X.media.publishing

Christian Wöhler

3D Computer Vision

Efficient Methods and Applications

Springer

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

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, photoclinometry 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

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.

C. Wöhler, *3D Computer Vision: Efficient Methods and Applications,*
DOI: 10.1007/978-3-642-01732-2_1, © Springer-Verlag Berlin Heidelberg 2009

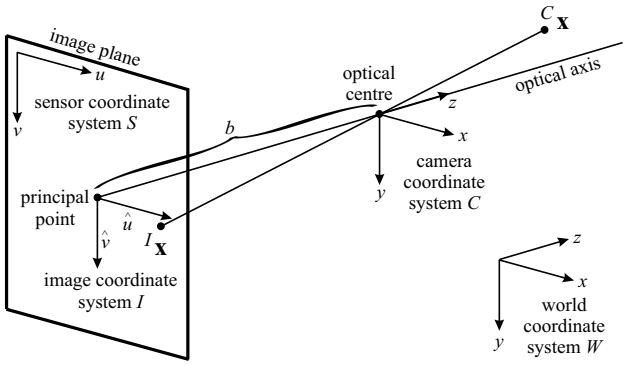

Fig. 1.1 The pinhole camera model. A scene point ${}^C\mathbf{x}$ defined in the camera coordinate system is projected into the image point ${}^I\mathbf{x}$ located in the image plane.

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 ${}^C_W T$, where the lower index defines the original coordinate system and the upper index the coordinate system into which the point is transformed. The transformation ${}^C_W T$ corresponds to an arbitrary rotation and translation. In this notation, the transformation is given by ${}^C\mathbf{x} = {}^C_W T\, {}^W\mathbf{x}$. 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 ${}^{C_i}_{W}T$, transforming a point ${}^{W}\mathbf{x}$ from the world coordinate system W into the camera coordinate system C_i. The transformation ${}^{C_i}_{W}T$ 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

$$ {}^{C_i}\mathbf{x} = {}^{C_i}_{W}T\left({}^{W}\mathbf{x}\right) = R_i\,{}^{W}\mathbf{x} + \mathbf{t}_i. \tag{1.2} $$

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

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

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

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

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

$$ {}^{S}\mathbf{x} = {}^{S}_{I}T\left({}^{I}\mathbf{x}\right) \tag{1.5} $$

from the image coordinate system into the sensor coordinate system is defined in the general case by an affine transformation ${}^{S}_{I}T$ (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 $\mathscr{P}\left({}^{C_i}_{W}T, \{c_j\}_i, {}^{W}\mathbf{x}\right)$ 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

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

Since Eq. (1.1) is based on Euclidean geometry, it is nonlinear in z, implying that the function $\mathscr{P}$ 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

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

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

$$[R_i \mid \mathbf{t}_i]. \tag{1.8}$$

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

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

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

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

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 $i = 1, \ldots, N$, 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 $i = 1,\ldots,N$ and $k = 1,\ldots,K$. The image coordinates inferred from the extrinsic camera parameters ${}^{C_i}_{W}T$, 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

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

The transformation by ${}^{S_i}_{I_i}T^{-1}$ 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 ($\theta = 90°$) and the pixels are quadratic ($\alpha_u = \alpha_v$). This special case corresponds to ${}^{S_i}_{I_i}T$ 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 ${}^{C_i}_{W}T$, $\{c_j\}_i$, 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 ${}^{C_i}_{W}T$ 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 "classicial" 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 ${}^{C_1}_{W}T$ corresponds to unity while the relative orientation of camera 2 with respect to camera 1 is given by ${}^{C_2}_{W}T$ 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 ${}^{I_1}\mathbf{x} = (\hat{u}_1, \hat{v}_1)^T$ in image 1 corresponds to a ray through the origin of the camera coordinate system according to

$$ {}^{C_1}\mathbf{x} = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} \hat{u}_1 s \\ \hat{v}_1 s \\ bs \end{pmatrix}, \tag{1.12} $$

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

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

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 ${}^{I_2}\mathbf{x} = (\hat{u}_2, \hat{v}_2)^T$ are given by

