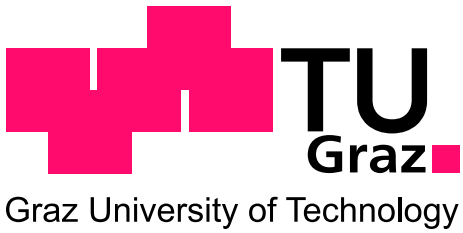


Self-calibrating Cameras in Video Surveillance

Roman Pflugfelder

Dissertation
Graz, May 2008
Submitted to the Faculty of Computer Science
Graz University of Technology, Austria

Adviser

Univ.-Prof. Dipl.-Ing. Dr.techn. Horst Bischof
Institute for Computer Graphics and Machine Vision

Referee

Dr. James Ferryman
Computational Vision Group
School of Systems Engineering
University of Reading, UK

The system should also be able to exploit multiple camera configurations with inter-camera integration. ... A well-developed cognitive vision surveillance system should also be portable to new contexts within the same application scenario as well as to new applications scenarios, for example, the ability to easily reconfigure a train-station surveillance system for an airport, either by design or autonomously.

Competencies of a Cognitive Surveillance System, An EU Research Roadmap of Cognitive Vision, 23. August 2005

Abstract

This thesis addresses the automatic calibration of two static surveillance cameras in a man-made world with orthogonal and parallel structures and a common ground plane. An approach is taken where the calibration of the interior orientation, the undistortion of the lens and the calibration of a camera's rotation to the world perform **before** calibrating the camera centers, which allows methods that work in slightly overlapping as in non-overlapping views.

We present a new **incremental** calibration composed of Expectation Maximization and Simulated Annealing that uses the uncertainties of noisy line segments to process a video stream instead of a single image. The advantage of video is that orthogonal and parallel edge information in a dynamic world can appear in different parts of the image at different time instants which can improve the estimates.

Following this approach, the thesis studies localization in slightly overlapping views by using point correspondences from ordinary people detectors. The success of Nister's 5-point algorithm is experimentally shown, whereas Hartley's 8-point algorithm fails. The reason is the more accurate rotation estimation with vanishing points as solely with point correspondences.

The most important insight of the thesis is a **linear** extended DRP-method for localization that works in non-overlapping **and** overlapping views. The method guarantees a **global** optimal solution for the simultaneously estimated exterior orientation and reconstructed 3-D positions of a **freely** moving object. To estimate rotation, the four basic 90 deg rotations around the ground plane normal are evaluated w.r.t the re-projection error.

Finally, the thesis applies the approach for people tracking between the cameras. Once the geometry is known, the matching can be solved by triangulation. The idea to expand triangulation by constraints that arise from a person's motion model is successful in non-overlapping views. Evaluations with the PETS 2006 dataset and two self-acquired datasets show promising results despite of large appearance changes, illumination changes, wide baseline and low resolution.

Kurzfassung

Diese Arbeit beschäftigt sich mit automatischen Methoden zur Kalibrierung von zwei fix montierten Überwachungskameras in künstlichen Umgebungen mit rechtwinkeligen und parallelen Strukturen und einer gemeinsamen Grundebene. Es wird ein Ansatz vorgestellt, der **zuerst** die innere Orientierung, die Linsenverzeichnung und die Rotation zu der Szene in jeder Kamera separat bestimmt und darauf aufbauend eine Bestimmung der Translation zwischen den beiden Kamerazentren durchführt. Diese Vorgangsweise erlaubt eine Methode, die in wenig überlappenden wie auch in nicht-überlappenden Sichtbereichen arbeitet.

Zuerst wird eine inkrementelle Methode bestehend aus "Expectation Maximization" und der "Simulierten Abkühlung" (engl. Simulated Annealing) unter Rücksichtnahme der Unsicherheiten in den zugrunde liegenden Daten (Bildkanten) vorgestellt. Nicht mehr ein Bild sondern eine ganze Bildfolge wird verarbeitet. Dies hat den Vorteil, dass Veränderungen über die Zeit wie Beleuchtungsveränderungen oder sich durch die Szene bewegende Objekte berücksichtigt werden. Diese Veränderungen können neue Informationen über die Strukturen in der Szene liefern, die die Parameter des Abbildungsmodells verbessern.

Dem Ansatz folgend, zeigen wir experimentell in wenig überlappenden Sichtbereichen den erfolgreichen Einsatz von Nisters 5-Punkt Algorithmus und den erfolglosen Einsatz von Hartleys 8-Punkt Algorithmus. Der Algorithmus scheitert, da die Rotationsbestimmung ausschließlich auf den unsicheren Punkt-Korrespondenzen des Personendetektors beruht. Im Gegensatz dazu basiert der Ansatz dieser Arbeit auf den durch die Bildkanten genauer bestimmbaren Fluchtpunkten der Szene. Die Arbeit zeigt, dass die verbleibende Mehrdeutigkeit in der Rotation durch diese ungenauen Punkt-Korrespondenzen aufgelöst werden kann.

Der wichtigste Beitrag dieser Arbeit ist eine Erweiterung einer bekannten **linearen** Methode (engl. DRP-method) zur Bestimmung der Kamerazentren, die in nicht-überlappenden **und** in überlappenden Sichtbereichen arbeitet. Die Methode garantiert eine **globale**, optimale Lösung der Kamerazentren und der 3-D Trajektorie eines sich **frei** bewegenden Objekts. Die Mehrdeutigkeit der Rotation wird durch eine Evaluierung des Rückprojektionsfehlers aus der endlichen Menge an 90 Grad Rotationen um den Normalenvektor der Grundebene ermittelt.

Schlußendlich zeigt die Arbeit, dass sich die entwickelte Kalibriermethode zum automatischen Verfolgen von Personen eingesetzen lässt. Sobald die Kamerageometrie bekannt ist, lässt sich das Verfolgen der Personen durch Triangulation lösen. Dies ist selbst in nicht-überlappenden Sichtbereichen unter Berücksichtigung des Bewegungsmodells der Personen möglich. Experimente mit dem PETS 2006 Datensatz und mit zwei selbst aufgenommenen Datensätzen zeigt vielversprechende Ergebnisse unter schwierigen Bedingungen wie großen Veränderungen in der Erscheinung der Personen, starker Beleuchtungsveränderungen, großen Abständen zwischen den Kameras und geringer Auflösung.

Acknowledgments

Nobody keeps up four years of researching without the motivation and help of others. One person I want to thank is my adviser Horst Bischof, who motivated me steadily to work in the field of Computer Vision for the last nine years.

Thanks to my former chief Helmut Schwabach. He has always believed in Intelligent Video and supported me in getting the funding for my research.

