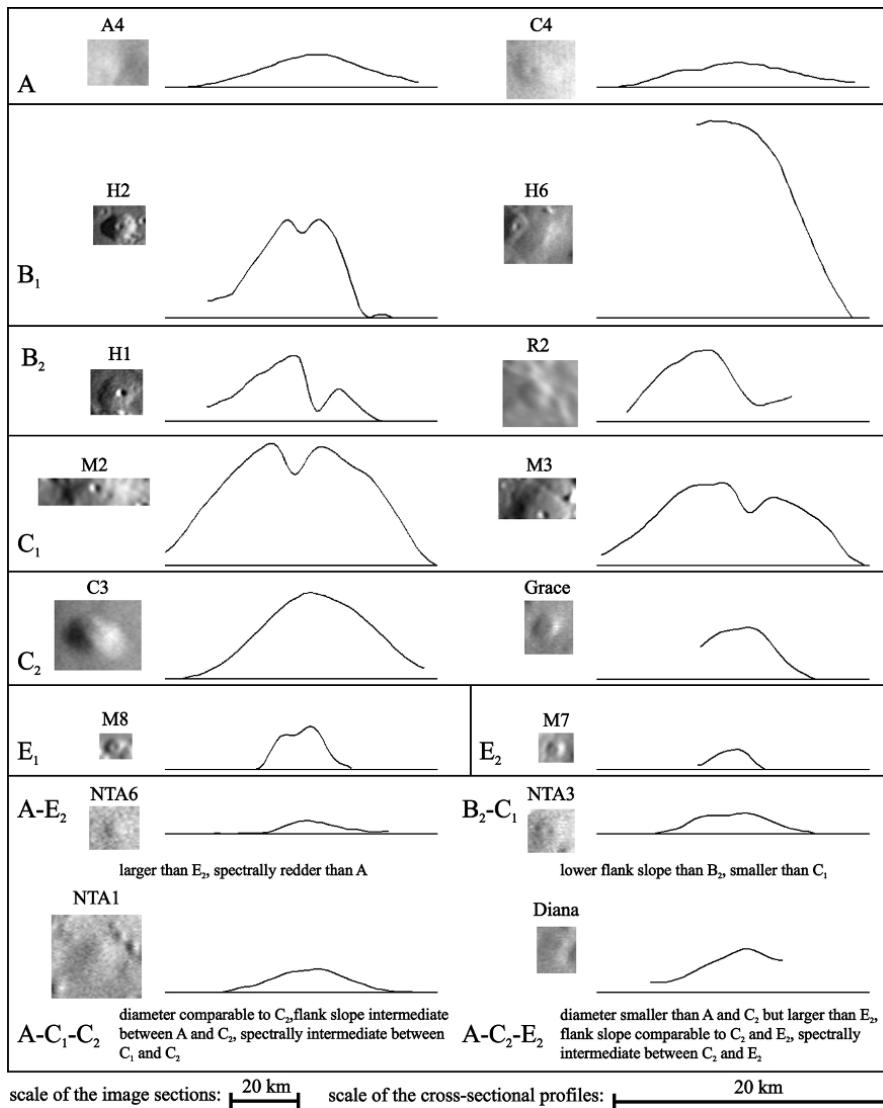


Fig. 7.31 Images and cross-sections of typical representatives of the monogenetic dome classes (cf. Table 7.2) and the continuum between them, adapted from Wöhler et al. (2007b). The image sections were extracted from the telescopic images shown in Figs. 7.18–7.22. The horizontal extension of the cross-sections is 20 km, their vertical axes are 30 times exaggerated.

7.4.4.2 Mechanisms of Dome Formation

The evident large variety of dome shapes raises broad questions concerning the source regions of the various dome types and the corresponding implications for

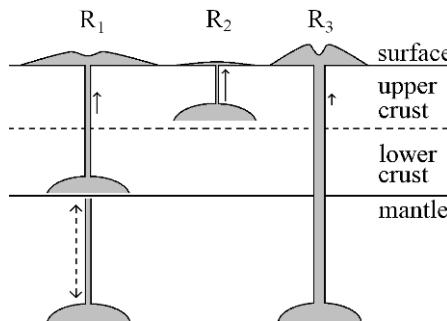


Fig. 7.32 Illustration of rheologic groups R_1 – R_3 . The indicated dome diameters and heights, magma source depths, and dike widths are not to scale but illustrate relative properties. Solid arrows indicate magma rise speed.

local and regional lunar volcanism, the reasons why certain types of lunar domes are concentrated in certain areas of the lunar surface, why domes tend to be aligned, and which differences in the lunar interior are responsible for the different lunar dome properties observed on the surface.

It is suggested by Wöhler et al. (2007b) that the ascent of dome-forming lavas was guided by the same internal stress fields induced by major basin impacts which also generated crustal fractures and faulting. While terrestrial crustal fractures are mostly due to tectonic processes, systems of crustal fractures on the Moon were generated by major impact events. Basin impacts caused shock waves that propagated through the lunar surface (Spudis, 1993). These shock waves induced faulting in the subsurface bedrock and reactivated faults caused by preceding impact events. Radial and concentric surface manifestations of faulting in the subsurface layers are apparent around several lunar impact basins (Wilhelms, 1987). Mechanisms of the intrusion of dikes, i.e. thin sheets of magma generating fractures in the crust during ascent, in the presence of crustal stress fields is modelled by Rubin (1993). Lunar Orbiter imagery (cf. Fig. 7.15) shows that the low domes in the Milichius/Tobias Mayer region display elongated summit pits or fissures oriented in parallel, radial to the Insularum basin and the Eastern Procellarum basin (Spudis, 1993). The locations of four of these domes are forming a linear chain of 210 km length in the same direction. Similarly, the eight low domes of the Northern Tranquillitatis Alignment form a linear chain of 100 km length aligned radial to the Imbrium basin. Six of these domes display elongated summit vents oriented in the same direction. Similarly, the domes Condorcet 1–3 in Mare Undarum are aligned radial to the Crisium basin (Lena et al., 2008). The explanation given by Wöhler et al. (2007b) for these observations is that the domes formed along crustal fractures generated by major impact events, hence running radially with respect to the basin locations (Rubin, 1993). In this context, the elongated summit pits are interpreted to indicate the direction of the dike through which the magma ascended to the lunar surface.

According to the work by Wilson and Head (2003), knowledge about the morphometric properties of lunar domes (diameter, height, volume) allows to estimate

Table 7.3 Cauchy, Arago, Hortensius, Milichius/Tobias Mayer dome fields, isolated domes.

Dome	long. (°)	lat. (°)	R_{750}	R_{415}/R_{750}	R_{950}/R_{750}	ζ (°)	D (km)	h (m)	V (km ³)	f	D_C (km)	d_C (m)	Name	Class
Effusive mare domes														
C1	37.48	7.11	0.0863	0.6508	1.0588	0.35	8.1	25	0.65	0.50	2.6	> 15	Cauchy 5	A
C2	38.32	7.23	0.0909	0.6377	1.0567	1.17	12.2	125	7.4	0.50	2.6	> 20	Cauchy 1 (Cauchy ω)	C ₂
C3	36.73	7.58	0.0919	0.6318	1.0608	1.28	17.0	190	12	0.30			Cauchy 2 (Cauchy τ)	C ₂
C4	36.78	8.85	0.0820	0.6621	1.0547	0.43	13.3	50	3.2	0.46	2.2	> 10		A
C5	33.02	10.56	0.0902	0.6452	1.0489	1.03	11.1	100	5.4	0.56				C ₂
C6	31.97	10.76	0.0842	0.6558	1.0685	0.74	7.7	50	0.99	0.43				A
C8	30.72	14.40	0.0975	0.6346	1.0465	2.5	12.5	270	17	0.51				C ₂
C12	37.2	12.37	0.0904	0.6165	1.0681	0.45	6.3	25	0.48	0.61				E ₂
A1	21.96	7.66	0.0863	0.6703	1.0577	0.88	5.6	45	0.32	0.29				A
A2	21.7	7.56	0.0886	0.6621	1.0565	1.5	25.4	330	67	0.40			Arago 2 (Arago α)	D
A3	20.07	6.24	0.0912	0.6738	1.0578	1.3	23.6	270	39.9	0.34			Arago 1 (Arago β)	D
A4	21.27	8.65	0.0806	0.6681	1.0762	0.66	11.1	65	3.2	0.50			Arago 3	A
A5	20.96	8.88	0.0816	0.6719	1.0731	0.59	8.4	45	1.4	0.58			Arago 4	A
A6	20.79	9.22	0.0805	0.6791	1.0751	0.58	9.5	50	1.8	0.50			Arago 5	A
H1	28.41	7.18	0.1308	0.6203	1.0278	1.89	8.48	140	3.4	0.44	2.6	> 95	Hortensius 1	B ₂
H2	-28.01	7.12	0.1281	0.6225	1.0131	3.45	7.63	230	6.3	0.60	1.9	> 45	Hortensius 2	B ₁
H3	-27.78	7.59	0.1174	0.6105	1.0197	2.05	12.3	220	17	0.65	1.7	> 65	Hortensius 3	B ₁
H4	-27.51	7.47	0.1209	0.6154	1.0303	3.21	6.78	190	5	0.73	1.9	> 40	Hortensius 4 (Hortensius σ)	B ₁
H5	-27.54	7.87	0.1183	0.6066	1.0287	5.39	8.48	400	18	0.81	1.6	> 50	Hortensius 5	B ₁
H6	-27.34	7.82	0.1237	0.6118	1.0213	3.57	12.5	390	32	0.66			Hortensius 6	B ₁
H7	-25.17	6.07	0.1463	0.6149	1.0156	1.47	7.82	100	2.4	0.51	2.4	> 45		B ₂
M1	-31.58	12.76	0.1090	0.6060	1.0444	0.86	13.4	100	8.2	0.58	2.4		Tobias Mayer 1	C ₁
M2	-30.05	12.79	0.1267	0.5837	1.0284	1.8	20.1	320	50	0.49	3.7	> 60	Tobias Mayer 4	C ₁
M3	-30.43	13.78	0.1063	0.5923	1.0538	1.4	15.6	190	12	0.34	3.1	> 70	Tobias Mayer 3	C ₁
M4	-27.39	12.04	0.1383	0.5980	1.0246	1.27	15.3	170	21	0.69	4.0	> 35	Tobias Mayer 6	B ₂
M5	-31.01	13.24	0.1181	0.5978	1.0184	0.6	15.3	80	6.8	0.46			Tobias Mayer 2	C ₁
M6	-32.74	11.48	0.1024	0.6068	1.0272	1.34	19.7	230	33	0.47	3.0	> 70		B ₁
M7	-30.96	13.75	0.1168	0.6151	1.0233	0.99	5.2	45	0.37	0.39	1.8			E ₂
M8	-29.4	14.00	0.1214	0.5951	1.0428	3.46	4.3	130	1.2	0.65	1.8			E ₁
M9	-29.4	14.00	0.1214	0.5951	1.0429	3.15	4.0	110	0.88	0.64	0.9			E ₁
M10	-31.7	14.06	0.1227	0.5976	1.0329	0.84	19.0	140	21	0.54	3.4	> 35	Tobias Mayer 7	C ₁
M11	-31.08	14.77	0.1186	0.5908	1.0297	2.8	6.0	150	2.3	0.54	2.4			B ₁
M12	-31.2	10.88	0.1156	0.6116	1.0117	2.72	9.7	230	11	0.63			Milichius 1 (Milichius π)	B ₁
M14	-32.13	12.76	0.1056	0.6094	1.0374	0.27	14.8	35	1.7	0.28				
M15	-25.17	6.07	0.1411	0.5920	1.0389	0.60	21.0	110	17	0.45	2.5	> 20		C ₁
Domes of possibly intrusive origin														
C9	34.66	7.06	0.0860	0.6053	1.0657	0.13	13.3	15	0.47	0.23				
C10	35.19	10.00	0.0902	0.6414	1.0582	0.3	19.2	50	10	0.72				
C11	36.75	11.06	0.0853	0.6656	1.0566	0.7	12.2	75	6.4	0.73				
M13	-31.53	11.68	0.1024	0.6119	1.0377	0.41	27.8	100	15	0.25				
Highland domes (for comparison)														
G1	-40.38	36.43	0.1720	0.5648	1.0679	8.3	19.0	1320	187	0.5			Gruithuisen γ	G
G2	-39.42	36.11	0.1552	0.5506	1.0583	6.9	27.0	1630	378	0.5			Gruithuisen δ	G
G3	-40.86	36.92	0.1492	0.5604	1.0647	9.1	7.5	1020	23	0.5			Northwest dome	G
Effusive mare domes in other lunar regions (for comparison)														
Y1	49.96	14.82	0.1588	0.5912	1.0654	1.31	9.6	110	4.8	0.60				B ₂
Ve1	57.83	-15.74	0.1410	0.5913	1.0628	0.55	16.8	80	7.2	0.41				C ₁
Ve2	59.00	-17.75	0.1589	0.6038	1.0567	0.25	13.5	30	2.1	0.50	3.0	> 7		C ₁
K1	-24.18	-26.84	0.1170	0.6298	1.0464	1.35	13.6	160	13	0.55	3.5	> 30	Kies 1 (Kies π)	C ₁
He1	-50.00	20.21	0.1085	0.6245	0.9753	2.5	14.4	230	21	0.56			Aristarchus 1 (Herodotus α)	B ₁
Lunar cones (for comparison)														
SC1	27.6	18.6	0.0882	0.6646	1.0362	3.54	2.3	72	0.16	0.52			Osiris	cone
SC2	27.5	18.9	0.0872	0.6609	1.0389	4.0	1.7	60	0.071	0.52			Isis	cone
MC1	-31.7	10.7	0.0977	0.6167	1.0426	3.9	1.8	61	0.081	0.52			cone	
Domes of possibly intrusive origin in other lunar regions (for comparison)														
V1	10.2	30.7	0.1195	0.5967	1.0134	0.55	30	130	42	0.46			Valentine	
V2	10.26	31.89	0.1117	0.5921	1.0343	0.82	11	80	1.9	0.34				

the rheologic properties of the magma which formed the dome, i.e. its viscosity and eruption rate as well as the duration of the effusion process. This approach is based on the model by Blake (1990) which assumes that domes were formed by the extrusion of viscous liquid magma onto a flat plane, spreading in all directions from the vent, in contrast to lava flows resulting from lava extrusion onto an inclined surface. Furthermore, Wilson and Head (2003) show that these rheologic parameters in turn yield the magma rise speed as well as the width and length of the feeder dike of the dome, where as a general rule the vertical extension of a dike into the crust approximately corresponds to its length (Jackson et al., 1997). Relying on such geophysical modelling, Wöhler et al. (2007b) find that dome morphology is directly related to magma rise speed and feeder dike dimensions. Accordingly, they

Table 7.4 Northern Tranquillitatis Alignment, domes in Mare Undarum, Mons Rümker, dome at the southern rim of Petavius.