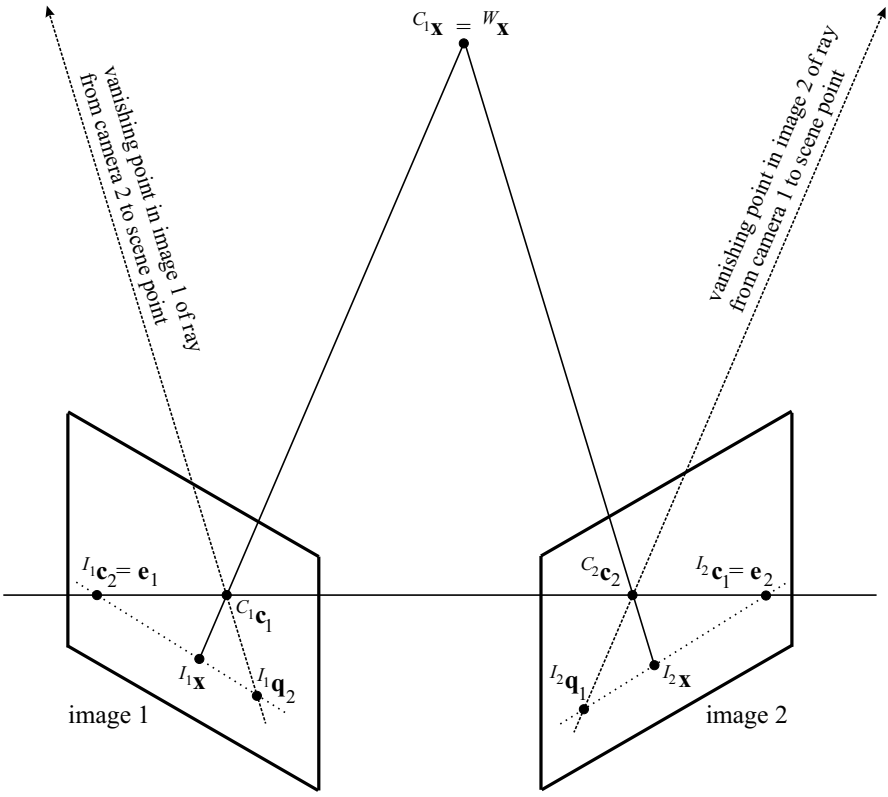

Fig. 1.2 Definition of epipolar geometry. The epipolar lines of the image points $^{I_1}\mathbf{x}$ and $^{I_2}\mathbf{x}$ are drawn as dotted lines, respectively.

$$\frac{\hat{u}_2}{b} = \frac{x_2}{z_2} \qquad \text{and} \qquad \frac{\hat{v}_2}{b} = \frac{y_2}{z_2}, \tag{1.14}$$

assuming identical principal distances for both cameras. With the abbreviations

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

we now obtain the relations

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

describing a straight line connecting the point $(p/r, q/r)^T$ for $s = 0$ with the point $(d/f, e/f)^T$ for $s \to \infty$. 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

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

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 ${}^{I_i}\mathbf{x} = ({}^{I_i}\hat{u}, {}^{I_i}\hat{v})^T$, which are related to the pixel coordinates given by ${}^{S_i}\mathbf{x} = ({}^{S_i}u, {}^{S_i}v)^T$ 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

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

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

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

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

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

where $E = [\mathbf{t}]_\times R$ 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

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

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

$$\begin{aligned} \left(A_2^{-1} \, {}^{S_2}\tilde{\mathbf{x}}\right)^T \left(\mathbf{t} \times R A_1^{-1} \, {}^{S_1}\tilde{\mathbf{x}}\right) &= 0 \\ {}^{S_2}\tilde{\mathbf{x}}^T A_2^{-T} \left(\mathbf{t} \times R A_1^{-1} \, {}^{S_1}\tilde{\mathbf{x}}\right) &= 0 \\ {}^{S_2}\tilde{\mathbf{x}}^T \, F \, {}^{S_1}\tilde{\mathbf{x}} &= 0, \end{aligned} \tag{1.20}$$

where $F = A_2^{-T} E A_1^{-1}$ 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