Thanks to all people at the Austrian Research Centers, who have helped me with their interest and discussions. Michael Nölle for his deep knowledge in statistics and German humor. I hope really that you will understand one day what "information" means. Gustavo Fernández for his proof-reading and his patience to endure me as room mate. Nabil Belbachir for the time during lunch where we had discussions and spinning new ideas. Norbert Brändle for the image dataset and for your mental support during our quarterly beer drinking nights. Angelika Hölbling and Bernhard Strobl for helping me with the financial and project-specific matters and finally Markus Clabian who has gracefully supported me as chief during my last year of research.

I owe much to the Institute for Computer Graphics and Vision, where I have never felt as external doctoral student. Thanks to Helmut Grabner, Peter Roth, Clemens Arth, Christian Leistner, Martin Winter and Michael Grabner for their discussions and for sharing software.

Many others supported me with their discussions. Some of them are Allen Hanbury, Georg Langs, Horst Wildenauer, Branislav Micusik, Thomás Pajdla, René Donner, Dave McKinnon, Hongdong Li and Brian Lovell.

A graceful "Thank you!" to Anja for her love and her invaluable support during the bad times. Last but not least I thank my family who have always supported me throughout my whole live.

This thesis was supported by the Austrian Research Centers GmbH - ARC and by the "Nachwuchsförderung des Forschungsförderungsfond FFG (former Forschungsförderungsfonds der gewerblichen Wirtschaft FFF)", projects 807.375, 809.674, 811.778 and 813.377. I acknowledge the funding of my dissertation by the Republic of Austria.

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

C-0	Initial calibration
C-1	Initial calibration with known rotation
C-2	Optimal calibration
C-3	Incremental calibration
EM	Expectation Maximization
IAC	Image of the Absolute Conic
LM	Levenberg-Marquardt
L-0	Localization with DRP
L-1	Localization with Nister's 5-point method
L-3	Extended DRP-method
MLE	Maximum Likelihood Estimator (Estimation)
RANSAC	Random Sample Consensus
SA	Simulated Annealing
SFM	Structure from Motion
SIFT	Scale Invariant Feature Transform
SLAT	Simultaneous Localization and Tracking
SVD	Singular Value Decomposition
VRML	Virtual Reality Modeling Language

Chapter 1

Introduction

Video surveillance plays a major role in today's security technologies. Hundreds of cameras observe public places such as train stations, airports and private properties. Security is one of our basic needs and especially in the last years a steady increase in security can be denoted. What makes video so demanding and competitive to other sensors is the rich information it offers - based on the well known slogan: "A picture says more than thousand words".

But which information is of actual interest in video surveillance? For sure not the everyday life, it is the suspicious behavior of someone or something, the unusual moments, the threads. Fortunately, such events happen rarely, but exactly for this reason such events are difficult to observe in time by a human operator. Simply the vast amount of video data overwhelms a human and exactly this reason calls for automatic systems.

Video surveillance is a promising branch of Computer Vision, a fascinating interdisciplinary field of Computer Science, Mathematics and Physics that tries to create machines that are able to see and understand. Although automatic systems for video surveillance are an upcoming Billion dollar market [Freeman, 2007], it is still in its infancy. In contrast to industrial vision, mature products are currently not existing, because the real world with all its unexpected incidents is hard to control.

Video surveillance covers many aspects that are current problems in Computer Vision research. Reliable object detection (recognition) is one of these problems. Robust object tracking under occlusion and in plane rotations in one or several cameras is another one. To solve these problems, several attempts were made in the past. For example, Coifman et al. [1998] showed in the PATH project a vehicle tracking system that operates in real-time and under difficult conditions like congestion, illumination change and shadows. A prominent attempt to bring research forward was the three years VSAM project. Object detection and object tracking in a forest of cameras was the objective [Collins et al., 2000]. After the end

of the project, the results run into a start-up company ObjectVideo* which is one of the commercial leaders in video surveillance today. Honeywell developed a system with name DETER that was able to fuse activities in several cameras into one overall view [Pavlidis et al., 2001].

All these projects were established in the US, but there are also activities in Europe. In this context we mention the ADVISOR project† [Siebel and Maybank, 2004] and the AVITRACK project‡ [Thirde et al., 2005], where the former treated video surveillance in metro stations and the latter emphasized the same technology at airport aprons. Object detection and object tracking are by far not the only themes in video surveillance, but on a lower level they are the most important ones. The analysis of activities or the detection of unusual events and behaviors is usually build on top of these two functionalities.

1.1 Self-calibration - A need in video surveillance?

Robust tracking of objects between two cameras is the ticket to video surveillance among many cameras. Such systems are attractive, because they offer an effective handling of object occlusions and they cover large areas of observation. [Shah et al., 2007].

Tracking is in its core a correspondence problem which is beside segmentation the main unsolved problem in Computer Vision. Fig. 1.1 illustrates this difficulty of matching the same person in three views of a train station. Known methods use either a combination of image features or geometry. Some methods use both information sources together as cues to establish a correspondence [Javed et al., 2008]. Unfortunately, the matching of images features between two different cameras is error-prone, because the illumination of the scene vary or the geometry and the appearance of the objects may drastically change; consider the difference in appearance between Camera 3 and 4.

The geometry, once known, does not suffer from these effects. The assumption that objects of interest, for example humans and vehicles, occupy a certain volume of space and that two objects cannot occupy the same volume of space is under practical considerations always valid, hence, the laws of physics and geometry reduce the correspondence problem to a combinatorial problem. This combinatorial (assignment) problem is unfortunately NP-complete§ [Nister et al., 2007] but as long as the number of objects and cameras is low, the Hungarian algorithm is an elegant solution [Kuhn, 1955]. Note that the number of objects also depends on the geometry itself, because the number of objects increases with larger views.

However, a manual calibration of the cameras needs an expensive expert and it is a tedious

*http://www.objectvideo.com; as at 13/02/2008.
†http://www-sop.inria.fr/orion/ADVISOR/finalreport.html; as at 13/02/2008.
‡http://www.cvg.rdg.ac.uk/projects/avitrack/index.html; as at 13/02/2008.
§They showed a proof with missing observations.

Figure 1.1: Three simultaneously taken images of the PETS 2006 dataset; Camera 3 (left), Camera 1 (middle) and Camera 4 (right). The same imaged person is framed in blue. Note the huge change in appearance which makes visual correspondence so hard. The correct match shows that once the geometry is known, the correspondence problem is simple to solve.

work - just imagine the calibration of hundreds of cameras. In the worst case, the calibration must be repeated each time a change in the environment happens which makes maintenance nearly impossible.