Dome	long. ($^{\circ}$)	lat. ($^{\circ}$)	R_{750}	R_{415}/R_{750}	R_{950}/R_{750}	ζ ($^{\circ}$)	D (km)	h (m)	V (km^3)	f	D_C (km)	d_C (m)	Name	Class
Northern Tranquillitatis Alignment (NTA) domes														
NTA1	35.10	14.66	0.1411	0.5920	1.0389	0.34	17.1	50	3.9 0.34				A-C ₁ -C ₂	
NTA2	35.40	14.30	0.0999	0.5995	1.0520	0.70	5.7	35	0.60 0.67				A-E ₂	
D	35.64	14.24	0.0980	0.6161	1.0362	1.31	6.1	70	1.1 0.54				Diana A-C ₂ -E ₂	
C7, G	35.86	14.18	0.0980	0.6161	1.0362	2.00	8.0	140	4.7 0.67				Grace C ₂	
NTA3	36.19	13.95	0.0937	0.6270	1.0482	0.62	9.2	50	2.0 0.60	2.6	> 7		B ₂ -C ₁	
NTA4	36.45	13.63	0.1052	0.6103	0.9916	0.62	8.3	45	0.83 0.34				B ₂ -C ₁	
NTA5	36.72	13.54	0.0985	0.6046	1.0366	0.50	5.7	25	0.35 0.55				A-E ₂	
NTA6	37.51	13.01	0.0942	0.6124	1.0570	0.49	7.0	30	0.55 0.48				A-E ₂	
Domes in Mare Undarum														
Co1	70.30	7.05	0.1162	0.5808	1.0690	1.8	9.7	150	9.5 0.85				Condorcet 1 B ₂ -C ₁	
Co2	70.30	6.72	0.1200	0.5747	1.0920	1.5	10.3	130	7.4 0.68				Condorcet 2 B ₂	
Co3	70.64	6.78	0.1143	0.5731	1.0826	1.1	11.2	110	5.3 0.49				Condorcet 3 C ₁	
Co4	70.93	6.67	0.1240	0.5751	1.0737	2.8	11.1	270	15.3 0.58				Condorcet 4 B ₁	
Du3	71.30	5.54	0.1376	0.5848	1.0648	0.9	11.7	90	3.0 0.31				Dubiago 3	
Individual domes on the Mons Rümker plateau														
R1	-58.10	41.40	0.1127	0.5667	1.0410	2.73	8.4	200	5.71 0.51				B ₁	
R2	-57.50	41.20	0.1166	0.5607	1.0305	1.64	9.1	130	4.31 0.51				B ₂	
R3	-58.30	40.40	0.1104	0.5643	1.0369	3.02	9.1	240	7.55 0.48				B ₁	
R4	-58.50	40.20	0.1124	0.5751	1.0013	2.67	7.3	170	3.68 0.52				B ₁	
R5	-57.10	40.80	0.1162	0.5955	0.9718	1.46	5.5	70	0.91 0.55				E ₂	
R6	-58.50	40.70	0.1072	0.5622	1.0428	1.66	6.9	100	2.17 0.58				B ₂	
Dome at the southern rim of Petavius														
Pel	60.68	-26.99	0.1830	0.6066	≈ 1.05	1.40	19.8	240	25.9 0.35	3.0	> 70		C ₁	

divide the examined set of lunar mare domes into three rheologic³ groups R₁–R₃ with distinct rheologic properties and feeder dike geometries, which are illustrated in Fig. 7.32.

Domes of rheologic group R₁ are characterised by high effusion rates, moderate to large erupted lava volumes, low to moderate lava viscosities of 10^4 – 10^6 Pa s, moderate magma rise speeds, dike widths around 10 m, and dike lengths of 20–150 km. In the classification scheme for lunar mare domes, this rheologic group is made up by domes of classes B₂, C, and E₂. Domes of rheologic group R₂ display similarly high effusion rates but much lower edifice volumes and thus shorter durations of the effusion process. The lavas of low viscosity between 10^2 and 10^4 Pa s created feeder dikes around 3 m wide and 7–16 km long. The small dikes and high effusion rates imply high magma rise speeds around 10^{-2} m s⁻¹. The rheologic group R₂ is made up by domes that belong to the spectral and morphometric classes A and E₂. Domes of group R₃ are characterised by relatively low lava effusion rates but large erupted lava volumes, implying long durations of the effusion process which typically amount to several years but may also be as long as 18 years. During effusion, the lava displayed high viscosities of $\sim 10^7$ Pa s, ascending at low speeds of $\sim 10^{-5}$ m s⁻¹ through feeder dikes of ~ 100 m width and 130–190 km length. This rheologic group consists of domes belonging to the spectral and morphometric classes B₁ and E₁. This rheologic scheme does not include the large and steep lunar highland domes. They are characterised by still higher magma viscosities, lower effusion rates, and longer durations of the effusion process.

By comparing the time scale of magma ascent with the time scale on which heat is conducted from the magma into the surrounding host rock, Wöhler et al. (2007b) find evidence that for mare domes, the importance of magma evolution processes during ascent such as cooling and crystallisation increases with lava viscosity. Dif-

³ The term rheology refers to the study of the deformation and flow of matter, i.e. rocks and viscous magma in the lunar domes scenario, under the influence of an applied stress.

ferent degrees of evolution of initially fluid basaltic magma are able to explain the broad range of lava viscosities inferred for the examined mare domes, thus being responsible for the observed large variety of three-dimensional dome shapes. On the other hand, for the lunar highland domes no significant cooling of the magma occurred in the dike during ascent despite the high viscosity. This supports the assumption of a specific phase of non-mare volcanism, characterised by the effusion of highly viscous lavas of a composition fundamentally different from the mare lavas, leading to the formation of large and steep volcanic edifices.

It is important to note that these fairly detailed geophysical conclusions about lunar domes and the mechanisms behind their formation could only be drawn based on morphometric properties obtained by image-based three-dimensional surface reconstruction methods.

Chapter 8

Conclusion

This work has described the state of the art in three-dimensional computer vision as well as achievements which have been made due to a variety of newly introduced methods. In this final chapter we discuss the main contributions of this work and possible future research directions.

The first part of this work has discussed three very general classes of three-dimensional computer vision methods. As a first class of methods for three-dimensional scene reconstruction, geometric approaches have been regarded. We have seen that three-dimensional scene reconstruction based on point correspondences between a set of images is a problem for which theoretically well-founded mathematical frameworks are available in the contexts of classical bundle adjustment and projective geometry. Similarly, intrinsic and extrinsic camera calibration and self-calibration are solved problems for which many different solutions exist. As a general rule, camera calibration based on linear methods in the framework of projective geometry should be refined by a bundle adjustment stage. The relative accuracies of linear projective geometry methods and bundle adjustment approaches for three-dimensional scene reconstruction and camera calibration strongly depend on the utilised camera system and on the application at hand, such that general rules are hard to obtain.

A major drawback even of recent camera calibration systems is the fact that the calibration rig has to be identified more or less manually in the images. At this point an important progress has been achieved in this work by introducing a graph-based method for the automatic detection of the calibration rig and its orientation. It has been shown that for wide-angle lenses with strong distortions or non-pinhole optical systems such as fisheye lenses or catadioptric omnidirectional cameras, it is preferable to use chequerboard patterns instead of photogrammetric retroreflective markers as calibration rigs. For such optical systems, nontrivial corrections need to be applied to the measured centres of circular markers while chequerboard corner locations are point features and thus bias-free. A method for the localisation of chequerboard corners at high accuracy based on a physical model of the point spread function of the lens has been introduced which yields an accuracy comparable to that of a circular marker detector under favourable illumination conditions while

showing a significantly more robust behaviour in the presence of low contrast or non-uniform illumination. At the same time, the proposed approach is clearly superior to previous chequerboard corner localisation techniques.

In the context of stereo vision, a variety of blockmatching, feature-based, dense, and spacetime methods has been described. A contour-based approach to the determination of accurate disparity values has been proposed for camera-based safeguarding of workspaces in the industrial production scenario. Spacetime stereo approaches which determine motion information along with the three-dimensional scene structure have recently emerged but are not yet often applied in computer vision systems. In this context, a novel spacetime stereo method based on local intensity modelling has been introduced. Along with the three-dimensional position of a scene point, the model adaptation yields information about the motion along the epipolar line and along the depth axis.

Furthermore, an introduction to methods for three-dimensional pose estimation of rigid, non-rigid, and articulated objects has been provided. A template based monocular approach for pose estimation of rigid objects has been proposed which relies on CAD data of the object and does not require a-priori information about the pose. To determine the three-dimensional structure of non-rigid objects in the scene, which display an infinite number of degrees of freedom, a multiple-view ziplock ribbon snake technique has been developed. For the pose estimation of articulated objects such as human body parts, consisting of several rigid parts related to each other by internal degrees of freedom, the multiocular contracting curve density (MOCCD) algorithm has been proposed as a robust appearance-based approach. Provided that an initial guess of the pose parameters is available, it allows to refine the pose of an object and track it over time at high accuracy in the presence of strongly cluttered scene background and low contrast between the object and the background. To address the problem of three-dimensional pose estimation of articulated objects in the absence of prior knowledge, a further proposed approach performs a model-based segmentation of a motion-attributed point cloud along with an estimation of the object pose and its temporal derivative. This method relies on the disparity values and the low-level motion information extracted by spacetime stereo. Another proposed point cloud segmentation approach relies on the extraction of motion-attributed clusters from spacetime stereo data, which are then used for generating and tracking object hypotheses based on a kernel particle filter framework. In this context, a general insight is that using motion-attributed three-dimensional point cloud data is favourable, if not essential for performing robust segmentation and tracking of objects in three-dimensional point clouds.

The second class of three-dimensional scene reconstruction methods regarded in this work are photometric methods, where we have concentrated on approaches related to shadow analysis, photoclinometry, shape from shading, photometric stereo, and shape from polarisation. It has been shown that photoclinometric approaches that aim for a reconstruction of cross-sectional surface profiles, which were originally developed in the domain of extraterrestrial remote sensing in the middle of the 20th century, are closely related to the shape from shading methods developed about three decades ago in the domain of computer vision. If information about the

reflectance properties of the surface is available, photometric approaches may provide fairly accurate information about the surface gradients, which in turn yield the three-dimensional surface shape. However, a drawback of most shape from shading methods is that they only converge towards a solution if additional constraints such as smoothness or integrability of the surface are applied. Furthermore, the determined solution is generally not unique. Shape from shading methods which yield a unique solution based on an eikonal equation approach are restricted to Lambertian surfaces and require a-priori knowledge about the position of local minima of the surface, which is not necessarily straightforward to obtain.

Some of these drawbacks are alleviated when several pixel-synchronous images acquired under different illumination conditions are available for an evaluation in terms of photometric stereo. The classical approach relies on three pixel-synchronous images of the scene acquired under different illumination conditions and yields a unique configuration of albedo and surface gradients for each pixel, which can be computed by a pixel-wise matrix inversion procedure as long as the surface is Lambertian. The described ratio-based photoclinometry and photometric stereo methods relying on two pixel-synchronous images are suitable for a more general class of reflectance functions. In contrast to the classical photometric stereo approach, they can be used in the presence of coplanar illumination vectors often encountered in remote sensing scenarios.

Furthermore, an extension of the shape from shading framework towards the determination of surface orientation from polarisation information has been described. Accurate and physically well-defined polarisation models based on the Fresnel theory are available for smooth dielectric surfaces. An approximate description can still be obtained according to the refraction law for smooth, specularly reflecting metallic surfaces. Rough metallic surfaces, however, have a more complex polarisation behaviour for which no accurate physical models are available so far. Hence, we have proposed an empirical framework to determine the dependence between the polarisation properties of rough metallic surfaces and the observing and illumination geometry.

The third examined class of three-dimensional reconstruction methods are real-aperture approaches. They exploit the dependence of the point spread function on the distance between the camera and the object. Depth from focus techniques search for the point of maximum sharpness by moving the camera or the object and are thus accurate but slow. Depth from defocus approaches measure the width of the point spread function and infer the object depth based on a previously determined relation between these two values. While classical real-aperture methods assume that this dependence can be well described in terms of geometric optics, it has been shown in this work that an empirical approach based on a suitable calibration of the lens (which may be performed simultaneously with geometric camera calibration) is preferable. The depth–defocus function introduced in this context has been shown to represent the observed relation between object depth and width of the point spread function much better than the relation inferred from geometric optics. A general property of the depth from defocus approach is that it yields dense but fairly inac-

curate and noisy depth maps. It has been demonstrated analytically that depth from defocus should be preferentially utilised in close-range scenarios.

The described classes of three-dimensional reconstruction methods all have their specific advantages and drawbacks. Some of the techniques have complementary properties—geometric methods determine three-dimensional point clouds describing textured parts of the scene while photometric methods may be able to reconstruct textureless regions. Hence, it has been demonstrated to be favourable for computer vision systems to integrate different three-dimensional scene reconstruction methods into a unifying framework. The first described integrated approach combines structure from motion and depth from defocus and yields a three-dimensional point cloud of the scene along with the absolute scaling factor without the need for a-priori knowledge about the scene or the camera motion. Several quantities that influence the accuracy of this approach, such as pixel noise, the nonlinearity of the depth-defocus function, and temperature effects, are discussed. Another integrated approach combines shadow and shading features for three-dimensional surface reconstruction, alleviating the ambiguity of the shape from shading solution. The shape from photopolarimetric reflectance and depth method integrates photopolarimetric information with depth information that can in principle be obtained from arbitrary sources. In this context, depth from defocus information can be favourably used to determine the large-scale properties of the surface, to appropriately initialise the surface gradients, and to estimate the surface albedo. Sparse depth information is incorporated by transforming it into dense depth difference information, such that the three-dimensional reconstruction accuracy is significantly increased especially on large scales. The shape from photopolarimetric reflectance and depth method has been extended towards an iterative scheme for stereo image analysis of non-Lambertian surfaces. This approach overcomes the general drawback of classical stereo approaches, which implicitly assume a Lambertian surface reflectance when establishing point correspondences between images. Disparity estimation is performed based on a comparison between the observation and the surface model, leading to a refined disparity map with a strongly reduced number of outliers. A further integrated approach has been introduced to address the problem of monocular three-dimensional pose refinement of rigid objects based on photopolarimetric, edge, and depth from defocus information. The first proposed technique relies on the nonlinear minimisation of an appropriate error term by gradient descent, while the second described method extends the Bayesian framework of the contracting curve density algorithm towards an integration of additional information such as depth from defocus. As a general result, it has been demonstrated that the combination of various monocular cues allows to determine all six pose parameters of a rigid object at high accuracy.