${}^{S_2}\tilde{\mathbf{l}}$ in image 2 which corresponds to a point ${}^{S_1}\tilde{\mathbf{x}}$ in image 1 is given by ${}^{S_2}\tilde{\mathbf{l}} = F\ {}^{S_1}\tilde{\mathbf{x}}$. 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 ${}^{S_2}\tilde{\mathbf{l}}$. Accordingly, the line ${}^{S_1}\tilde{\mathbf{l}} = F^T\ {}^{S_2}\tilde{\mathbf{x}}$ 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 ${}^{S_2}\tilde{\mathbf{l}} = F\ {}^{S_1}\tilde{\mathbf{x}}$ 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 $\tilde{\mathbf{e}}_2^T \left(F\ {}^{S_1}\tilde{\mathbf{x}}\right) = \left(\tilde{\mathbf{e}}_2^T F\right)\ {}^{S_1}\tilde{\mathbf{x}} = 0$ 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 $\left(P_1, P_2, \{{}^{W}\tilde{\mathbf{x}}_i\}\right)$, 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 ambigous up to a projective transformation H, where H is an arbitrary 4×4 matrix. Hence, the projective reconstruction given by $\left(P_1, P_2, \{{}^{W}\tilde{\mathbf{x}}_i\}\right)$ is equivalent to the one defined by $\left(P_1 H, P_2 H, \{H^{-1}\ {}^{W}\tilde{\mathbf{x}}_i\}\right)$.

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

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

A more general form of P_2 is

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

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 ${}^{S_1}\tilde{\mathbf{x}} \times (P_1\ {}^{W}\tilde{\mathbf{x}}) = \mathbf{0}$, which allows to express the matrix G as

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

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 jth 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

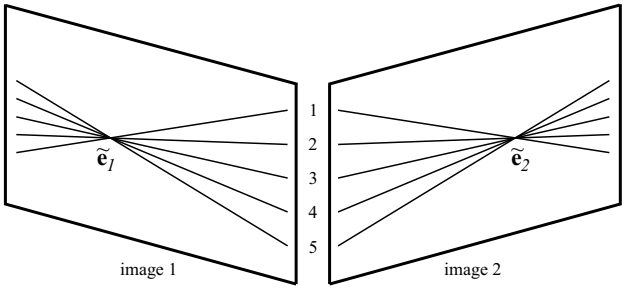

Fig. 1.3 In each of the two images, the epipolar lines form a pencil of lines. The intersection points correspond to the epipoles $\tilde{\mathbf{e}}_1$ and $\tilde{\mathbf{e}}_2$. Corresponding pairs of epipolar lines are numbered consecutively.

point (cf. Fig. 1.3). For the pencils of epipolar lines in image 1 and 2, the intersection points correspond to the epipoles $\tilde{\mathbf{e}}_1$ and $\tilde{\mathbf{e}}_2$. 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 ${}^{S_1}\mathbf{x}^{(e)}$ and ${}^{S_2}\mathbf{x}^{(e)}$. 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 $\mathbf{t}$ 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 stereoplotting 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

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

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

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

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

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

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:

$$\begin{aligned} u &= \frac{L_1x+L_2y+L_3z+L_4}{L_9x+L_{10}y+L_{11}z+1} \\ v &= \frac{L_5x+L_6y+L_7z+L_8}{L_9x+L_{10}y+L_{11}z+1}. \end{aligned} \tag{1.27}$$

If we use the abbreviations $b_u = b/k_u$, $b_v = b/k_v$, and $D = -(x_0r_{31}+y_0r_{32}+z_0r_{33})$, the parameters $L_1 \ldots L_{11}$ can be expressed as

$$\begin{aligned} L_1 &= \frac{u_0r_{31}-b_ur_{11}}{D} \\ L_2 &= \frac{u_0r_{32}-b_ur_{12}}{D} \\ L_3 &= \frac{u_0r_{33}-b_ur_{13}}{D} \\ L_4 &= \frac{(b_ur_{11}-u_0r_{31})x_0+(b_ur_{12}-u_0r_{32})y_0+(b_ur_{13}-u_0r_{33})z_0}{D} \\ L_5 &= \frac{v_0r_{31}-b_vr_{21}}{D} \\ L_6 &= \frac{v_0r_{32}-b_vr_{22}}{D} \\ L_7 &= \frac{v_0r_{33}-b_vr_{23}}{D} \\ L_8 &= \frac{(b_vr_{21}-v_0r_{31})x_0+(b_vr_{22}-v_0r_{32})y_0+(b_vr_{23}-v_0r_{33})z_0}{D} \end{aligned}$$