Another serious problem of well known calibration methods [Hartley and Zisserman, 2004; Svoboda et al., 2005] is the need for point correspondences that are uniformly distributed within the images. Unfortunately, cameras for video surveillance have usually slight overlapping views which provide point correspondences in small image areas. Sometimes cameras are still within a few meters distance, but they have non-overlapping views and do not provide a single point correspondence [Pflugfelder and Bischof, 2006a].

The lack of approaches for self-calibrating cameras with slight and non-overlapping views asks for a solution and this thesis tries to give an answer.

1.2 The self-calibration problem

Let us consider for a moment a similar problem that is Structure from Motion (SFM) [Ma et al., 2004]. A single, moving camera captures images of a static 3-D structure, whereas the objective is to reconstruct this structure by using the images and to self-calibrate the cameras in the consecutive 3-D positions. One could imagine the self-calibration of static surveillance cameras with the help of moving objects as the dual problem.

The standard SFM approach is to use putative point correspondences in the first few frames to estimate the epipolar geometry by using efficient algebraic methods like the well known 8-point algorithm. Putting such methods into a Random Sample Consensus (RANSAC) framework [Fischler and Bolles, 1981] assures robustness. This projective reconstruction is then upgraded to a metric reconstruction by using auto-calibration [Hartley and Zisserman, 2004]. The root of auto-calibration is that the interior orientation stays the same while

the camera is capturing images. Finally, the initialization is updated over the whole image sequence by Bundle Adjustment techniques [Engels et al., 2006].

Our approach of self-calibration is a dual approach. The static cameras extract orthogonal and parallel structures in the scene and estimate the interior orientation by using single-view calibration. This calibrated cameras are then used to estimate the missing similarity transformation, that is, the rotation, the translation and the scale between the cameras. At least five putative point correspondences are sufficient for this localization.

The problem of self-calibration is defined more mathematically as follows: Given the synchronous image sequences of two cameras $i \in \{1, 2\}$ that show moving objects. Each object j at each time instant t in 3-D position $\mathbf{X}_j^t$ is represented by a single image point $\mathbf{x}_{ij}^t$. Assume that the scene captured by the two cameras is the same rigid scene with orthogonal and parallel elements. Self-calibration is the step-wise estimation of the camera matrices P_1 with

$$\lambda \mathbf{x}_{1j}^t = P_1 \begin{pmatrix} \mathbf{X}_j^t \\ 1 \end{pmatrix} = K_1 R_1 \bar{R} \left[I_3 \; - \mathbf{C_1} \right] \begin{pmatrix} \mathbf{X}_j^t \\ 1 \end{pmatrix} \tag{1.1}$$

and P_2 with

$$\lambda \mathbf{x}_{2j}^t = P_2 \begin{pmatrix} \mathbf{X}_j^t \\ 1 \end{pmatrix} = K_2 R_2 \left[I_3 \; - \mathbf{C_2} \right] \begin{pmatrix} \mathbf{X}_j^t \\ 1 \end{pmatrix}. \tag{1.2}$$

The interior orientation matrices K_1, K_2 and the rotation matrices R_1, R_2 are estimated by orthogonal and parallel structures in the images. As the scene is the same in both views, it is used as a large calibration object. However, a rotational ambiguity $\bar{R}$ still remains that is solved together with the estimation of the camera centers $\mathbf{C}_1$ and $\mathbf{C}_2$ in a second step by using putative point correspondences of the moving objects.

The estimation of the $\mathbf{X}_j^t$ is basically not in the consideration of self-calibration, however, the reader will see later that an estimation is necessary in case of non-overlapping cameras which leads to the recent field of Simultaneous Localization and Tracking (SLAT) [Taylor et al., 2006].

Our approach to solve the self-calibration problem contains several assumptions about the cameras, the scene and the moving object. In the following, these assumptions are explicitly summarized:

Manhattan world: The environment viewed by the cameras must contain orthogonal directions. These directions are given by static or dynamic objects with straight lines that are partly imaged as line segments. Usually, man-made structures are build in an orthogonal manner. Most of the line segments must be the images of such orthogonal directions [Coughlan and Yuille, 1999].

Natural camera: The skew of the image sensor's pixel is assumed to be zero and the aspect

ratio of the pixel's side lengths is assumed to be constant and known. Most digital cameras are natural [Liebowitz and Zisserman, 1999]. In some critical cases, vanishing points lie close to infinity or only two prominent directions are present. Under such conditions the principal point is set to the image center.

Natural pose: The mounting of the camera must be rigid and the pose of the camera must be in an upright position, that is, the abscissa of the image is nearly parallel to the ground plane.

Synchronization: The cameras are synchronized by an external trigger.

The reader can convince himself that these assumptions are often valid in practice. Hence, the approach is applicable in many indoor and outdoor environments.

1.3 Arguments for this approach

Our interest are camera settings where the views are slightly overlapping or where the views are non-overlapping. The former case has the speciality that the images of moving objects are present in small parts of the image. However, standard multi-view calibration assumes evenly distributed points. One can show that for example the 8-point algorithm fails quickly with a significant amount of noise in the points, that is, when the error is comparable to the distance of a neighboring point, the computation of the Fundamental matrix fails [Pflugfelder and Bischof, 2006b]. Unfortunately, image points on moving objects are always afflicted by substantial errors. Such errors are usually several pixels large. Especially the estimation of the rotation matrix suffers from this fact. The advantage of our approach is that beside the estimation of the interior orientation, the rotation matrix except for the final ambiguity is also estimated from the line segments. The line segments are part of the background and they are detected with sub-pixel accuracy. The ambiguity is limited to a finite number of possible rotation matrices that are a priori known, because we assume a Manhattan world and a Natural pose.

The reader could argue that at least in the case of overlapping views, Scale Invariant Feature Transform (SIFT) [Lowe, 2004], Maximally Stable Extremal Regions (MSER) [Matas et al., 2002] or Harris corner detectors [Forsyth and Ponce, 2002] could deliver sub-pixel accurate points - so why points on moving objects? Unfortunately, an empirical evaluation of several real scenes debilitates this argument. For example, consider the result of SIFT point matching in Fig. 1.2. The matching code is available on the web*. Only four point correspondences between camera 1 and 3 are found; too few for any calibration approach†. Apart from this

*http://www.cs.ubc.ca/ lowe/keypoints/; as at 14/02/2008.

†The estimation of the Essential matrix needs at least five point correspondences [Nister, 2004].