The second part of this work has addressed several scenarios in which three-dimensional computer vision methods are favourably applied. The first regarded application scenario is quality inspection of industrial parts. For the three-dimensional pose estimation of rigid parts, the proposed combined methods have turned out to yield rotational pose parameters of an accuracy of a few tenths of a degree. It has been shown that these accuracies are significantly higher than those obtained with

previously existing methods mostly relying on edge information. This favourable behaviour can be attributed to the inclusion of photopolarimetric information. The accuracy of the determined lateral object position is largely given by the pixel resolution of the image. The depth accuracy obtained from defocus information comes close to the accuracies reported in the literature for systems performing pose estimation in stereo image pairs.

The three-dimensional pose estimation of non-rigid parts (cables and tubes) has been demonstrated based on the multiocular ziplock ribbon snake method. It has been employed in a robot-based system in which the result of pose estimation is utilised for gripping the non-rigid object. The localisation accuracy of this approach has been shown to be comparable to the pixel resolution of the image.

The integrated frameworks involving the combination of shadow and shading features, shape from photopolarimetric reflectance and depth, and specular stereo have been applied to the three-dimensional reconstruction of rough metallic surfaces. The accuracy of depth differences on the surface is typically better than the lateral pixel resolution of the utilised images, corresponding to an absolute depth accuracy of 30–100 µm at a lateral pixel resolution of 86 µm. For poorly known reflectance parameters, a graceful degradation of the reconstruction accuracy is observed. By comparing these results to typical accuracies of active scanning systems, it has been shown that the proposed integrated frameworks provide accurate, cost-efficient, and fast methods for three-dimensional surface reconstruction in largely controlled environments.

A different application scenario is safe human–robot interaction in the industrial production environment. We have given an introduction to existing systems which are either still under investigation or already commercially available. Furthermore, an overview of hand tracking and gesture recognition methods for human–robot interaction has been provided. This discussion has led to the conclusion that most of the existing methods are not suitable for safety systems. Their performance tends to decrease in the presence of cluttered background, many of them depend on skin colour cues which are unsuitable for detecting humans in the industrial production environment, and a high accuracy of the localisation and pose estimation of the hand is often only obtained if a user-specific model is available. Hence, we have evaluated the proposed methods for segmentation of three-dimensional point clouds and model-based object localisation and pose estimation in the context of human–robot interaction. Merely relying on coarse object models, the motion of a human through an industrial workspace as well as the motion of the hand–forearm limb through a scene with a cluttered background can be reliably tracked. It is not necessary to adapt these systems to the individual human being regarded. Similarly, encouraging results have been obtained for the MOCCD algorithm, which is used for tracking the hand–forearm limb through highly complex cluttered scenes.

The third application scenario refers to the generation of topographic maps in the domain of lunar remote sensing. Methods based on radar and laser altimetry, shadow length measurements, stereo and multi-image photogrammetry, and photoclinometry are well-established with respect to the extraction of regional or global topographic data of various solar system bodies. However, the coverage of the lu-

nar surface by high-quality topographic data is fairly sparse, and height differences smaller than about 100 m can hardly be extracted from the existing maps due to the lack of images showing an amount of detail and variation in viewing geometry which are sufficient for highly accurate photogrammetric evaluation. Hence, we have applied the proposed photometric methods to the three-dimensional reconstruction of lunar craters, low wrinkle ridges and faults, and domes, mainly based on telescopic lunar images acquired with CCD cameras but also utilising recent images by the SMART-1 spacecraft. In this context we have given an introduction to the reflectance behaviour of planetary regolith surfaces. The presented digital elevation maps of lunar craters are of higher resolution than the available topographic maps while the extracted basic features (e.g. crater depth) are consistent with the existing data. Most of the examined lunar wrinkle ridges, faults, and domes have not been previously mapped. Only one catalogue of lunar domes containing height values for a substantial number of these objects exists so far, which was compiled during the mid-1960s based on telescopic photographic lunar images. However, the analysis in this work has shown that the height values stated therein are systematically several times too large. The lunar dome data set presented in this work is thus the first collection of its kind containing reasonably accurate morphometric values for a statistically significant number of objects. A direct comparison has shown that for nearly all examined lunar domes, the accuracy of the height values obtained with the proposed photometric methods based on telescopic CCD imagery is higher than that of dome heights inferred from the Lunar Topographic Orthophotomaps derived by photogrammetric evaluation of orbital Apollo imagery. A novel classification scheme for lunar domes has been developed based on the extracted morphometric data set, and a brief outline of the achieved geophysical insights has been provided.

In the light of the described progress in the field of three-dimensional computer vision methods and their extensive evaluation in several strongly different real-world application scenarios, future research directions emerge.

For example, the proposed frameworks for three-dimensional reconstruction of metallic surfaces should be integrated with active scanning systems e.g. relying on coded structured light in order to exploit the advantages of both approaches. In such surface inspection systems, the scanning component, which requires a comparably high instrumental effort, will only need to operate at low lateral resolution since the data can be densified by the image-based photometric reconstruction techniques.

The encouraging results obtained with the proposed integrated geometric and photometric frameworks for three-dimensional surface reconstruction furthermore suggest that photometric information should be utilised in combination with classical correspondence-based photogrammetric evaluation techniques for the purpose of regional or global topographic mapping of planetary bodies in the context of future mapping activities. In the photogrammetry community, the high potential of photometric approaches for the improvement or refinement of traditionally prepared topographic maps has only recently been addressed by Dorrer et al. (2007). For the Moon, the orbital imagery currently being acquired by the Japanese lunar orbiting spacecraft Kaguya and the Chinese mission Cheng'e is better resolved by at least one order of magnitude than all previously existing orbital data sets. Provided that

images acquired under suitably oblique illumination angles are available, these image data will then allow to construct regional or even global topographic maps of the Moon with a lateral and vertical resolution of a few metres.

In the domain of safe human–robot interaction, the computer vision system should not stop at the measurement level but utilise the reconstruction results for subsequent cognitive stages. The extracted three-dimensional representation of the scene will thus allow the recognition of actions performed by the user (cf. e.g. Hofemann (2007)). Furthermore, it will be possible to utilise action recognition results for a reliable long-term prediction of the human motion, which is otherwise difficult to predict due to its inherent non-uniformity. Relying on the scene segmentation and three-dimensional pose estimation methods described in this work, Hahn et al. (2008a) have undertaken a step towards action recognition and long-term motion prediction in the context of safe human–robot interaction, which should be continued by future research activities.

References

- Abdel-Aziz, Y. I., Karara, H. M., 1971. Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. Proc. of Symp. on Close-Range Photogrammetry, American Society of Photogrammetry, pp. 1–18.
- Abshire, J. B., Sun, X., Afzal, R. S., 2000. Mars Orbiter Laser Altimeter: Receiver model and performance analysis. *Applied Optics* 39, pp. 2440–2460.
- Adams, J. B., McCord, T. B., 1970. Remote sensing of lunar surface mineralogy: implication from visible and near infrared reflectivity of Apollo 11 samples. Proc. Apollo 11 Lunar Sci. Conf., pp. 1937–1945.
- Agrawal, A., Chellappa, R., Raskar, R., 2005. An algebraic approach to surface reconstruction from gradient fields. Proc. Int. Conf. on Computer Vision, Beijing, vol. 1, pp. 174–181.
- Akima, H., 1970. A new method of interpolation and smooth curve fitting based on local procedures. *J. Assoc. for Computing Machinery* 17(4), pp. 589–602.
- Arulampalam, M. S., Maskell, S., Gordon, N., Clapp, T., 2002. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Trans. on Signal Processing* 50(2), pp. 173–188.
- Aschwanden, P. F., 1993. Experimenteller Vergleich von Korrelationskriterien in der Bildanalyse. Hartung-Gorre-Verlag, Konstanz.
- Ashbrook, J., 1961. Dimensions of the Lunar Dome Kies 1. *Journal of the Association of Lunar and Planetary Observers* 15(1–2).
- Aström, K., 1996. Algebraic varieties in multiple view geometry. Proc. Europ. Conf. on Computer Vision, vol. II, pp. 671–682.
- Atkinson, G. A., Hancock, E. R., 2005a. Analysis of Directional Reflectance and Surface Orientation using Fresnel Theory. Proc. Iberoamerican Congress on Pattern Recognition, Beijing, pp. 103–111.
- Atkinson, G. A., Hancock, E. R., 2005b. Multi-view Surface Reconstruction Using Polarization. Proc. Int. Conf. on Computer Vision, Beijing, pp. 309–316.
- Bachler, G., Berger, M., Röhrer, R., Scherer, S., Pinz, A., 1999. A Vision Driven Automatic Assembly Unit. Proc. International Conference on Computer Analysis of Images and Patterns, Ljubljana, Slovenia, pp. 375–382.

- Baerveldt, A.-J., 1992. Cooperation between Man and Robot: Interface and Safety. Proc. IEEE International Workshop on Robot and Human Communication, Tokyo, Japan.
- Baker, H. H., Binford, T. O, 1981. Depth from Edge and Intensity Based Stereo. Proc. Int. Joint Conf. on Artificial Intelligence, Vancouver, Canada, pp. 631–636.
- Baldwin, R. B., 1963. The Measure of the Moon. University of Chicago Press, Chicago, pp. 390–394.
- Barrois, B., Wöhler, C., 2007. 3D Pose Estimation Based on Multiple Monocular Cues. ISPRS Workshop Towards Benchmarking Automated Calibration, Orientation and Surface Reconstruction from Images (BenCOS), held in conjunction with CVPR 2007, Minneapolis, USA.
- Barrois, B., Wöhler, C., 2008. Spatio-temporal 3D Pose Estimation of Objects in Stereo Images. In: Gasteratos, A., Vincze, M., Tsotsos, J. (eds.), Proc. Int. Conf. on Computer Vision Systems, Santorini, Greece. Lecture Notes in Computer Science 5008, pp. 507–516, Springer-Verlag Berlin Heidelberg.
- Barrow, H., 1977. Parametric correspondence and chamfer matching: two new techniques for image matching. Proc. Int. Joint Conf. on Artificial Intelligence, pp. 659–663.
- Bar-Shalom, Y., Li, X. R., 1993. Estimation and Tracking: Principles, Techniques, and Software. Artech House.
- Battl, J., Mouaddib, E., Salvi, J., 1998. Recent progress in coded structured light as a technique to solve the correspondence problem: a survey. Pattern Recognition 31(7), pp. 963–982.
- Bauckhage, C., Hanheide, M., Wrede, S., Käster, T., Pfeiffer, M., Sagerer, G., 2005. Vision Systems with the Human in the Loop. EURASIP Journal on Applied Signal Processing 2005(14), pp. 2375–2390.
- Baumgardner, J., Mendillo, M., Wilson, J. K., 2000. A digital high definition imaging system for spectral studies of extended planetary atmospheres, 1. Initial result in white light showing features on the hemisphere of Mercury unimaged by Mariner 10. Astronomical Journal 119, pp. 2458–2464.
- Beaudet, P. R., 1978. Rotationally invariant image operators. Proc. Int. Conf. on Pattern Recognition, pp. 579–583.
- Beraldin, J.-A., 2004. Integration of Laser Scanning and Close-Range Photogrammetry – The Last Decade and Beyond. Proc. 20th Int. Soc. for Photogrammetry and Remote Sensing Congress, Commission VII, pp. 972–983.
- Besl, P. J., McKay, N. D., 1992. A method for registration of 3-D shapes. IEEE Trans. on Pattern Analysis and Machine Intelligence 14(2), 239–256.
- Beyer, R. A., McEwen, A. S., 2002. Photoclinometric measurements of meter-scale slopes for the potential landing sites of the 2003 Mars Exploration Rovers. Proc. Lunar Planet. Sci. XXXIII, abstract #1443.
- Bhat, D. N., Nayar, S. K., 1998. Stereo and Specular Reflection. Int. J. of Computer Vision 26(2), pp. 91–106.
- Birchfield, S., 1998. An introduction to projective geometry (for computer vision). <http://www.ces.clemson.edu/stb/projective/> (accessed October 16, 2007).

- Black, M. J., Jepson, A. D., 1998. A Probabilistic Framework for Matching Temporal Trajectories: CONDENSATION-Based Recognition of Gestures and Expressions. Proc. Europ. Conf. on Computer Vision, LNCS 1406, Springer-Verlag Berlin Heidelberg, pp. 909–924.
- Blake, S., 1990. Viscoplastic models of lava domes. In: Fink, J. (ed.), Lava Flows and Domes: Emplacement Mechanisms and Hazard Implications, IAVCEI Proc. on Volcanology, vol. 2, pp. 88–128, Springer-Verlag New York.
- Blake, A., Isard, M., 1998. Active Contours. Springer-Verlag, London, 1998.
- Blasko, G., Fua, P., 2001. Real-time 3d object recognition for automatic tracker initialization. Proc. Int. Symp. on Augmented Reality, New York, USA, pp. 175–176.
- Bock, H. H., 1974. Automatische Klassifikation. Vandenhoeck & Ruprecht, Göttingen, Germany.
- Bouguet, J.-Y., 1999. Visual methods for three-dimensional modeling. PhD thesis, California Institute of Technology, Pasadena.
- Bouguet, J.-Y., 2007. Camera Calibration Toolbox for Matlab.
http://www.vision.caltech.edu/bouguetj/calib_doc/ (accessed September 04, 2007).
- Bray, M., Koller-Meier, E., van Gool, L., 2004. Smart particle filtering for 3D hand tracking. Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 675–680.
- Brenner, C., 2000. Dreidimensionale Gebäuderekonstruktion aus digitalen Oberflächenmodellen und Grundrissen. Doctoral dissertation, Institute of Photogrammetry, Stuttgart University. Deutsche Geodätische Kommission, Reihe C, no. 530, München.
- Brenner, C., 2005. Constraints for modelling complex objects. In: Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXXIII-3/W24, pp. 49–54.
- Bronstein, I. N., Semendjajew, K. A., 1989. Taschenbuch der Mathematik. Verlag Harri Deutsch, Frankfurt a. M., Germany.
- Brown, D. C., 1958. A solution to the general problem of multiple station analytical stereotriangulation. Technical report rca-mtp data reduction technical report no. 43 (or afmtc tr 58-8), Patrick Airforce Base, Florida, 1958.
- Brown, D. C., 1966. Decentering distortion of lenses. Photometric Engineering 32(3), pp. 444–462.
- Brown, D. C., 1971. Close-range camera calibration. Photometric Engineering 37(8), pp. 855–866.
- Brox, T., Rosenhahn, B., Cremers, D., 2008. Contours, Optic Flow, and Prior Knowledge: Cues for Capturing 3D Human Motion from Videos. In: Rosenhahn, B., Klette, R., Metaxas, D. (eds.), Human Motion: Understanding, Modelling, Capture and Animation, Springer-Verlag, Dordrecht, The Netherlands.
- Brungart, D. L., 1964. The Origin of Lunar Domes. MSc. thesis, Airforce Institute of Technology, Wright Patterson Air Force Base, Ohio, USA.