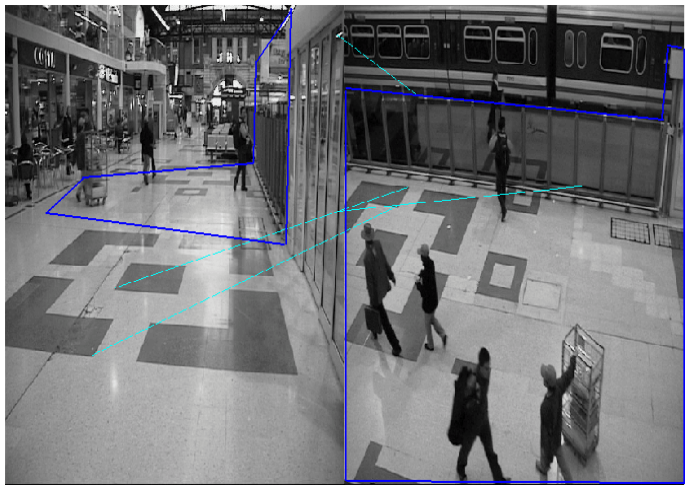


Figure 1.2: SIFT point matching is not always successful. The images of two neighboring cameras are shown. The whole image was automatically generated by Lowe's SIFT method. The cyan lines depict the wrong matches. The blue contours show the opposite camera's visible field of view. Although large image areas overlap, these areas are characterized by low resolution, high perspective distortion and low texture which makes a correct matching with features of the image background extremely difficult.

problem, these point correspondences are all wrong, because SIFT does not work reliable in less textured images [Bay et al., 2005]. This example shows large homogeneous areas due to the floor tiles. But less textured images that suffer under severe perspective distortion with low resolution in many image areas are the normality and not the exception. Unfortunately, we got a similar result with MSER so that these features are also not usable in this scene.

We mentioned in the last section that self-calibration does not need to reconstruct the 3-D positions of the moving objects. Rahimi et al. [2004] showed that a calibration without this reconstruction of the positions is impossible when we consider cameras with non-overlapping views which is the previously mentioned SLAT problem; more about SLAT in Chap. 3. Their iterative method optimizes a non-linear objective function with all the disadvantages of non-linear optimization.

This thesis proposes a linear and efficient method that does not suffer from these deficiencies. It localizes the cameras and reconstructs the positions simultaneously, but assumes the interior orientation and the rotation matrices to be known which is true when the calibration follows our step-wise approach.

1.4 Previous work

All previous systems assume the interior orientation as known and a common ground plane. Image points are rectified up to a similarity transformation which are the images of coplanar points on this ground plane [Liebowitz and Zisserman, 1998].

The first work that is related to ours is from Lee et al. [2000]. They assume overlap between three cameras. The system tracks vehicles and generates trajectories by the vehicle's center of gravity. Then the plane-induced inter-camera homographies are estimated by an alignment process which is a non-linear optimization in the ground plane. An automatic synchronization is embedded by assuming that the minimum of the objective function is found at the right time offset between the video sequences. Then the rotation matrices and the translations between three overlapping cameras are recovered from these homographies.

The quality of the solution relies on the accuracy of the tracker, therefore Stein [1999] experimented with a successive dense alignment of the images which is inefficient and does not scale well to more than three cameras.

Instead of a metric recovery of the calibration parameters, Black and Ellis [2001]; Black et al. [2002] content themselves with the plane-induces homographies. Synchronization is done in the same way as in the work of Stein et al.

Caspi et al. [2006] were the first who discovered the advantages of trajectories instead of static points for the matching and then camera recovery. The restriction of planar motion is also relaxed and instead of considering a homography, they consider the fundamental matrix. The approach that the temporal information in trajectories compensate for the lack of point correspondences in non-overlapping cameras was for the first time mentioned by them [Caspi and Irani, 2002] and independently mentioned again two years later by Rahimi et al. [2004].

A further work in the same direction is the one of Stauffer and Tieu [2003]. He concentrated more on a reliable model for point correspondence. He also pointed out that vehicles mostly drive along a straight line which will lead to a degenerate solution for the homography and that the choice of the points must contribute for that.

Perhaps the most similar approach to ours is by Jaynes [2004]. He uses orthogonal and parallel line segments to find vanishing points, hence recovering the rotation of the camera to the common ground plane. Line segments allow a more accurate estimation as trajectories of moving objects. However, he used walking people to generate trajectories that are rectified and aligned within a non-linear optimization similar to Stein et al. Where a Manhattan world does not exist, he draws manually line segments. Synchronization is done by an external trigger.

1.5 Contribution

This thesis contributes in various ways to existing work.

Regarding the calibration of single cameras:

- A method that estimates the interior orientation, the rotation to the world and the

lens distortion in one framework by using line segments. The method measures the error in the endpoints [Liebowitz, 2001] and optimizes in the parameters of the geometric model [Schindler and Dellaert, 2004] instead of optimizing the vanishing points. The nonlinear optimization is embedded in an Expectation Maximization (EM) framework [Dempster et al., 1977] that guarantees a locally optimal estimation with respect to the error in the endpoints which are assumed to be Gaussian.

- Instead of a brute-force initialization, we combine RANSAC with the EM framework. This decision is encouraged by Forsyth and Ponce [2002]. A plausibility test with the focal length of the lens guarantees a correct initialization.

- The method is incremental which has two advantages: (i) many images of the same static scene help to integrate out image noise, which can be of particular interest, especially, when noise-levels are high, and (ii) the method can account for new line segments that can suddenly appear, because illumination changes or new objects move through the scene. We propose a novel idea to update the parameters only when they improve and prohibit the updating when line segments, which are basically the source of information, temporally disappear. This update is based on the uncertainty of the data and an approach motivated by Simulated Annealing (SA) [Duda et al., 2001].

Regarding two slightly overlapping views, we propose a localization based on the independent self-calibration of each single camera that is possible with at least three point matches instead of five [Nister, 2004], All these points can be collinear.

Finally we propose:

- A method that is able to localize cameras with overlapping or non-overlapping views.

- This method is the first one that does not assume objects moving on a common plane or/and at constant speed. The only assumption is a smooth motion.

- It is computationally attractive, because the problem is solved in closed-form by Singular Value Decomposition (SVD) [Golub and van Loan, 1996] which guarantees also to find a global minimum.

1.6 Thesis outline

The overall layout of the thesis is as follows:

Chap. 2: This chapter describes a new approach to calibrate the interior orientation and the rotation to the scene of a single camera automatically and incrementally from an

image sequence or even on-line from a video stream. It is a consequent continuation and meaningful combination of previous work, which considered only single images.

Chap. 3: This chapter describes methods that will estimate the missing camera centers and the final rotational ambiguity between two cameras under different conditions and assumptions. The emphasis is laid on distant cameras with slight overlapping views and with non-overlapping views, where our approach of successive calibration and then localization has significant advantages compared to methods in the literature. The main result is a method that works in overlapping and non-overlapping situations.

Chap. 4: This chapter discusses the possible accuracy of the approach under realistic conditions by experiments with synthetic and real image data. During all the experiments we had people matching as application in mind. Thus, we expected the worst case reconstruction error of a person smaller than the average volume a person occupies in space which is in width approximately between half a meter and a meter.

Finally, the chapter shows two applications in video surveillance for self-calibrating cameras. On the one hand we show that kernel-based tracking is significantly improved by the use of geometry. On the other hand, we show that the matching across two cameras is solely possible with slight and with non-overlapping views.

Chap. 5: This chapter summarizes the work and the contributions that are done within this thesis. We discuss the conclusions and outline avenues of future work.

1.7 Publications

The publications [Pflugfelder and Bischof, 2003, 2005, 2006a,b, 2007a,b] emerged during the PhD work with Horst Bischof and contributed to this thesis.

1.8 Notation

Vectors are written by bold, small arabic letters, for example $\mathbf{a}$, throughout the thesis. The vector's coordinates have the same letter, but are written normally and with consecutively numbered subscripts, for example, a_1, a_2, ...

Matrices are written by italic, large arabic letters, for example A, and their elements are the same small letters with consecutively numbered subscripts, for example, a_{11}, a_{12}, ..., whereas the first subscript denotes the row and the second subscript the column.

Functions are upright, small letters, for example, f$(\cdot)$. Sets are written in calligraphic style, for example, $\mathcal{S}$. Greek symbols usually denote thresholds or parameters.

Chapter 2

Incremental auto-calibration of single cameras

This chapter describes a new approach to calibrate a single camera automatically and incrementally from an image sequence or even on-line from a video stream. It is a consequent continuation and meaningful combination of previous work, which considered only single images. Sec. 2.1 will describe this previous work in detail and will elaborate the important issues carefully. Exactly these issues are tackled by the approach which is treated in the Sec. 2.2 - 2.8.

The novelties of this approach are as follows: (i) all important issues are considered by one single method, (ii) many images of the same static scene help to integrate out image noise, which can be of particular interest, especially, when noise-levels are high, and (iii) the method can account for new edge information that can suddenly appear, that is, a light goes on, or objects with valuable edge information move into or through the scene.

2.1 Previous and related work

The work in this chapter owes much to the previous work of others, because many approaches in research fields like vanishing point detection, lens distortion and camera calibration were gathered and combined to form a novel method. This section traces these approaches from the past until today and uncovers their strengths to emulate and weaknesses to avoid. Despite the deep understanding of the geometric principles in these research fields [Faugeras and Luong, 2001; Hartley and Zisserman, 2004; Ma et al., 2004; McGlone, 2004; Semple and Kneebone, 1998], this section sheds light onto the important issues, which have still remained unclear in literature. One can imagine that the literature in these research fields is widely ramified within Computer Vision and Photogrammetry and sometimes in some issues contradictory.

No attempt has been made to provide an exhaustive and complete list. All references were chosen, because they either document an important step forward in research or contributed significantly to our work.

The literature review is divided into four sections. Sec. 2.1.1 traces the progress of vanishing point detection over nearly 50 years. In Sec. 2.1.2 the focus shifts to lens distortion which has significant influence on the accuracy of calibration and is unjustifiably neglected by many authors. Sec. 2.1.3 shows the steady evolution of single camera calibration over the past. Finally, Sec. 2.1.4 will summarize the important issues that must be tackled by an accurate and robust calibration method.

2.1.1 Detection of vanishing points

It is known since the 16th century that vanishing points contain valuable information about image formation [Kemp, 1992]; their localization enables the geometric interpretation of scenes such as the calibration of a single camera. Criminisi et al. [2000] explains vanishing points, their geometric properties and their history in greater detail. The first attempts to develop methods for the detection of vanishing points date back to the late 1970s; a time, Computer Vision was in its early roots. Since then, a steady activity in this research field can be denoted. Vanishing points are often distant to the image center. Nearly all methods use line segments as source of information to compute vanishing points and small errors in the line segments forming concurrent lines have a drastic effect on the location of vanishing points; a fact that makes an accurate detection method difficult to develop.

2.1.1.1 Search space

Line segments and vanishing points need a convenient parametrization in a search space, which allows a robust and accurate detection, even in critical cases such as vanishing points at or close to infinity. The literature has discussed so far the image plane [Caprile and Torre, 1990; Coughlan and Yuille, 2003; McLean and Kotturi, 1995; Minagawa et al., 1999; Sekita, 1994], the polar space [Ballard and Brown, 1982; Cantoni et al., 2001; Nakatani and Kitahashi, 1980], Tuytelaars's line parameter space [Seo et al., 2006; Tuytelaars, 1998], the sphere* [Antone and Teller, 2000; Badler, 1974; Barnard, 1983; Brillault-O'Mahony, 1991; Cipolla and Boyer, 1998; Collins and Weiss, 1990; Gallagher, 2002; Kosecka and Zhang, 2002; Lutton et al., 1994; Magee and Aggarwal, 1984; Quan and Mohr, 1989; Shufelt, 1999; Van den Heuvel, 1998], the space of all pairwise concurrent lines [Rother, 2002b] and the projective plane [Kanatani, 1996; Liebowitz and Zisserman, 1998; Pflugfelder et al., 2005; Rother, 2003;

*Instead of sphere, the name Gaussian sphere occurs frequently in the literature. The term Gaussian was coined by the famous German mathematician Carl Friedrich Gauss (1777 - 1855) and refers to a spherical surface stated in the Gauss' law of electric flux.

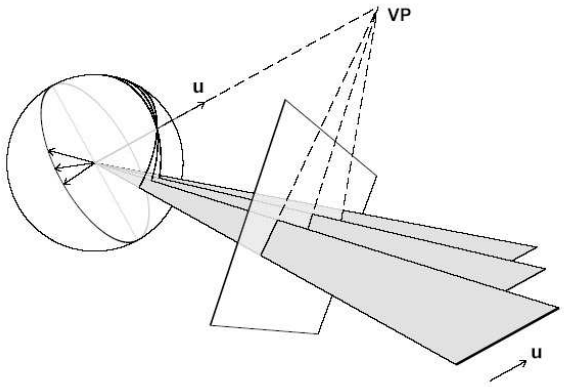


Figure 2.1: The Gaussian sphere [Collins, 1993]. Direction **u** is a point on the sphere which is formed by the great circles of parallel image lines. VP is the vanishing point on the image plane.

Schaffalitzky and Zisserman, 2000; Sturm and Maybank, 1999].