- Bruss, A., 1989. The Eikonal Equation: Some Results Applicable to Computer Vision. In: Horn, B. K. P., Brooks, M. (eds.). *Shape from Shading*. MIT Press, Cambridge, USA.
- Burns, R. G., Parkin, K. M., Loeffler, B. M., Leung, I. S., Abu-Eid, R. M., 1976. Further characterization of spectral features attributable to titanium on the moon. Proc. 7th Lunar Science Conference, Houston, Texas, vol. 3 (A77-34651 15-91), pp. 2561–2578, Pergamon Press, New York, USA.
- Bussey, C., Spudis, P., 2004. *The Clementine Atlas of the Moon*. Cambridge University Press, Cambridge, UK, 2004.
- Calow, R., Gademann G., Krell G., Mecke R., Michaelis B., Riefenstahl N., Walke M., 2002. Photogrammetric measurement of patients in radiotherapy. IS-PRS Journal of Photogrammetry and Remote Sensing 56(5–6), pp. 347–359.
- Cañero, C., Radeva, P., Toledo, R., Villanueva, J. J., Mauri, J., 2000. 3D curve reconstruction by biplane snakes. Proc. Int. Conf. on Pattern Recognition, Barcelona, vol. 4, pp. 4563–4567.
- Canny, J., 1986. A computational approach to edge detection. IEEE Trans. on Pattern Analysis and Machine Intelligence 8, pp. 679–698.
- Caselles, V., Kimmel, R., Sapiro, G., 1995. Geodesic active contours. Proc. Int. Conf. on Computer Vision, Boston, USA, pp. 694–699.
- Cattermole, P., 1996. *Planetary Volcanism*. 2nd edition, John Wiley and Sons, Chichester.
- Chandrasekhar, S., 1950. *Radiative Transfer*. Oxford University, London, UK.
- Charette, M. P., McCord, T. B., Pieters, C., Adams, J. B., 1974. Application of remote spectral reflectance measurements to lunar geology classification and determination of titanium content of lunar soils. J. Geophys. Res. 79(11), pp. 1605–1613.
- Chaudhuri, S., Rajagopalan, A. N., 1999. *Depth From Defocus: A Real Aperture Imaging Approach*. Springer-Verlag, New York, 1999.
- Chen, D., Zhang, G., 2005. A new sub-pixel detector for x-corners in camera calibration targets. Proc. 13th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision.
- Chua, C. S., Guan, H. Y., Ho, Y. K., 2000. Model-based finger posture estimation. Proc. Asian Conf. on Computer Vision, pp. 43–48.
- Cintala, M. J., Head, J. W., Mutch, T. A., 1976. Craters on the Moon, Mars, and Mercury: a comparison of depth/diameter characteristics. Proc. Lunar Planet. Sci. VII, pp. 149–151.
- Cipolla, R., Drummond, T., Robertson, D., 1999. Camera calibration from vanishing points in images of architectural scenes. Proc. 10th British Machine Vision Conference, Nottingham, UK, pp. 382–391.
- Clark, R. N., Roush, T. L., 1984. Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. J. Geophys. Res. 89, pp. 6329–6340.
- Clarke, T. A., Fryer, J. F., 1998. The development of camera calibration methods and models. Photogrammetric Record 16(91), pp. 51–66.
- Cohen, L. D., 1991. On active contour models and balloons. Computer Vision, Graphics, and Image Processing. Image Understanding 53(2), pp. 211–218.

- Cohen, L. D., Cohen, I., 1993. Finite element methods for active contour models and balloons for 2d and 3d images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 15(11), pp. 1131–1147.
- Cook, A. C., 2007. Lunar Digital Elevation Models.
<http://users.aber.ac.uk/atc/dems.html> (accessed November 05, 2007).
- Cook, A. C., Spudis, P. D., Robinson, M. S., Watters, T. R., Bussey, D. B. J., 1999. The topography of the lunar poles from digital stereo analysis. *Proc. Lunar Planet. Sci. XXX*, abstract #1154.
- Cordero-Tercero, G., Mendoza-Ortega, B. E., 2001. Study of Ridges' Topographical Profiles on Europa by Means of Photoclinometric Data from E4 Galileo Orbit. *Proc. Spring Meeting of the American Geophysical Union*, abstract #P41B-07.
- Cox, I., Hingorani, S., Rao, S., 1996. A Maximum Likelihood Stereo Algorithm. *Computer Vision and Image Understanding* 63(3), pp. 542–567.
- Craig, J. J., 1989. Introduction to Robotics, Mechanics and Control. Addison-Wesley.
- Cryer, J.E., Tsai, P.-S., Shah, M., 1995. Integration of shape from shading and stereo. *Pattern Recognition*, 28(7), pp. 1033–1043.
- d'Angelo, P., 2007. 3D scene reconstruction by integration of photometric and geometric methods. Doctoral dissertation, Technical Faculty, Bielefeld University, Germany.
<http://bieson.ub.uni-bielefeld.de/volltexte/2007/1145/> (accessed December 04, 2007).
- d'Angelo, P., Wöhler, C., Krüger, L., 2004. Model based multi-view active contours for quality inspection. *Proc. Int. Conf. on Computer Vision and Graphics*, Warsaw, Poland.
- d'Angelo, P., Wöhler, C., 2005a. 3D Reconstruction of Metallic Surfaces by Photopolarimetric Analysis. In: Kalviainen, H., Parkkinen, J., Kaarna, A. (eds.), *Proc. 14th Scand. Conf. on Image Analysis*, Joensuu, Finland. Lecture Notes in Computer Science 3540, pp. 689–698, Springer-Verlag Berlin Heidelberg.
- d'Angelo, P., Wöhler, C., 2005b. 3D surface reconstruction based on combined analysis of reflectance and polarisation properties: a local approach. *ISPRS Workshop Towards Benchmarking Automated Calibration, Orientation and Surface Reconstruction from Images (BenCOS)*, Beijing, China.
- d'Angelo, P., Wöhler, C., 2005c. 3D Surface Reconstruction by Combination of Photopolarimetry and Depth from Defocus. In: Kropatsch, W., Sablatnig, R., Hanbury, A. (eds.). *Pattern Recognition, Proc. 27th DAGM Symposium*, Vienna, Austria. Lecture Notes in Computer Science 3663, pp. 176–183, Springer-Verlag Berlin Heidelberg.
- d'Angelo, P., Wöhler, C., 2006. Image-based 3D surface reconstruction by combination of sparse depth data with shape from shading and polarisation. *ISPRS Conf. on Photogrammetric Computer Vision*, Bonn, Germany.
- d'Angelo, P., Wöhler, C., 2008. Image-based 3D surface reconstruction by combination of photometric, geometric, and real-aperture methods. *ISPRS Journal of Photogrammetry and Remote Sensing* 63(3), pp. 297–321.

- Davis, J., Nehab, D., Ramamoorthi, R., Rusinkiewicz, S., 2005. Spacetime Stereo: A Unifying Framework for Depth from Triangulation. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 27(2), pp. 296–302.
- Davis, J., Shah, M., 1999. Toward 3D gesture recognition. *Int. J. of Pattern Recognition and Artificial Intelligence* 13(3), pp. 381–393.
- Decaudin, P., 1996. Cartoon looking rendering of 3D scenes. INRIA Research Report 2919.
- Delamarre, Q., Faugeras, O., 2001. 3D articulated models and multiview tracking with physical forces. *Computer Vision and Image Understanding* 81(3), pp. 328–357.
- Demant, C., 1999. *Industrial Image Processing*. Springer-Verlag, Berlin.
- Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society, Series B*, 39(1), pp. 1–38.
- Deriche, R., Faugeras, O., 1990. Tracking line segments. *Proc. Europ. Conf. on Computer Vision*, pp. 259–268.
- Deutscher, J., Blake, A., Reid, I., 2001. Articulated body motion capture by annealed particle filtering. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 2126–2133, Kauai, Hawaii, USA.
- Donner, W., 1995. Verfahren und Vorrichtung zur optischen Scharfeinstellung. German patent DE 4413368.5.
- Dorrer, E., Ostrovskiy, A., Kerimov, A., Neukum, G., 2007. SFS with DTM. Symposium of ISPRS Commission IV/7, Goa, India.
- Downs, G. S., Goldstein, R. M., Green, R. R., Morris, G. A., 1971. Mars radar observations: A preliminary report. *Science* 174, pp. 1324–1327.
- Downs, G. S., Reichley, P. E., Green, R. R., 1975. Radar measurements of Mars topography and surface properties: The 1971 and 1973 oppositions. *Icarus* 26, pp. 273–312.
- Dreschler, L., Nagel, H. H., 1982. On the selection of critical points and local curvature extrema of region boundaries for interframe matching. *Proc. Int. Conf. on Pattern Recognition*, pp. 542–544, Munich, Germany.
- Duda, R. O., Hart, P. E., 1973. *Pattern classification and scene analysis*. Wiley-Interscience.
- Duric, Z., Li, F., Sun, Y., Wechsler, H., 2002. Using Normal Flow for Detection and Tracking of Limbs in Color Images. *Proc. Int. Conf. on Pattern Recognition*, vol. 4, pp. 268–271, Quebec City, Canada.
- Durucan, E., 2001. Low computational cost illumination invariant change detection for video surveillance by linear independence. Thèse no. 2454, Ecole Polytechnique Fédérale de Lausanne.
- Ebert, D., 2003. Bildbasierte Erzeugung kollisionsfreier Transferbewegungen für Industrieroboter. Doctoral dissertation, Faculty of Computer Science, Kaiserslautern University, Germany.
- Ebert, D., Henrich, D., 2003. SIMERO: Sichere Mensch-Roboter-Koexistenz. *Proc. Workshop für OTS-Systeme in der Robotik – Mensch und Roboter ohne trennende Schutzsysteme*, Stuttgart, Germany, pp. 119–134.

- Ekman, P., Friesen, W. V., 1969. The repertoire of nonverbal behavior: categories, origins, usage and coding. *Semiotica* 1, pp. 50–98.
- Eliason, E., Isbell, C., Lee, E., Becker, T., Gaddis, L., McEwen, A., Robinson, M., 1999. Mission to the Moon: the Clementine UVVIS global mosaic. PDS Volumes USA_NASA_PDS_CL_4001 4078.
<http://pdsmaps.wr.usgs.gov> (accessed September 04, 2007).
- Ellenrieder, M. M., 2004. Shape reconstruction of flexible objects from monocular images for industrial applications. Proc. SPIE Electronic Imaging, SPIE vol. 5299, Computational Imaging II, pp. 295–303, San Jose, USA.
- Ellenrieder, M. M., 2005. Optimal Viewpoint Selection for Industrial Machine Vision and Inspection of Flexible Objects. Doctoral dissertation, Technical Faculty, Bielefeld University, Germany. VDI Fortschritt-Berichte, Reihe 10, no. 763, VDI-Verlag, Düsseldorf.
- Erol, A., Bebis, G., Nicolescu, M., Boyle, R. D., Twombly, X., 2007. Vision-Based Hand Pose Estimation: A Review. *Computer Vision and Image Understanding* 108(1–2), pp. 52–73.
- European Space Agency, 2007. Kepler Crater as seen by SMART-1.
http://www.esa.int/SPECIALS/SMART-1/SEMBGLVT0PE_2.html (accessed September 04, 2007).
- Fassold, H., Danzl, R., Schindler, K., Bischof, H. 2004. Reconstruction of Archaeological Finds using Shape from Stereo and Shape from Shading. Proc. 9th Computer Vision Winter Workshop, Piran, Slovenia, pp. 21–30.
- Faugeras, O., 1993. Three-Dimensional Computer Vision (Artificial Intelligence). MIT Press, Cambridge, USA.
- Feddemra, J. T., Novak, J. L., 1994. Whole Arm Obstacle Avoidance for Teleoperated Robots. Proc. IEEE Int. Conf. on Robotics and Automation, pp. 3303–3309.
- Fenton, L. K., Herkenhoff, K. E., 2000. Topography and Stratigraphy of the Northern Martian Polar Layered Deposits Using Photoclinometry, Stereogrammetry, and MOLA Altimetry. *Icarus* 147(2), pp. 433–443.
- Fermüller, C., Aloimonos, Y., 1997. On the Geometry of Visual Correspondence. *Int. J. of Computer Vision* 21(3), pp. 223–247.
- Finsterwalder, S., 1899. Die geometrischen Grundlagen der Photogrammetrie. *Jahresbericht der Deutschen Mathematischen Vereinigung*, Teubner, Leipzig.
- Fischler, M. A., Bolles, R. C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24(6), pp. 381–395.
- Fischler, M. A., Bolles, R. C., 1983. Perceptual organization and the curve partitioning problem. Proc. Int. Joint Conf. on Artificial Intelligence, vol. 2, pp. 1014–1018.
- Foley, J. D., van Dam, A., Feiner, S. K., Hughes, J. F., Phillips, R. L., 1993. Introduction to Computer Graphics. Addison-Wesley Professional.
- Fowlkes, C., Belongie, S., Chung, F., Malik, J., 2004. Spectral Grouping Using the Nyström Method. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 26(2), pp. 214–225.

- Franke, U., Gavrila, D., Görzig, S., Lindner, F., Paetzold, F., Wöhler, C., 1999. Autonomous Driving Approaches Downtown. *IEEE Intelligent Systems* 13(6), pp. 40–48.
- Franke, U., Joos, A., 2000. Real-time Stereo Vision for Urban Traffic Scene Understanding. *Proc. IEEE Conf. on Intelligent Vehicles*, Detroit, pp. 273–278.
- Franke, U., Kutzbach, I., 1996. Fast Stereo based Object Detection for Stop&Go Traffic, *IEEE Int. Conf. on Intelligent Vehicles*, Tokyo, pp. 339–344.
- Franke, U., Rabe, C., Badino, H., Gehrig, S. K., 2005. 6D-Vision: Fusion of Stereo and Motion for Robust Environment Perception. In: Kropatsch, W., Sablatnig, R., Hanbury, A. (eds.). *Pattern Recognition*, Proc. 27th DAGM Symposium, Vienna, Austria. Lecture Notes in Computer Science 3663, pp. 216–223, Springer-Verlag Berlin Heidelberg.
- Frankot, R. T., Chellappa, R., 1988. A method for enforcing integrability in shape from shading algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 10(4), pp. 439–451.
- Fritsch, J., Hofemann, N., Sagerer, G., 2004. Combining Sensory and Symbolic Data for Manipulative Gesture Recognition. *Proc. Int. Conf. on Pattern Recognition*, Cambridge, UK, vol. 3, pp. 930–933.
- Fua, P., Brechbühler, C., 1996. Imposing Hard Constraints on Soft Snakes. *Proc. Europ. Conf. on Computer Vision*, Cambridge, UK, vol. 2, pp. 495–506.
- Fua, P., Leclerc, Y. G., 1990. Model driven edge detection. *Machine Vision and Applications* 3(1), pp. 45–56.
- Fusiello, A., Trucco, E., Verri, A., 2000. A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications* 12, pp. 16–22.
- Gaskell1, R. W., Barnoiun-Jha, O. S., Scheeres, D. J., 2007. Modeling Eros with stereophotoclinometry. *Proc. Lunar Planet. Sci. XXXVIII*, abstract #1333.
- Gavrila, D. M., Davis, L. S., 1996. 3D Model-based tracking of humans in action: a multi-view approach. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 73–80.
- Gavrila, D. M., Philomin, V., 1999. Real-time Object Detection for “Smart” Vehicles. *Proc. Int. Conf. on Computer Vision*, pp. 87–93, Kerkyra, Greece.
- Gecks, T., Henrich, D., 2005. Human–robot Cooperation: Safe Pick-and-Place Operations. *Proc. IEEE Int. Workshop on Robot and Human Interactive Communication*, Nashville, USA.
- Gehrke, S., Lehmann, H., Wahlsch, M., Albertz, J., 2006. New Large-Scale Topographic Maps of Planet Mars. *Proc. Europ. Planetary Science Congress*, Berlin, Germany, p. 228.
- Germer, T. A., Rinder, T., Rothe, H., 2000. Polarized light scattering measurements of polished and etched steel surfaces. *Scattering and Surface Roughness III*, SPIE vol. 4100, pp. 148–155.
- Gillis, J. J., Lucey, P. G., 2005. Evidence that UVVIS ratio is not a simple linear function of TiO₂ content for lunar mare basalts. *Proc. Lunar Planet. Sci. XXXVI*, abstract #2252.