The polar space is a line parameter space in (r,ϕ) with r the normal distance between a line and the origin on the image plane and ϕ the enclosed angle [Duda and Hart, 1972]. Instead of the polar parametrization of a line, the incidence relation between a point $(u\ v)^\top \in \mathbb{R}^2$ and a line $(a\ b)^\top \in \mathbb{R}^2$ with

$$au + b + v = 0, \tag{2.1}$$

provides the second well known line parametrization. Tuytelaars [1998] described a clever partitioning into three bounded subspaces, that is, the (a,b)-space for $|a| \leq 1$ and $|b| \leq 1$, the $(1/a,b/a)$-space for $|a| \geq 1$ and $|b| \leq |a|$ and finally the $(1/b,a/b)$-space for $|b| \geq 1$ and $|a| \leq |b|$. In contrast to the polar space, the sphere is centered at the origin of the camera coordinate system (Fig. 2.1). The origin and each point on the sphere's surface defines a directional vector in the camera coordinate system. The sphere is frequently mentioned in the literature as a unit sphere, that is, a sphere with radius one, because the length of a directional vector has no relevance and is usually normalized to length one. Parallel straight lines in the camera coordinate system are partly imaged as line segments onto the image plane. These line segments project onto circles on the sphere. Such circles on the sphere are called great circles and always intersect in two points. These points represent the two directions in which the straight lines are orientated. The straight line passing through both points pierces the image plane in one point, that is, the vanishing point of the line segments. The plane spanned by a line segment and the origin is called the interpretation plane.

In a nutshell, a proper search space shall fulfill the following six properties [Barnard, 1983; Rother, 2002b]:

Property 1 All points and lines in the image plane have a finite parametrization.

Property 2 The search space is bounded. No singularities exist.

Property 3 No correlations in the parametrization exist.

Property 4 A metric to measure error distances exists.

Property 5 A point in the image plane maps preferably to a point or a line in the search space.

Property 6 The metric in the search space is locally nearly linear in terms of its effect to a metric in the image plane. The mapping does not distort the original metric.

Rother [2002b] pointed out that no search space can simultaneously fulfill Property 2 and 6. He explained that the original distances in the image plane are not preserved by the mapping. This well known fact is a consequence of Girard's theorem in spherical geometry [Henderson and Taimina, 2005]. The image plane itself obviously does not fulfill Property 1 and 2, thus, is very problematic in critical cases. The polar space does not fulfill Property 1, 2 and 5, because lines outside of the image have no parametrization and points in the image are mapped to cosines in the polar space, respectively. The literature always mentions the polar space as bounded space, however, it is limited to lines clipping the image. Tuytelaars's line parameter space and the sphere fulfill all properties except Property 6, because the mapping from the image to both spaces introduces a distortion. Especially, the sphere is popular in the literature. The projective space of all pairwise concurrent lines and the projective plane does not fulfill Property 4 and 6. Additionally, the projective plane does not fulfill Property 3, because points up to a scalar value are equivalent. However, a normalization before distances are measured can circumvent this metric lack.

2.1.1.2 Grouping and estimation

Vanishing points are computed from concurrent lines which are found by a grouping of corresponding line segments. Many authors [Badler, 1974; Caprile and Torre, 1990; Cipolla and Boyer, 1998; Kanatani, 1996; Liebowitz and Zisserman, 1998; Nakatani and Kitahashi, 1980; Sturm and Maybank, 1999] did the grouping of line segments by hand and considered only the estimation of the vanishing points. For example, concurrent lines are detected by the Hough transform [Hough, 1962] and the vanishing point is determined by the mean of all possible intersections either on the sphere [Badler, 1974] or in the image plane [Caprile and Torre, 1990; Nakatani and Kitahashi, 1980]. Another way is to exploit the incidence axiom of concurrent lines also either on the sphere [Cipolla and Boyer, 1998] or in the image plane [Kanatani, 1996; Sturm and Maybank, 1999]. Liebowitz and Zisserman [1998] formulated the dual perception of incidence by the vanishing point incident to all lines which fit the line segments. Collins and Weiss [1990] describes a potential detection method in two steps.

A first step groups line segments to concurrent lines and a second, successive step estimates the location of the intersection, that is, the vanishing point.

Clustering approaches

Gallagher [2002]; Kosecka and Zhang [2002]; Magee and Aggarwal [1984]; Van den Heuvel [1998] used clustering on the sphere to perform these two steps. Gallagher [2002]; Magee and Aggarwal [1984] determined the interpretation planes of each line segment and then computed the cross products of their normal vectors. Their approach finds then clusters of cross products supporting a vanishing point by the k-means algorithm [Duda et al., 2001]. Clustering has several advantages in contrast to the Hough transform. On the one hand clustering does not need any quantization and on the other hand the clustering process is able to detect cross products as outliers. Van den Heuvel [1998] tested the triple incidence of cross products statistically. His approach runs through all possible combinations and initializes clusters when the test succeeds. After a final clustering the cluster with the highest support is chosen as a vanishing point. Kosecka and Zhang [2002]; McLean and Kotturi [1995] proposed an approach that generates a histogram over the orientation of the line segments. They found out that the orientation of line segments in images of man-made scenes form peaks in the histogram.

The Hough transform

The literature has frequently emphasized the Hough transform* [Hough, 1962] as a possible approach to perform either both grouping and estimation at once [Ballard and Brown, 1982; Barnard, 1983; Cantoni et al., 2001; Lutton et al., 1994; Quan and Mohr, 1989; Shufelt, 1999; Tuytelaars, 1998] or only the grouping [Antone and Teller, 2000; Brillault-O'Mahony, 1991; Collins and Weiss, 1990]. The Hough transform is a well known method to detect image primitives, originally lines, in a quantized Hough space, originally the polar space of line parameters. It increments for each point in the image particular accumulator cells in the Hough space. Cells with local maxima indicate detected image primitives. The Hough transform requires a bounded Hough space, hence, the image plane is out of the question.

Ballard and Brown [1982]; Cantoni et al. [2001] studied the polar space as Hough space. In a sense, this approach seems obvious, however, parallel line segments produce collinear points in the Hough space. This fact lets the cosine collapse into a straight line. The equation system will quickly become ill-conditioned the more distant a vanishing point is, because the line parameter points are packed in a very small area of the Hough space.

Bräuer-Burchardt and Voss [2002] had the same idea as Kosecka and Zhang [2002]; McLean and Kotturi [1995], namely to group line segments using their orientation. Despite a histogram his approach uses a Hough transform into the polar space. Line segments with

*The Hough transform was originally developed and patented by the IBM engineer Paul Hough.

approximately the same orientation in the image create clusters in the polar space that can be detected as peaks.

In contrast to the polar space, the Hough transform is able to perform alternatingly between Tuytelaars's Hough space and the image plane; an elegant solution to detect first lines then vanishing points and finally vanishing lines. The quantization of the subspaces has the consequence that accumulator cells more distant from the image grow larger and larger. Tuytelaars justified this irregular quantization with the argument that points lying further away are normally less accurately determined anyhow and that shifts in their position have less impact in the image; an argument that we mentioned at the beginning of this section. We agree that the uncertainty in the estimation grows with the distance, but the accuracy of the estimation should depend only on the line segments and the numerical accuracy of the estimation method and not on a growing quantization.

Barnard [1983] had the idea to use the sphere as Hough space. The sphere is popular, because it is a convenient Hough space at a first glance. However, one of its shortcomings is similar to the drawback of Tuytelaars's space. The inverse gnomonic projection introduces a distortion. Cells more distant to the meridian cover larger and larger areas on the image plane. Hence, Barnard's work initiated improvements of the sphere's tessellation [Lutton et al., 1994; Quan and Mohr, 1989]. Lutton et al. proposed an irregular quantization of the spherical coordinates along the longitude and a regular quantization along the latitude. She also described a probability mask that weights the accumulator with the aim to compensate for the bias introduced by the finite extent of the image; the Hough transform to a sphere tend to increment cells in the visible image. However, vanishing points are also likely outside the image. Lutton et al. further modeled the error in line segments by a swath model, which allows to increment all cells falling in a swath about the great circle. The swath size is defined by the possible interpretation planes passing through the noisy endpoints of the line segments. Quan and Mohr tried to overcome the other well known problem of the Hough transform, which is the cell size. Smaller cells lead to higher accuracy, but at the price of loosing the computational benefits of the Hough transform. Thus, Quan and Mohr sampled the sphere with respect to a particular point on the sphere from a coarse to a fine resolution by using a hierarchical Hough transform.

Collins and Weiss [1990] appreciated the Hough transform as elegant solution for the clustering, but saw the estimation of vanishing points as statistical problem. Inspired by Magee and Aggarwal [1984], they interpreted the cross products as random samples of a Bingham distribution on the sphere [Bingham, 1974]. The mean of the Bingham distribution is the expected vanishing point. Later, Brillault-O'Mahony [1991] improved the statistical error model by introducing an error model in the line segments. This approach combines in a beneficial way the amenities of the Hough transform to solve the combinatorial explosion in

the number of line segments and the accuracy in the estimation.

Further approaches

Grammatikopoulos et al. [2003]; Rother [2002b] replaced the image plane with the space of all possible intersections of line segments. This space is bounded, so it is able to generate a histogram over all intersections whereas the frequencies are proportional to the probability that a particular intersection is a vanishing point. They used a voting technique similar to the Hough transform to compute these frequencies; an intersection point will receive more votes, if more line segments intersect. Methods using RANSAC were also proposed [Pflugfelder et al., 2005; Rother, 2003; Schaffalitzky and Zisserman, 2000]. RANSAC acts here as a search engine. Originally invented by Fischler and Bolles [1981], it is a random algorithm for robust fitting of models in the presence of outliers. The algorithm searches a consensus set within the total number of line segments that intersect all in a single point. The assumption is here that intersection points with a large support of line segments are vanishing points. RANSAC has several advantages over the previously mentioned method, for example, no computation of all possible intersection points is necessary anymore. Pflugfelder et al. [2005] used RANSAC consecutively for the detection of several vanishing points in a scene. After each run the consensus set is removed from all line segments. This new set of line segments is then used in the next run.

An interesting approach that uses walking people to extract the vanishing points was proposed by Lv et al. [1999]. As the estimation of the vertical rotation axis of a person is difficult, their approach is inaccurate. Three years later, Krahnstoever and Mendoca [2005] presented a Bayesian formulation that showed promising improvements. Krahnstoever and Mendonca [2006] also showed that the assumption of constant velocity will further improve the accuracy.

An EM framework

Antone and Teller [2000]; Kosecka and Zhang [2002]; Minagawa et al. [1999]; Pflugfelder et al. [2005]; Sekita [1994] did not see the grouping and estimation as consecutive steps. They rather recognized grouping and estimation as an alternating process, which is formulated in a statistical framework assuming a specific error model in the line segments. Antone and Teller [2000] mentioned that a wrong grouping of line segments done by the Hough transform will have no chance to be corrected during the estimation process, if grouping and estimation are seen separately. Sekita [1994] was probably one of the first who showed that the EM algorithm is an adequate solution in this case. He used EM to group concurrent lines in the image plane. The EM algorithm [Dempster et al., 1977] performs clustering and estimation by alternating between finding the best clustering given the current estimate of the vanishing point (E-step), and finding the best estimate of the vanishing point given the current clustering (M-step). The advantage of EM is that it works in a continuous space and

it is guaranteed to converge to the optimal solution given an initialization close to the optimal solution. In case of a bad initialization the EM algorithm can stuck in a local minima. This is the main drawback of EM, because such an initialization is sometimes difficult to achieve. Minagawa et al. [1999] extended this approach to detect not only vanishing points but also vanishing lines. Antone and Teller [2000]; Kosecka and Zhang [2002] showed that the EM algorithm works on the sphere. They showed that the Hough transform is adequate for initialization. Pflugfelder et al. [2005] initialized EM with RANSAC, which seems to be an elegant and quite sophisticated statistical solution in particular for the problem of vanishing point detection but also in general. The latter fact is also mentioned in Forsyth and Ponce [2002].

A priori knowledge about the camera and the scene

The prerequisite of a calibrated camera in sphere based approaches allows to guide the search for vanishing points by exploiting the orthogonality between points on the sphere. Magee and Aggarwal [1984] pointed out that the focal length must not be known exactly, because orthogonality is invariant to a change in focal length. A priori knowledge about the scene also helps to increase robustness and accuracy [Shufelt, 1999]. Instead of a single image, Pflugfelder et al. [2005] processed many images, which is the traditional statistical approach to improve accuracy even further. A deep investigation of statistical methods in geometry is given by Kanatani [1996]. Kosecka and Zhang [2002] showed that a calibrated camera is not mandatory to search for vanishing points on the sphere, because concurrent lines in the image plane will remain concurrent under a projective transformation. However, the orthogonality between points as a constraint in the estimation cannot be exploited. Gallagher [2002] utilized ground truth data of typical indoor and outdoor images to establish prior probabilities and proposed the detection as Bayesian inference on the sphere.