- Gövert, T., 2006. Konzeption und Implementierung eines Systems zur raumzeitlichen konturbasierten 3D-Stereoanalyse im Produktionsszenario. Diplom thesis, Technical Faculty, Bielefeld University, Germany.
- Goguen, J. D., 1981. A Theoretical and Experimental Investigation of the Photometric Functions of Particulate Surfaces. PhD thesis, Cornell University, Ithaca, USA.
- Goldstein, R. M., Melbourne, W. G., Morris, G. A., Downs, G. S., O'Handley, D. A., 1970. Preliminary radar results of Mars. *Radio Science* 5, pp. 475–478.
- Gonçalves, N., Araújo, H., 2002. Estimation of 3D Motion from Stereo Images – Differential and Discrete Formulations. Proc. Int. Conf. on Pattern Recognition, vol. 1, pp. 335–338, Quebec City, Canada.
- Gottesfeld Brown, L., 1992. A Survey of Image Registration Techniques. *ACM Computing Surveys* 24(4), pp. 325–376.
- Grammatikopoulos, L., Karras, G., Petsa, E., 2004. Camera calibration combining images with two vanishing points. In: Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXXV-5, pp. 99–104.
- Grammatikopoulos, L., Karras, G., Petsa, E., Kalisperakis, I., 2006. A unified approach for automatic camera calibration from vanishing points. In: Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXXVI-5.
- Grebner, K., 1994. Wissensbasierte Entwicklungsumgebung für Bildanalysesysteme aus dem industriellen Bereich. Doctoral dissertation, Technical Faculty, Bielefeld University, Germany.
- Grest, D., Koch, R., 2008. Motion Capture for Interaction Environments. In: Rosenhahn, B., Klette, R., Metaxas, D. (eds.), *Human Motion: Understanding, Modelling, Capture and Animation*, Springer-Verlag, Dordrecht, The Netherlands.
- Groß, H.-M., Richarz, J., Mueller, S., Scheidig, A., Martin, C., 2006. Probabilistic Multi-modal People Tracker and Monocular Pointing Pose Estimator for Visual Instruction of Mobile Robot Assistants. Proc. IEEE World Congress on Computational Intelligence and Int. Conf. on Neural Networks, pp. 8325–8333.
- Haala, N., Becker, S., Kada, M., 2006. Cell decomposition for the generation of building models at multiple scales. In: Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXXVI-3, pp. 19–24.
- Hafezi, K., Wöhler, C., 2004. A general framework for three-dimensional surface reconstruction by self-consistent fusion of shading and shadow features and its application to industrial quality inspection tasks. Photonics Europe, Strasbourg, SPIE vol. 5457, pp. 138–149.
- Hahn, M., Krüger, L., Wöhler, C., Groß, H.-M., 2007. Tracking of Human Body Parts using the Multiocular Contracting Curve Density Algorithm. Proc. Int. Conf. on 3-D Digital Imaging and Modeling, Montréal, Canada.
- Hahn, M., Krüger, L., Wöhler, C., 2008a. 3D Action Recognition and Long-term Prediction of Human Motion. In: Gasteratos, A., Vincze, M., Tsotsos, J. (eds.), Proc. Int. Conf. on Computer Vision Systems, Santorini, Greece. Lecture Notes in Computer Science 5008, pp. 23–32, Springer-Verlag Berlin Heidelberg.

- Hahn, M., Krüger, L., Wöhler, C., 2008b. Spatio-temporal 3D Pose Estimation of Human Body Parts using the Shape Flow Algorithm. Proc. Int. Conf. on Pattern Recognition, Tampa, USA.
- Hanek, R., 2004. Fitting Parametric Curve Models to Images Using Local Self-adapting Separation Criteria. Doctoral dissertation, Faculty of Computer Science, Technical University of Munich.
- Hapke, B. W., 1981. Bidirectional reflectance spectroscopy 1: Theory. *J. Geophys. Res.* 86, pp. 3039–3054.
- Hapke, B. W., 1984. Bidirectional reflectance spectroscopy 3: Correction for macroscopic roughness. *Icarus* 59, pp. 41–59.
- Hapke, B. W., 1986. Bidirectional reflectance spectroscopy 4: The extinction coefficient and the opposition effect. *Icarus* 67, 264–280.
- Hapke, B. W., 1993. Theory of reflectance and emittance spectroscopy. Cambridge University Press, Cambridge, UK.
- Hapke, B. W., 2002. Bidirectional reflectance spectroscopy 5: The coherent backscatter opposition effect and anisotropic scattering. *Icarus* 157, pp. 523–534.
- Haralick, R., Joo, H., Lee, C., Zhuang, X., Vaidya, V., Kim, M., 1989. Pose estimation from corresponding point data. *IEEE Trans. on Systems, Man, and Cybernetics* 19(6), pp. 1426–1446.
- Harmon, J. K., Campbell, D. B., 1988. Radar observations of Mercury. In: Vilas, F., Chapman, C. R., Shapley Matthews, M. (eds.), *Mercury*, The University of Arizona Press, Tucson, USA.
- Harris, C., Stephens, M., 1988. A combined corner and edge detector. Proc. 4th Alvey Vision Conf., pp. 189–192.
- Hartley, R., 1997. Kruppa's equations derived from the fundamental matrix. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 21, pp. 133–135.
- Hartley, R., Zisserman, A., 2003. *Multiple View Geometry in Computer Vision* (Second Edition). Cambridge University Press, Cambridge, UK.
- Hatzitheodorou, M., 1989. The derivation of 3-d surface shape from shadows. Proc. Image Understanding Workshop, Palo Alto, pp. 1012–1020.
- Head, J. W., Gifford, A., 1980. Lunar mare domes: classification and modes of origin. *The Moon and Planets* 22, pp. 235–257.
- Heap, T., Hogg, D., 1996. Toward 3D hand tracking using a deformable model. Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 140–145.
- Hecht, E., 2001. *Optics* (4th Edition). Addison Wesley.
- Heikkilä, J., Silvén, O., 1997. A four-step camera calibration procedure with implicit image correction. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1106–1112.
- Heisele, B., 1998. Objektdetektion in Straßenverkehrsszenen durch Auswertung von Farbbildfolgen. Doctoral dissertation, Faculty of Electrical Engineering, Stuttgart University. Fortschritt-Berichte VDI, Reihe 10, no. 567.
- Helfenstein, P., 1988. The geological interpretation of photometric surface roughness. *Icarus* 73, pp. 462–481.
- Helfenstein, P., Veverka, J., Hillier, J., 1997. The Lunar Opposition Effect: A Test of Alternative Models. *Icarus* 128, pp. 2–14.

- Hel-Or, Y., Werman, M., 1996. Constraint Fusion for Recognition and Localization of Articulated Objects. *International Journal of Computer Vision* 19(1), pp. 5–28.
- Henyey, L. G., Greenstein, J. L., 1941. Diffuse radiation in the Galaxy. *Astrophysical Journal* 93, pp. 70–83.
- Herkenhoff, K. E., Soderblom, L. A., Kirk, R. L., 2002. MOC photoclinometry of the north polar residual cap on Mars. *Proc. Lunar Planet. Sci. XXXIII*, abstract #1714.
- Hertz, J. A., Krogh, A., Palmer, R. G., 1991. *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, USA.
- Hinz, M., Toennies, K. D., Grohmann, M., Pohle, R., 2001. Active double-contour for segmentation of vessels in digital subtraction angiography. *Proc. of the SPIE*, vol. 4322, pp. 15541562.
- Hirschmüller, H., 2001. Improvements in Real-Time Correlation-Based Stereo Vision. *Proc. IEEE Workshop on Stereo and Multi-Baseline Vision*, Kauai, pp. 141–148.
- Hirschmüller, H., Innocent, P. R., Garibaldi, J., 2002. Real-Time Correlation-Based Stereo Vision with Reduced Border Errors. *Int. J. of Computer Vision* 47(1/2/3), pp. 229–246.
- Hirschmüller, H., 2006. Stereo Vision in Structured Environments by Consistent Semi-Global Matching. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 2386–2393.
- Hirschmüller, H., Mayer, H., Neukum, G., and the HRSC CoI team, 2007. Stereo processing of HRSC Mars Express images by semi-global matching. *Symposium of ISPRS Commission IV/7*, Goa, India.
- Hofemann, N., 2007. Videobasierte Handlungserkennung für die natürliche Mensch-Maschine-Interaktion. Doctoral dissertation, Technical Faculty, Bielefeld University, Germany.
- Horn, B. K. P., 1986. *Robot Vision*. MIT Press, Cambridge, USA.
- Horn, B. K. P., 1989. Height and Gradient from Shading. MIT technical report, AI memo no. 1105A.
- Horn, B. K. P., 2000. Tsai's camera calibration method revisited. MIT technical report.
http://people.csail.mit.edu/bkph/articles/Tsai_Revisited.pdf (accessed September 04, 2007).
- Horn, B. K. P., Brooks, M. (eds.), 1989. *Shape from Shading*. MIT Press, Cambridge, USA.
- Horn, B. K. P., Schunck, B. G., 1981. Determining optical flow. *Artificial Intelligence* 17(1–3), pp. 185–203.
- Horovitz, I., Kiryati, N., 2004. Depth from Gradient Fields and Control Points: Bias Correction in Photometric Stereo. *Image and Vision Computing* 22, pp. 681–694.
- Hothmer, J., 1958. Possibilities and limitations for elimination of distortion in aerial photographs. *Photogrammetric Record* 2(12), pp. 426–445.
- Hu, H., Gao, X., Li, J., Wang, J., Liu, H., Calibrating human hand for teleoperating the hit/dlr hand. *Proc. IEEE Int. Conf. on Robotics and Automation*, vol. 5, pp. 4571–4576.

- Huguet, F., Devernay, F., 2007. A Variational Method for Scene Flow Estimation from Stereo Sequences. Proc. Int. Conf. on Computer Vision, pp. 1–7.
- Isard, M., Blake, A., 1998. CONDENSATION – conditional density propagation for visual tracking. Int. J. of Computer Vision 29, pp. 5–28.
- Jackson, P. A., Wilson, L., Head, J. W., 1997. The use of magnetic signatures in identifying shallow intrusions on the moon. Proc. Lunar Planet. Sci. XXVIII, abstract #1429.
- Jähne, B., 2005. Digitale Bildverarbeitung. Springer-Verlag, Berlin.
- Jaumann, R., and 25 coauthors, 2007. The high-resolution stereo camera (HRSC) experiment on Mars Express: instrument aspects and experiment conduct from interplanetary cruise through the nominal mission. Planetary and Space Science 55(7–8), pp. 928–952.
- Jennings, C., 1999. Robust finger tracking with multiple cameras. Proc. IEEE Int. Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, pp. 152–160.
- Jiang, X., Bunke, H., 1997. Dreidimensionales Computersehen. Springer-Verlag, Berlin.
- Jin, H., Soatto, S., Yezzi, A., 2003. Multi-View Stereo Beyond Lambert. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 171–178.
- Jo, K. H., Kuno, Y., Shirai, Y., 1998. Manipulative hand gesture recognition using task knowledge for human computer interaction. Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 468–473.
- Joshi, M. V., Chaudhuri, S., 2004. Photometric stereo under blurred observations. Proc. Int. Conf. on Pattern Recognition, Cambridge, UK, vol. 3, pp. 169–172.
- Kapral, C., Garfinkle, R., 2005. GLR Lunar Dome Catalog.
<http://digilander.libero.it/glrgroup/kapralcatalog.pdf> (accessed April 26, 2009).
- Kass, M., Witkin, A., Terzopoulos, D., 1988. Snakes: Active Contour Models. International Journal of Computer Vision 1(4), pp. 321–331.
- Kender, J. R., Smith, E. M., 1987. Shape from Darkness: Deriving Surface Information from Dynamic Shadows. Int. Conf. on Computer Vision, London, UK, pp. 539–546.
- Kim, H., Fellner, D. W., 2004. Interaction with hand gesture for a back-projection wall. Proc. Computer Graphics International, pp. 395–402.
- Kimmel, R., Bruckstein, A. M., 1995. Global shape from shading. Computer Vision and Image Understanding 62(3), pp. 360–369.
- Kimmel, R., Sethian, J. A., 2001. Optimal Algorithm for Shape from Shading and Path Planning. Journal of Mathematical Imaging and Vision 14(3), pp. 237–244.
- Kirk, A. G., O'Brien, J. F., Forsyth, D. A., 2005. Skeletal parameter estimation from optical motion capture data. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 782–788.
- Kitchen, L., Rosenfeld, A., 1982. Gray-level corner detection. Pattern Recognition Letters 1, pp. 95–102.
- Klette, R., Koschan, A., Schlüns, K., 1996. Computer Vision: Räumliche Information aus digitalen Bildern. Vieweg Verlag, Braunschweig.

- Klette, R., Kozera, R., Schlüns, K., 1999. Shape from shading and photometric stereo methods. In: Jähne, B., Haussecker, H., Geissler, P. (eds.), *Handbook of Computer Vision and Applications*, vol. 2, *Signal Processing and Pattern Recognition*, Academic Press, San Diego, USA, pp. 532–590.
- Klette, R., Schlüns, K., 1996. Height data from gradient fields. Proc. Photonics East, SPIE vol. 2908, *Machine Vision Applications, Architectures, and Systems Integration V*, Boston, pp. 204–215.
- Knoop, S., Vacek, S., Dillmann, R., 2005. Modeling Joint Constraints for an Articulated 3D Human Body Model with Artificial Correspondences in ICP. Proc. Int. Conf. on Humanoid Robots, Tsukuba, Japan.
- Kölzow, T., 2002. System zur Klassifikation und Lokalisation von 3D-Objekten durch Anpassung vereinheitlichter Merkmale in Bildfolgen. Doctoral dissertation, Technical Faculty, Bielefeld University, Germany.
- Krauß, M., 2006. Integration von Depth-from-Defocus und Shape-from-Polarisation in einen Pose-Estimation-Algorithmus. Diplom thesis, Faculty of Computer Science and Automation, Technical University of Ilmenau.
- Krotkov, E., 1987. Focusing. *Int. J. of Computer Vision* 1, pp. 223–237.
- Krüger, L., 2007. Model Based Object Classification and Localisation in Multiocular Images. Doctoral dissertation, Technical Faculty, Bielefeld University, Germany.
- Krüger, L., Ellenrieder, M. M., 2005. Pose estimation using the multiocular contracting curve density algorithm. Proc. 10th Int. Fall Workshop on Vision, Modeling, and Visualization, Erlangen, Germany.
- Krüger, L., Wöhler, C., Würz-Wessel, A., Stein, F., 2004. In-factory calibration of multiocular camera systems. Proc. SPIE Photonics Europe (Optical Metrology in Production Engineering), Strasbourg, pp. 126–137.
- Krüger, L., Wöhler, C., 2009. Accurate chequerboard corner localisation for camera calibration and scene reconstruction. Submitted to *Pattern Recognition Letters*.
- Kruppa, E., 1913. Zur Ermittlung eines Objektes aus zwei Perspektiven mit innerer Orientierung. *Sitzungsberichte der Mathematisch Naturwissenschaftlichen Kaiserlichen Akademie der Wissenschaften* 122, pp. 1939–1948.
- Kuch, J. J., Huang, T. S., 1994. Human computer interaction via the human hand: A hand model. Proc. Asilomar Conference on Signal, Systems, and Computers, pp. 1252–1256.
- Kuhl, A., 2005. Spatial Scene Reconstruction by Combined Depth-from-Defocus and Shape-from-Motion. Diplom thesis, Faculty of Computer Science and Automation, Technical University of Ilmenau.
- Kuhl, A., Wöhler, C., Krüger, L., Groß, H.-M., 2006. Monocular 3D Scene Reconstruction at Absolute Scales by Combination of Geometric and Real-Aperture Methods. In: Franke, K., Müller, K.-R., Nickolay, B., Schäfer, R. (eds.), *Pattern Recognition*, Proc. 28th DAGM Symposium, Heidelberg, Germany. Lecture Notes in Computer Science 4174, pp. 607–616, Springer-Verlag Berlin Heidelberg.

- Kuhn, S., Gecks, T., Henrich, D., 2006. Velocity control for safe robot guidance based on fused vision and force/torque data. Proc. IEEE Conf. on Multisensor Fusion and Integration for Intelligent Systems, Heidelberg, Germany.
- Kuiper, G. P., 1961. Orthographic atlas of the moon. University of Arizona Press, Tucson, USA.
- Kwon, Y.-H., 1998. DLT Method.
<http://www.kwon3d.com/theory/dlt/dlt.html> (accessed October 16, 2007).
- Lacey, A. J., Thacker, N. A., Courtney, P., Pollard, S. B., 2002. TINA 2001: The Closed Loop 3D Model Matcher. Imaging Science and Biomedical Engineering Division, Medical School, University of Manchester.
- Lamdan, Y., Wolfson, H., 1988. Geometric Hashing: A General and Efficient Model-Based Recognition Scheme. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 238–249.
- Lange, C., Hermann, T., Ritter, H., 2004. Holistic Body Tracking for Gestural Interfaces. In: Gesture-Based Communication in Human-Computer Interaction, selected and revised papers of the 5th Int. Gesture Workshop, Genova, Italy. Lecture Notes in Computer Science 2915, pp. 132–139, Springer Verlag Berlin Heidelberg.
- Lee, J., Kunii, T., 1993. Constraint-based hand animation. In: Models and Techniques in Computer Animation, Springer Verlag, Tokyo, pp. 110–127.
- Lena, R., Wöhler, C., Bregante, M. T., Fattinnanzi, C., 2006. A combined morphometric and spectrophotometric study of the complex lunar volcanic region in the south of Petavius. *J. Royal Astronomical Society of Canada* 100(1), pp. 14–25.
- Lena, R., Wöhler, C., Bregante, M. T., Lazzarotti, P., Lammel, S., 2008. Lunar domes in Mare Undarum: Spectral and morphometric properties, eruption conditions, and mode of emplacement. *Planetary and Space Science* 56, pp. 553–569.
- Li, M., Lavest, J.-M., 1995. Some aspects of zoom-lens camera calibration. Technical Report ISRN KTH/NA/P-95/03-SE, Royal Institute of Technology (KTH), Stockholm, Sweden.
- Lim, H. S., Binford, T. O., 1987. Stereo correspondence: A hierarchical approach. Proc. of DARPA Image Understanding Workshop, pp. 234–241.
- Lim, J., Jeffrey, H., Yang, M., Kriegman, D., 2005. Passive Photometric Stereo from Motion. Proc. IEEE Int. Conf. Computer Vision, vol. II, pp. 1635–1642.
- Lin, J. Y., Wu, Y., Huang, T. S., 2000. Modeling the constraints of human hand motion. Proc. IEEE Human Motion Workshop, pp. 121–126.
- Lin, J. Y., Wu, Y., Huang, T. S., 2004. 3D Model-based hand tracking using stochastic direct search method. Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition, Seoul, Korea, pp. 693–698.
- Lohse, V., Heipke, C., 2003. Derivation of digital terrain models by means of Multi-Image Shape-from-Shading: Results using Clementine images. ISPRS Workshop High Resolution Mapping from Space, Hannover, Germany.
- Lohse, V., Heipke, C., 2004. Multi-image shape-from-shading. Derivation of Planetary Digital Terrain Models Using Clementine Images. Int. Archives of Photogrammetry and Remote Sensing XXXV/B4, pp. 828–833.

- Lohse, V., Heipke, C., Kirk, R. L., 2006. Derivation of planetary topography using multi-image shape-from-shading. *Planetary and Space Science* 54, pp. 661–674.
- Longuet-Higgins, H. C., 1981. A computer algorithm for reconstructing a scene from two projections. *Nature* 293, pp. 133–135.
- Lourakis, M., Argyros, A., 2004. The design and implementation of a generic sparse bundle adjustment software package based on the Levenberg–Marquardt algorithm. Technical Report 340, Institute of Computer Science – FORTH, Heraklion, Crete, Greece.
- Lowe, D. G., 1987. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence* 31(3), pp. 355–395.
- Lowe, D. G., 1991. Fitting parameterized three-dimensional models to images. *IEEE Trans. Pattern Analysis and Machine Intelligence* 13(5), pp. 441–450.
- Lowitzsch, S., Kaminski, J., Knauer, M. C., Häusler, G., 2005. Vision and Modeling of Specular Surfaces. Proc. 10th Int. Fall Workshop on Vision, Modeling, and Visualization, Erlangen, Germany.
- Lu, Y., Zhang, J. Z., Wu, Q. M. J., Li, Z. N., 2004. A Survey of Motion-Parallax-Based 3-D Reconstruction Algorithms. *IEEE Trans. on Systems, Man, and Cybernetics – Part C: Applications and Reviews* 34(4), pp. 532–548.
- Lucas, B. D., Kanade, T., 1981. An iterative image registration technique with an application to stereo vision. Proc. Int. Joint Conf. on Artificial Intelligence, Vancouver, pp. 674–679.
- Lucey, P. G., Blewett, D. T., Hawke, B. R., 1998. Mapping the FeO and TiO₂ content of the lunar surface with multispectral imagery. *J. Geophys. Res.* 103(E2), pp. 3679–3699.
- Lucchese, L., Mitra, S., 2002. Using saddle points for subpixel feature detection in camera calibration targets. Proc. Asia-Pacific Conference on Circuits and Systems, pp. 191–195.
- Luhmann, T., 2003. Nahbereichsphotogrammetrie. Grundlagen, Methoden und Anwendungen. Wichmann, Heidelberg.
- Lumelsky, V., Cheung, E., 1993. Real-Time Collision Avoidance in Teleoperated Whole-Sensitive Robot Arm Manipulators. *IEEE Trans. on Systems, Man and Cybernetics* 23(1), pp. 194–203.
- Lumme, K., Bowell, E., 1981. Radiative transfer in the surfaces of atmosphereless bodies. I. Theory. *Astronomical Journal* 86, pp. 1694–1704.
- MacQueen, J. B., 1967. Some Methods for classification and Analysis of Multivariate Observations. Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297, University of California Press, Berkeley, USA.
- Magda, S., Zickler, T., Kriegman, D., Belhumeur, P., 2001. Beyond Lambert: Reconstructing Surfaces with Arbitrary BRDFs. Proc. Int. Conf. on Computer Vision, pp. 291–302.
- Malik, S., Laszlo, J., 2004. Visual touchpad: A two-handed gestural input device. Proc. Int. Conf. on Multimodal Interfaces, pp. 289–296.
- Mallon, J., Whelan, P. F., 2006. Which pattern? Biassing aspects of planar calibration patterns and detection methods. *Pattern Recognition Letters* 28(8), pp. 921–930.

- Mandler, E., Oberländer, M., 1990. One Pass Encoding of Connected Components in Multi-Valued Images. Proc. IEEE Int. Conf. on Pattern Recognition, Atlantic City, pp. 64–69.
- Marr, D., Poggio, T., 1979. A Computational Theory of Human Stereo Vision. Proc. Royal Society of London. Series B, Biological Sciences, vol. 204, no. 1156, pp. 301–328.
- Mason, S., 1994. Expert system based design of photogrammetric networks. PhD thesis, ETH Zürich.
- McCord, T. B., Adams, J. B., 1973. Progress in optical analysis of lunar surface composition. *The Moon* 7, pp. 453–474.
- McCord, T. B., Charette, M. P., Johnson, T. V., Lebofsky, L. A., Pieters, C., Adams, J. B., 1972. Lunar spectral types. *J. Geophys. Res.* 77, pp. 1349–1359.
- McCord, T. B., Pieters, C., Feierberg, M. A., 1976. Multispectral mapping of the lunar surface using ground-based telescopes. *Icarus* 29, pp. 1–34.
- McEwen, A. S., 1985. Albedo and Topography of Ius Chasma, Mars. *Proc. Lunar Planet. Sci.* XVI, pp. 528–529.
- McEwen, A. S., 1991. Photometric Functions for Photoclinometry and Other Applications. *Icarus* 92, pp. 298–311.
- McEwen, A. S., 1996. A precise lunar photometric function. *Proc. Lunar Planet. Sci.* XXVII, pp. 841–842.
- McKay, D. S., Heiken, G., Basu, A., Blanford, G., Simon, S., Reedy, R., French, B. M., Papike, J., 1991. The Lunar regolith. In: Heiken, G., Vaniman, D., French, B. M. (eds.), *Lunar Sourcebook*, Cambridge University Press, Cambridge, UK.
- McGuire, A. F., Hapke, B. W., 1995. An Experimental Study of Light Scattering by Large, Irregular Particles. *Icarus* 113, pp. 134–155.
- Meister, G., 2000. Bidirectional Reflectance of Urban Surfaces. Doctoral dissertation, Hamburg University.
- Melendrez, D. E., Johnson, J. R., Larson, S. M., Singer, R. B., 1994. Remote sending of potential lunar resources. 2. High spatial resolution mapping of spectral reflectance ratios and implications for nearside mare TiO₂ content. *J. Geophys. Res.* 99(E3), pp. 5601–5619.
- Merrit, E. L., 1948. Field camera calibration. *Photogrammetric Engineering* 14(2), pp. 303–309.
- Miyazaki, D., Tan, R. T., Hara, K., Ikeuchi, K., 2003. Polarization-based Inverse Rendering from a Single View. Proc. IEEE Int. Conf. on Computer Vision, Nice, vol. 2, pp. 982–987.
- Miyazaki, D., Kagesawa, M., Ikeuchi, K., 2004. Transparent Surface Modeling from a Pair of Polarization Images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 26(1), pp. 73–82.
- Miyazaki, D., Ikeuchi, K., 2005. Inverse Polarization Raytracing: Estimating Surface Shape of Transparent Objects. Proc. Int. Conf. on Computer Vision and Pattern Recognition, San Diego, vol. 2, pp. 910–917.
- Moeslund, T. B., Hilton, A., Krüger, V., 2006. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding* 104(2), pp. 90–126.

- Mouginis-Mark, P. J., Wilson, L., 1979. Photoclinometric measurements of Mercurian landforms. *Proc. Lunar Planet. Sci.* X, pp. 873–875.
- Morel, O., Meriaudeau, F., Stoltz, C., Gorria, P., 2005. Polarization imaging applied to 3D reconstruction of specular metallic surfaces. *Machine Vision Applications in Industrial Inspection XIII. SPIE* vol. 5679, pp. 178–186.
- Mühlmann, K., 2002. Design und Implementierung eines Systems zur schnellen Rekonstruktion dreidimensionaler Modelle aus Stereobildern. Doctoral dissertation, Faculty of Science, Mannheim University.
- Mündermann, L., Corazza, S., Andriacchi, T. P., 2008. Markerless Motion Capture for Biomechanical Applications. In: Rosenhahn, B., Klette, R., Metaxas, D. (eds.), *Human Motion: Understanding, Modelling, Capture and Animation*, Springer-Verlag, Dordrecht, The Netherlands.
- Mursky, G., 1996. *Introduction to Planetary Volcanism*. Prentice Hall, Upper Saddle River.
- Myles, Z., da Vitoria Lobo, N., 1998. Recovering affine motion and defocus blur simultaneously. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 20(6), pp. 652–658.
- Nayar, S., 1989. Shape from Focus. Carnegie Mellon University, technical report CMU-RI-TR-89-27.
- Nayar, S. K., Ikeuchi, K., Kanade, T., 1991. Surface reflection: Physical and Geometrical Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13(7), pp. 611–634.
- Nayar, S. K., Fang, X.-S., Boult, T., 1993. Removal of specularities using color and polarization. *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, New York, pp. 583–590.
- Nayar, S. K., Bolle, R. M., 1996. Reflectance based object recognition. *Int. J. of Computer Vision* 17(3), pp. 219–240.
- Nelder, J. A., Mead, R., 1965. A simplex method for function minimization. *Computer Journal* 7, pp. 308–313.
- Nehaniv, C. L., 2005: Classifying Types of Gesture and Inferring Intent. *Proc. Symp. on Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction*, pp. 74–81. The Society for the Study of Artificial Intelligence and the Simulation of Behaviour.
- Neuenschwander, W., Fua, P., Iverson, L., Szekely, G., Kubler, O., 1997. Ziplock Snakes. *Int. J. of Computer Vision* 25(3), pp. 191–201.
- Neumann, G. A., 2001. Some aspects of processing extraterrestrial LIDAR data: Clementine, NEAR, MOLA. *Int. Archives of Photogrammetry and Remote Sensing* 34:3/W4, pp. 73–80.
- Nevatia, R., Babu, K. R., 1980. Linear Feature Extraction and Description. *Computer Graphics and Image Processing* 13, pp. 257–269.
- Nickel, K., Seemann, E., Stiefelhagen, R., 2003. 3D-Tracking of Head and Hands for Pointing Gesture Recognition in a Human-Robot Interaction Scenario. *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Seoul, Korea, pp. 565–570.

- Nickel, K., Stiefelhagen, R., 2004. Real-Time Person Tracking and Pointing Gesture Recognition for Human-Robot Interaction. Proc. Europ. Conf. on Computer Vision, Workshop on HCI, Prague, Czech Republic. Lecture Notes in Computer Science 3058, pp. 28-38, Springer-Verlag Berlin Heidelberg.
- Niemann, H., Hornegger, J., 2001. A novel probabilistic model for object recognition and pose estimation. Int. J. of Pattern Recognition and Artificial Intelligence 15(2), pp. 241–253.
- Nister, D., 2004. An efficient solution to the five-point relative pose problem. IEEE Trans. on Pattern Analysis and Machine Intelligence 26(6), pp. 756–777.
- Nölker, C., Ritter, H., 1999. Grefit: Visual recognition of hand postures. In: Braffort, A., Gherbi, R., Gibet, S., Richardson, J., Teil, D. (eds.), Gesture-Based Communication in Human-Computer Interaction, Proc. Int. Gesture Workshop, Lecture Notes in Artificial Intelligence 1739, pp. 61–72, Springer Verlag Berlin Heidelberg.
- Nomura, Y., Zhang, D., Sakaida, Y., Fujii, S., 1996. 3-d object pose estimation based on iterative image matching: Shading and edge data fusion. Proc. Int. Conf. on Computer Vision and Pattern Recognition, pp. 866–871.
- Nourbakhsh, I. R., Andre, D., Tomasi, C., Genesereth, M. R., 1997. Mobile robot obstacle avoidance via depth from focus. Robotics and Autonomous Systems 22, pp. 151–158.
- Novak, J. L., Feddema, J. T., 1992. A Capacitance-Based Proximity Sensor for Whole Arm Obstacle Avoidance. Proc. IEEE Int. Conf. on Robotics and Automation, pp. 1307–1314.
- O'Brien, J. F., Bodenheimer, R. E., Brostow, G. J., Hodgins, J. K., 2000. Automatic joint parameter estimation from magnetic motion capture data. Proc. Graphics Interface Conf., pp. 53-60.
- Oka, K., Sato, Y., Koike, H., 2002. Real-time tracking of multiple fingertips and gesture recognition for augmented desk interface systems. Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 429–434.
- Olague, G., Hernández, B., 2005. A new accurate and flexible model based multi-corner detector for measurement and recognition. Pattern Recognition Letters 26(1), pp. 27–41.
- Pavlovic, V., Sharma, R., Huang, T. S., 1997. Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review. IEEE Trans. on Pattern Analysis and Machine Intelligence 19(7), pp. 677695.
- Pentland, A., 1987. A new sense for depth of field. IEEE Trans. Pattern Analysis and Machine Intelligence 9, pp. 523–531.
- Pettengill, G. H., Eliason, E., Ford, P. G., Loriat, G. B., 1980. Pioneer Venus Radar Results: Altimetry and Surface Properties. J. Geophysical Research 85, pp. 8261–8270.
- Pike, R. J., Clow, G., 1981. Revised Classification of Terrestrial Volcanoes and Catalogue of Topographic Dimensions, With new Results of Edifice Volume. US Geological Survey Open-File Report 81-1038.
- Phong, B. T., 1975. Illumination for computer generated pictures. Commun. ACM 18(6), pp. 311–317, ACM Press, New York.

- Phong, T. Q., Horaud, R., Yassine, A., Tao, P. D., 1996. Object pose from 2-D to 3-D point and line correspondences. *Int. J. of Computer Vision* 15(3), pp. 225–243.
- Pike, R. J., 1978. Volcanoes on the inner planets: Some preliminary comparisons of gross topography. *Proc. Lunar Planet. Sci. IX*, pp. 3239–3273.
- Pike, R. J., 1980. Control of crater morphology by gravity and target type: Mars, Earth, Moon. *Proc. Lunar Planet. Sci. XI*, pp. 2159–2189.
- Pike, R. J., 1988. Geomorphology of impact craters on Mercury. In: Vilas, F., Chapman, C. R., Shapley Matthews, M. (eds.), *Mercury*, The University of Arizona Press, Tucson, USA.
- Plänkers, R., Fuà, P., 2003. Articulated soft objects for multiview shape and motion capture. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 25(9), pp. 1182–1187.
- Pollefeyns, M., van Gool, L., 1999. Stratified self-calibration with the modulus constraint. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 21, pp. 707–724.
- Pons, J.-P., Keriven, R., Faugeras, O., 2005. Modelling dynamic scenes by registering multi-view image sequences. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 822–827.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., 1992. *Numerical Recipes in C*. Cambridge University Press, Cambridge, UK.
- Rahmann, S., 1999. Inferring 3D scene structure from a single polarization image. *Polarization and Color Techniques in Industrial Inspection*, SPIE vol. 3826, pp. 22–33.
- Rahmann, S., Canterakis, N., 2001. Reconstruction of Specular Surfaces using Polarization Imaging. *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, Kauai, vol. I, pp. 149–155.
- Rey, W. J. J., 1983. *Introduction to Robust and Quasi-Robust Statistical Methods*. Springer-Verlag, Berlin, Heidelberg.
- Rindfleisch, T., 1966. Photometric Method for Lunar Topography. *Photogrammetric Engineering* 32(2), pp. 262–277.
- Rogers, D. F., 2001. *An Introduction to NURBS*. Academic Press, San Diego, USA.
- Rosenhahn, B., Perwass, C., Sommer, G., 2003. Pose estimation of free-form surface models. In: Michaelis, B., Krell, G. (eds.), *Pattern Recognition*, Proc. 25th DAGM Symposium, Magdeburg, Germany. Lecture Notes in Computer Science 2781, pp. 574–581, Springer-Verlag Berlin Heidelberg.
- Rosenhahn, B., Kersting, U., Smith, A., Gurney, J., Brox, T., Klette, R., 2005. A system for marker-less human motion estimation. In: Kropatsch, W., Sablatnig, R., Hanbury, A., (eds.), *Pattern Recognition*, Proc. 27th DAGM Symposium, Vienna, Austria. Lecture Notes in Computer Science 3663, pp. 230–237, Springer-Verlag Berlin Heidelberg.
- Rosenhahn, B., Brox, T., Cremers, D., Seidel, H.-P., 2006. A comparison of shape matching methods for contour based pose estimation. In: *Combinatorial Image Analysis*, Lecture Notes in Computer Science 4040, pp. 263–276, Springer-Verlag Berlin Heidelberg.

- Rosenhahn, B., Kersting, U. G., Powell, K., Brox, T., Seidel, H.-P., 2008. Tracking Clothed People. In: Rosenhahn, B., Klette, R., Metaxas, D. (eds.), *Human Motion: Understanding, Modelling, Capture and Animation*, Springer-Verlag, Dordrecht, The Netherlands.
- Rother, C., 2000. A new approach for vanishing point detection in architectural environments. Proc. 11th British Machine Vision Conference, Bristol, UK, pp. 382–391.
- Rottensteiner, F., Trinder, J., Clode, S., Kubik, K., 2005. Automated delineation of roof planes in LIDAR data. In: Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXXVI-3/W19, pp. 221–226.
- Rottensteiner, F., 2006. Consistent estimation of building parameters considering geometric regularities by soft constraints. In: Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXXVI-3, pp. 13–18.
- Roy, S., Cox, L., 1998. A Maximum-Flow Formulation of the N -camera Stereo Correspondence Problem. Proc. Int. Conf. on Computer Vision, Bombay, pp. 492–499.
- Rubin, A. S., 1993. Tensile fracture of rock at high confining pressure: Implications for dike propagation. *J. Geophys. Res.* 98, pp. 15919–15935.
- Rucklidge, W., 1996. Efficient Visual Recognition Using the Hausdorff Distance. Lecture Notes in Computer Science, vol. 1173, Springer-Verlag Berlin Heidelberg.
- Rükl, A., 1999. *Mondatlas*. Verlag Werner Dausien, Hanau, Germany.
- Rouy, E., Tourin, A., 1992. A viscosity solutions approach to shape-from-shading. *SIAM Journal of Numerical Analysis* 29(3), pp. 867–884.
- Sagerer, G., 1985. Darstellung und Nutzung von Expertenwissen für ein Bildanalyssystem. Springer-Verlag Berlin Heidelberg.
- Samaras, D., Metaxas, D., Fua, P., Leclerc, Y.G., 2000. Variable Albedo Surface Reconstruction from Stereo and Shape from Shading. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol. I, pp. 480–487.
- Savarese, S., Rushmeier, H., Bernardini, F., Perona, P., 2002. Implementation of a Shadow Carving System for Shape Capture. Proc. Int. Symp. on 3D Data Processing, Visualization and Transmission, Padua, pp. 107–114.
- Schaper, D., 2002. Automated quality control for micro-technology components using a depth from focus approach. Proc. 5th IEEE Southwest Symp. on Image Analysis and Interpretation, pp. 50–54.
- Scharstein, D., Szeliski, R., 2001. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *Int. J. Computer Vision* 47(1/2/3), pp. 7–42.
- Schenk, P. M., Pappalardo, R. T., 2002. Stereo and Photoclinometric Topography of Chaos and Anarchy on Europa: Evidence for Diapiric Origin. Proc. Lunar Planet. Sci. XXXIII, abstract #2035.
- Schlüns, K., 1997. Shading Based 3D Shape Recovery in the Presence of Shadows. Proc. First Joint Australia & New Zealand Biennial Conference on Digital Image & Vision Computing: Techniques and Applications, Auckland, New Zealand, pp. 195–200.

- Schmidt, J., Fritsch, J., Kwolek, B., 2006. Kernel particle filter for real-time 3d body tracking in monocular color images. Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 567–572.
- Schmidt, J., Wöhler, C., Krüger, L., Gövert, T., Hermes, C., 2007. 3D Scene Segmentation and Object Tracking in Multiocular Image Sequences. Proc. Int. Conf. on Computer Vision Systems, Bielefeld, Germany.
- Schmid, H. H., 1974. Stellar calibration of the orbigon lens. Photogrammetric Engineering 40(1), pp. 101–111.
- Schulz, O., 2003. Image-based 3D-surveillance in Human-Robot-Cooperation. Proc. Int. CAMT Conference, Modern Trends in Manufacturing, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, Poland, pp. 327–334.
- Schunck, B. G., 1989. Image Flow Segmentation and Estimation by Constraint Line Clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence 11(10), pp. 1010–1027.
- Sethian, J., 1999. Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science. Cambridge University Press, Cambridge, UK.
- Shi, J., Malik, J., 1997. Normalized Cuts and Image Segmentation. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 731–737, San Juan, Puerto Rico.
- Shi, J., Malik, J., 1998. Motion Segmentation and Tracking using Normalized Cuts. Proc. Int. Conf. on Computer Vision, Bombay, India, pp. 1154–1160.
- Shi, J., Malik, J., 2000. Normalized Cuts and Image Segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence 22(8), pp. 888–905.
- Shi, J., Tomasi, C., 1994. Good features to track. IEEE Conf. on Computer Vision and Pattern Recognition, Seattle, USA, pp. 593–600.
- Shimada, N., Shirai, Y., Kuno, Y., Miura, J., 1998. Hand gesture estimation and model refinement using monocular camera-ambiguity limitation by inequality constraints. Proc. IEEE Int. Conf. on Face and Gesture Recognition, pp. 268–273.
- Shimada, N., Kimura, K., Shirai, Y., 2001. Real-time 3D hand posture estimation based on 2D appearance retrieval using monocular camera. Proc. ICCV workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, Vancouver, Canada, pp. 23–30.
- Simchony, T., Chellappa, R., Shao, M., 1990. Direct Analytical Methods for Solving Poisson Equations in Computer Vision Problems. IEEE Trans. Pattern Analysis and Machine Intelligence 12(5), pp. 435–446.
- Sminchisescu, C., 2008. 3D Human Motion Analysis in Monocular Video: Techniques and Challenges. In: Rosenhahn, B., Klette, R., Metaxas, D. (eds.), Human Motion: Understanding, Modelling, Capture and Animation, Springer-Verlag, Dordrecht, The Netherlands.
- Smith, E. I., 1974. Rümker Hills: A Lunar Volcanic Dome Complex. The Moon 10(2), pp. 175–181.
- Socher, G., 1997. Qualitative Scene Descriptions from Images for Integrated Speech and Image Understanding. Doctoral dissertation, Technical Faculty, Bielefeld University, Germany.

- Som, F., 2005. Sichere Steuerungstechnik fr den OTS-Einsatz von Robotern. Proc. Workshop für OTS-Systeme in der Robotik (Sichere Mensch-Roboter-Interaktion ohne trennende Schutzsysteme), Stuttgart, Germany.
- Spudis, P. D., 1993. The Geology of Multi-Ring Impact Basins. Cambridge University Press, Cambridge, UK.
- Spurr, J. E., 1945. Geology Applied to Selenology, vol. I. Science Press, Lancaster, Pa., USA.
- Staid, M. I., Pieters, C. M., Head, J. W., 1996. Mare Tranquillitatis: Basalt emplacement history and relation to lunar samples. *J. Geophys. Res.* 101(E10), pp. 213–227.
- Stein, F., 2004. Efficient Computation of Optical Flow Using the Census Transform. In: Rasmussen, C. E., Bühlhoff, H. H., Giese, M. A., Schölkopf, B. (eds.), Pattern Recognition, Proc. 26th DAGM Symposium, Tübingen, Germany. Lecture Notes in Computer Science 3175, pp. 79–86, Springer-Verlag Berlin Heidelberg.
- Stenger, B., Mendonça, P. R. S., Cipolla, R., 2001. Model-based hand tracking using an unscented kalman filter. Proc. British Machine Vision Conference, vol. I, pp. 63–72, Manchester, UK.
- Stenger, B., Thayananthan, A., Torr, P. H. S., Cipolla, R., 2003. Filtering using a tree-based estimator. Proc. Int. Conf. on Computer Vision, pp. 1063–1070.
- Stenger, B., Thayananthan, A., Torr, P., Cipolla, R., 2004. Hand pose estimation using hierarchical detection. Proc. Int. Workshop on Human-Computer Interaction, Lecture Notes in Computer Science 3058, Springer-Verlag Berlin Heidelberg, pp. 102–112.
- Stößel, D., 2007. Automated Visual Inspection of Assemblies from Monocular Images. Doctoral dissertation, Technical Faculty, Bielefeld University, Germany.
- Subbarao, M., 1988. Parallel depth recovery by changing camera parameters. Proc. Int. Conf. on Computer Vision, pp. 149–155.
- Subbarao, M., Surya, G., 1994. Depth from Defocus: A Spatial Domain Approach. *Int. J. on Computer Vision* 13(3), pp. 271–294.
- Subbarao, M., Wei, T.-C., 1992. Depth from Defocus and Rapid Autofocusing: A Practical Approach. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 773–776.
- Szczepanski, W., 1958. Die Lösungsvorschläge für den räumlichen Rückwärtseinschnitt. In: Deutsche Geodätische Kommission, Reihe C: Dissertationen, Heft Nr. 29, pp. 1–144.
- Taycher, L., Trevor, J., 2002. Recovering articulated model topology from observed motion. Proc. Neural Information Processing Systems, pp. 1311–1318.
- Thayananthan, A., Stenger, B., Torr, P. H. S., Cipolla, R., 2003a. Learning a kinematic prior for tree-based filtering. Proc. Brit. Machine Vision Conf., vol. 2, pp. 589–598.
- Thayananthan, A., Stenger, B., Torr, P. H. S., Cipolla, R., 2003b. Shape context and chamfer matching in cluttered scenes. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol. I, pp. 127–133.
- Thiemermann, S., 2003. Direkte Mensch-Roboter-Kooperation in der Kleinteilmontage mit einem SCARA-Roboter. Doctoral dissertation, Stuttgart University.

- Torrance, K., Sparrow, E., 1967. Theory for Off-Specular Reflection from Rough Surfaces. *Journal of the Optical Society of America* 57(9), pp. 1105–1114.
- Triggs, W., 1997. Auto-calibration and the absolute quadric. *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, pp. 609–614.
- Tsai, R. Y., 1987. A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses. *IEEE J. of Robotics and Automation RA* 3(4), pp. 323–344.
- Tsechpenakis, G., Metaxas, D., Neidle, C., 2008. Combining Discrete and Continuous 3D Trackers. In: Rosenhahn, B., Klette, R., Metaxas, D. (eds.), *Human Motion: Understanding, Modelling, Capture and Animation*, Springer-Verlag, Dordrecht, The Netherlands.
- Turk, M., 2005. Multimodal Human Computer Interaction. In: Kisacanin, B., Pavlovic, V., Huang, T. S. (eds.), *Real-Time Vision for Human-Computer Interaction*, Springer Verlag Berlin Heidelberg, pp. 269–283.
- Turtle, E. P., Jaeger, W. L., Schenk, P. M., 2007. Ionian mountains and tectonics: Insights into what lies beneath Io's lofty peaks. In: Lopes, R. M., Spencer, J. R. (eds.), *Io after Galileo, A New View of Jupiter's Volcanic Moon*, Springer Praxis Publishing, Chichester, UK.
- Van der Mark, W., Gavrila, D. M., 2006. Real-Time Dense Stereo for Intelligent Vehicles. *IEEE Trans. on Intelligent Transportation Systems* 7(1), pp. 38–50.
- Vedula, S., Baker, S., 2005. Three-dimensional scene flow. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 27(3), pp. 475–480.
- Everka, J., Helfenstein, P., Hapke, B. W., Goguen, J. D., 1988. Photometry and polarimetry of Mercury. In: Vilas, F., Chapman, C. R., Shapley Matthews, M. (eds.), *Mercury*, The University of Arizona Press, Tucson, USA.
- Viéville, T., Lingrand, D., 1996. Using singular displacements for uncalibrated monocular vision systems. INRIA Technical Report 2678.
- Viola, P. A., Jones, M. J., 2001. Rapid object detection using a boosted cascade of simple features. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 511–518.
- Vögtle, T., Steinle, E., 2000. 3D modelling of building using laser scanning and spectral information. In: *Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XXXIII-B3B, pp. 927–933.
- von Bank, C., Gavrila, D. M., Wöhler, C., 2003. A Visual Quality Inspection System Based on a Hierarchical 3D Pose Estimation Algorithm. In: Michaelis, B., Krell, G. (eds.), *Pattern Recognition, Proc. 25th DAGM Symposium*, Magdeburg, Germany. Lecture Notes in Computer Science 2781, pp. 179–186, Springer-Verlag Berlin Heidelberg.
- Wachsmuth, S., Wrede, S., Hanheide, M., Bauckhage, C., 2005. An active memory model for cognitive computer vision systems. *KI Journal* 19(2), Special Issue on Cognitive Systems, pp. 25–31.
- Wang, L., Yang, R., Davis, J. E., 2007. BRDF Invariant Stereo Using Light Transport Constancy. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 29(9), pp. 1616–1626.

- Warell, J., 2004. Properties of the Hermean regolith: IV. Photometric parameters of Mercury and the Moon contrasted with Hapke modelling. *Icarus* 167(2), pp. 271–286.
- Wei, T., Klette, R., 2004. Fourier transform based methods for height from gradients. Proc. Control, Automation, Robotics and Vision Conference, Kunming, China, vol. 1, pp. 85–91.
- Weiss, R., Boldt, M., 1986. Geometric grouping applied to straight lines. Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, pp. 489–495.
- Weitz, C. M., Head, J. W., 1999. Spectral properties of the Marius Hills volcanic complex and implications for the formation of lunar domes and cones. *J. Geophys. Res.* 104(E8), pp. 18933–18956.
- Weng, J., Huang, T. S., Ahuja, N., 1989. Motion and structure from two perspective views: algorithms, error analysis and error estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 11(5), pp. 451–476.
- Westling, M., Davis, L., 1996. Object recognition by fast hypothesis generation and reasoning about object interactions. Proc. Int. Conf. on Pattern Recognition, vol. IV, pp. 148153.
- Whitaker, E. A., 1999. Mapping and Naming the Moon. A History of Lunar Cartography and Nomenclature. Cambridge University Press, Cambridge, UK.
- Wieczorek, M. A., Zuber, M. T., Phillips, R. J., 2001. The role of magma buoyancy on the eruption of lunar basalts. *Earth and Planetary Science Letters* 185, pp. 71–83.
- Wildey, R. L., 1975. Generalized photoclinometry for Mariner 9. *Icarus* 25, pp. 613–626.
- Wilhelms, D. E., 1964. A photometric technique for measurement of lunar slopes. In: Astrogeologic Studies, Annual Progress Report, Part D: Studies for Space Flight Program, USGS preliminary report, pp. 1–12.
- Wilhelms, D. E., 1987. The geologic history of the Moon. USGS Prof. Paper 1348, USGS, Flagstaff, USA.
- Wilhelms, D. E., McCauley, J. F., 1971. Geologic Map of the Near Side of the Moon. USGS, Flagstaff, Arizona.
- Williams, D. J., Shah, M., 1992. A fast algorithm for active contours and curvature estimation. *Computer Vision, Graphics Image Processing (Image Understanding)* 55, pp. 14–26, 1992.
- Wilson, L., Head, J. W., 1996. Lunar linear rilles as surface manifestations of dikes: theoretical considerations. Proc. Lunar Planet. Sci. XXVII, abstract #1445.
- Wilson, L., Head, J. W., 2002. Tharsis-radial graben systems as the surface manifestations of plume-related dike intrusion complexes: models and implications. *J. Geophys. Res.* 107(E8), p. 5057.
- Wilson, L., Head, J. W., 2003. Lunar Gruithuisen and Mairan domes: Rheology and mode of emplacement. *J. Geophys. Res.* 108(E2), pp. 5012–5018.
- Winkler, K. (ed.), 2006. Three Eyes Are Better than Two. SafetyEYE uses technical image processing to protect people at their workplaces. In: DaimlerChrysler Hightech Report 12/2006, DaimlerChrysler AG Communications, Stuttgart, Germany.

- Wöhler, C., 2008. Image-based 3D surface reconstruction in the macroscopic and microscopic domain using geometric and photopolarimetric cues. Proc. Oldenburger 3D-Tage, pp. 244–253, Oldenburg, Germany.
- Wöhler, C., Anlauf, J. K., 2001. Real-time object recognition on image sequences with the adaptable time delay neural network algorithm – applications for autonomous vehicles. *Image and Vision Computing* 19(9–10), pp. 593–618.
- Wöhler, C., d'Angelo, P., 2009. Stereo image analysis of non-Lambertian surfaces. *Int. J. of Computer Vision* 81(2), pp. 172–190.
- Wöhler, C., d'Angelo, P., Krüger, L., Kuhl, A., Groß, H.-M., 2009. Monocular 3D scene reconstruction at absolute scale. *ISPRS Journal of Photogrammetry and Remote Sensing*, in press, DOI 10.1016/j.isprsjprs.2009.03.004.
- Wöhler, C., Hafezi, K., 2005. A general framework for three-dimensional surface reconstruction by self-consistent fusion of shading and shadow features. *Pattern Recognition* 38(7), pp. 965–983.
- Wöhler, C., Krüger, L., 2003. A Contour Based Stereo Vision Algorithm for Video Surveillance Applications. *SPIE Visual Communication and Image Processing*, Lugano, vol. 5150(3), pp. 102–109.
- Wöhler, C., Lena, R., Bregante, M. T., Lazzarotti, P., Phillips, J., 2006a. Vertical studies about Rupes Cauchy. *Selenology* 25(1), pp. 7–12.
- Wöhler, C., Lena, R., Lazzarotti, P., Phillips, J., Wirths, M., Pujic, Z., 2006b. A combined spectrophotometric and morphometric study of the lunar mare dome fields near Cauchy, Arago, Hortensius, and Milichius. *Icarus* 183, pp. 237–264.
- Wöhler, C., Lena, R., Pau, K. C., 2007a. The lunar dome complex Mons Rümker: Morphometry, rheology, and mode of emplacement. *Proc. Lunar and Planet. Sci. XXXVIII*, abstract #1091.
- Wöhler, C., Lena, R., Phillips, J., 2007b. Formation of lunar mare domes along crustal fractures: Rheologic conditions, dimensions of feeder dikes, and the role of magma evolution. *Icarus* 189(2), pp. 279–307.
- Wolff, L. B., 1987. Shape from polarization images. *Computer Vision Workshop* 87, pp. 79–85.
- Wolff, L. B., 1989. Surface orientation from two camera stereo with polarizers. *Optics, Illumination, Image Sensing for Machine Vision IV*, SPIE vol. 1194, pp. 287–297.
- Wolff, L. B., 1991. Constraining Object Features Using a Polarization Reflectance Model. *IEEE Trans. Pattern Analysis and Machine Intelligence* 13(7), pp. 635–657.
- Wolff, L. B., Angelopoulou, E., 1994. Three-Dimensional Stereo by Photometric Ratios. *Journal of the Optical Society of America* 11(11), pp. 3069–3078.
- Wood, C. A., 1973. Moon: Central peak heights and crater origins. *Icarus* 20, pp. 503–506.
- Wood, C. A., Andersson, L., 1978. New morphometric data for fresh lunar craters. *Proc. Lunar Planet. Sci. IX*, pp. 3369–3389.
- Woodham, R. J., 1980. Photometric Method for Determining Surface Orientation from Multiple Images. *Optical Engineering* 19(1), pp. 139–144.

- Wu, S. S. C., Elassal, A. A., Jordan, R., Schafer, F. J., 1982. Photogrammetric application of Viking orbital photography. *Planetary and Space Science* 30(1), pp. 45–55.
- Wu, S. S. C., Schafer, F. J., Jordan, R., Howington, A. E., 1987. Topographic map of Miranda. *Proc. Lunar Planet. Sci. XVIII*, pp. 110–111.
- Wu, S. S. C., Doyle, F. J., 1990. Topographic mapping. In: Greeley, R., Batson, R. M. (eds.), *Planetary Mapping*, Cambridge University Press, Cambridge, UK.
- Xiong, Y., Shafer, S., 1993. Depth from Focusing and Defocusing. *DARPA Image Understanding Workshop*, Palo Alto, California, USA, pp. 967–976.
- Xu, C., Prince, J. L., 1998. Snakes, shapes, and gradient vector flow. *IEEE Trans. on Image Processing* 7(3), pp. 359–369.
- Ye, Q. Z., Ong, S. H., Han, X., 2001. A stereo vision system for the inspection of IC bonding wires. *Int. J. of Imaging Systems and Technology* 11(4), pp. 254–262.
- Yoon, Y., DeSouza, G. N., Kak, A. C., 2003. Real-time tracking and pose estimation for industrial objects using geometric features. *Proc. Int. Conf. on Robotics and Automation*, pp. 3473–3478.
- Yu, Y., Chang, J. T., 2002. Shadow Graphs and Surface Reconstruction. *Europ. Conf. on Computer Vision*, Copenhagen, pp. 31–45.
- Zabih, R., Woodfill, J., 1994. Non-parametric transforms for computing visual correspondence. *Proc. Europ. Conf. on Computer Vision*, pp. 151–158.
- Zhang, Z., 1992. Iterative point matching for registration of free-form curves. Technical report no. 1658, Institut National de Recherche en Informatique et en Automatique (INRIA) Sophia Antipolis, France.
- Zhang, Z., 1998. A Flexible New Technique for Camera Calibration. Microsoft Research Technical Report MSR-TR-98-71.
- Zhang, Z., 1999a. Flexible Camera Calibration By Viewing a Plane From Unknown Orientations. *Proc. Int. Conf. on Computer Vision*, pp. 666–673.
- Zhang, Z., 1999b. Iterative point matching for registration of free-form curves and surfaces. *Int. J. on Computer Vision* 13(2), pp. 119–152.
- Zhang, L., Curless, B., Seitz, S., 2003. Spacetime Stereo: Shape recovery for dynamic scenes. *Proc. Computer Vision and Pattern Recognition*, vol. 2, pp. 367–374.
- Zhou, H., Huang, T. S., 2003. Tracking articulated hand motion with eigen dynamics analysis. *Proc. Int. Conf. on Computer Vision*, Washington, DC, USA, pp. 1102–1109.
- Zickler, T., Belhumeur, P. N., Kriegman, D. J., 2002. Helmholtz Stereopsis: Exploiting Reciprocity for Surface Reconstruction. *Proc. Europ. Conf. on Computer Vision*, pp. 869–884.
- Zickler, T., Belhumeur, P. N., Kriegman, D. J., 2003. Toward a Stratification of Helmholtz Stereopsis. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 548–554.
- Zickler, T., Ho, J., Kriegman, D. J., Ponce, J., Belhumeur, P. N., 2003. Binocular Helmholtz Stereopsis. *Proc. Int. Conf. on Computer Vision*, pp. 1411–1417.

- Ziegler, J., Nickel, K., Stiefelhagen, R., 2006. Tracking of the Articulated Upper Body on Multi-View Stereo Image Sequences. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 774–781.
- Zuniga, O. A., Haralick, R. M., 1983. Corner detection using the facet model. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 30–37.