2.1.1.3 Error distance

The grouping of the line segments and the estimation of a vanishing point need the definition of an appropriate error distance between a line segment and a potential vanishing point. The error distance yields the cost function for the grouping and the estimation, hence, it is of great importance. The arc length and the euclidian distance is usually used on the sphere and the planar search spaces respectively. Most people measure these error distances between pairwise intersections of great circles or straight lines, which are fitted to the line segments, except Grammatikopoulos et al. [2003]; McLean and Kotturi [1995]; Rother [2002b], who measure the angle between the straight line and the line passing trough the vanishing point and the midpoint of the line segment. Kanatani [1996] used the normal distance between the straight line and the vanishing point as error distance, however, this error distance has

no geometric meaning. In contrast, Liebowitz and Zisserman [1998]; Pflugfelder et al. [2005]; Rother [2003]; Schaffalitzky and Zisserman [2000] proposed the normal distance between the endpoints of a line segment and a line passing trough the vanishing point and a particular point on the line, not necessarily the midpoint of the line segment, as error distance. In fact, this error distance is the only error distance between a line segment and a vanishing point mentioned in the literature*. One can show that this error distance originally proposed by Liebowitz is independent of the location of vanishing points in the image plane (property 6). The cost function that is minimized during the estimation of a vanishing point will (i) have no singularities, (ii) be optimal under a Gaussian error mode, (iii) be locally continuous to the second order and (iv) be locally quadratic.

Except for the pure Hough transform approaches, the estimation of vanishing points is on the one hand usually formulated as linear, least-squares problem with a closed-form solution by SVD† [Kanatani, 1996] and on the other hand as a non-linear, least-squares problem that can be solved iteratively by the Levenberg-Marquardt algorithm [Liebowitz and Zisserman, 1998].

2.1.2 Lens distortion

Many authors ignore the distortion of lenses and concentrate solely on vanishing point detection. However, lenses of practical interest with a small to moderate focal length usually possess a noneligible distortion. The detection of vanishing points is rather sensible to small distortions which yield to severe errors; especially in the case of distant vanishing points.

Several models of distortion have been discussed in the literature [Clarke and Fryer, 1998; Freyer and Brown, 1986]. The most common model for modern lenses is the radial distortion model. The distortion increases radially with the distance to a radial center. The common opinion that the radial center conforms to either the image center or the principal point cannot generally be accepted, because the lens is never perfectly centered above the image sensor [Hartley and Kang, 2005]. Moreover, Devernay and Faugeras [2001] showed that the aspect ratio of the image sensor is not necessarily the same as the pixel aspect ratio used in the radial distortion model. This difference accounts for a tangential distortion of the lens. The consideration of the first order coefficients of the radial model is entirely sufficient in the case of lenses with good quality [Zhang, 2000]. He showed that first order polynomials can achieve an average accuracy of $\frac{1}{10}$ pixel in the image.

Interestingly, the effect of radial distortion is independent of the imaging process. Hence, the estimation of the parameters of the radial lens distortion model can be completely decoupled

*All other error distances are measured between a straight line and a vanishing point.

†The reader is advised to look at http://web.mit.edu/18.06/www/Video/video-fall-99-new.html as at 06/02/2008 for an excellent explanation of the SVD by Gilbert Strang's video lecture.

from the calibration of the internal parameters [Bräuer-Burchardt, 2004; Cornelis et al., 2002; Devernay and Faugeras, 2001; Li and Hartley, 2005]. To find the parameters of the distortion model, Devernay and Faugeras [2001] used the invariance of straightness with respect to a projective transformation, that is, following a pinhole camera model straight lines in space are projected to straight lines in the image. The approach detects short edges and fits a polygon onto them with high tolerance. Then, the image and edges are straightened out iteratively by following the distortion model and optimizing its parameters until a best fit transforms edges to line segments. The author noted that their method is completely automatic and makes only the assumption of a man-made world where only straight lined edges exist. The handicap of this method is the question after the tolerance level in the polygon fitting. If the value is too large, the algorithm will try to transform edges to line segments which perhaps are not images of straight lines. Contrary, if the value is too small the algorithm will fail to succeed.

Another well-known approach is to use the fact that parallel lines intersect all in a vanishing point after a projective transformation [Ahmed and Farag, 2005; Becker and Bove, 1995]. The roots of this idea are in Photogrammetry and can be traced back to the so-called "plumb-line method" introduced by [Brown, 1971]. Especially with simultaneous vanishing point detection this approach is useful [Bräuer-Burchardt, 2004; Grammatikopoulos et al., 2003; Van den Heuvel, 1999]. Bräuer-Burchardt [2004], for example, computed vanishing points and the distortion alternatingly. He constructed an ideal line passing through the vanishing point and the midpoint of a line segment. Then, the points composing the line segments are projected onto the line. These projected points and the original ones allow for an estimation of the coefficients of the polynomial models and the radial center.

2.1.3 Calibration

The vanishing points of three orthogonal directions are quite common in man-made environments. Thus, vanishing points are a powerful geometric information to calibrate a camera internally in so-called "Manhattan worlds" [Coughlan and Yuille, 2000].

The seminal work of single-view calibration based on vanishing points in the Computer Vision community is the work done by Caprile and Torre [1990]. This paper was the starting point for more interest; although earlier groundwork exist within the Photogrammetry community [Gracie, 1968; Grammatikopoulos et al., 2003; Karras and Petsa, 1999]. Caprile and Torre used the projected edges of a cuboid to compute the three finite, orthogonal vanishing points in the image. As the world has three main directions, i.e. two horizontal and one vertical, three orthogonal vanishing points are always present - at least at infinity. The orthogonality provides constraints on the cuboid to image plane transformation. Together with a further constraint of a natural camera, i.e. zero skew, constant aspect ratio, Caprile and

Torre were able to estimate the remaining three internal parameters, i.e. focal length and principal point.

[Cipolla et al., 1999] used the idea of Caprile and Torre to reconstruct partial 3D models of architectural scenes, e.g. a church. Thereby, they used pairs of uncalibrated images to estimate (i) the internal parameters by Caprile and Torre's method (ii) exploit the epipolar geometry to find the full projection matrices and (iii) created a VRML-model by using triangulation and texture mapping. In contrast to Caprile and Torre, they did a re-estimation of the vanishing points after knowing the internal parameters. They showed that this approach converges on the one hand to an optimal estimation of the vanishing points and on the other hand to an optimal estimation of the internal parameters.

Another groundwork was done by Faugeras et al. [1992]. He and his colleagues embedded the calibration problem within a projective geometry framework. They found a direct interrelationship between the image of the absolute conic and the internal camera parameters. Thus, it is possible to calibrate a camera internally by knowing the image of the absolute conic and vice versa. Based on this fundamental discovery, Liebowitz and Zisserman [1998] described the internal camera calibration as an application of metric plane rectification. If length ratios, angles or equal angles on a plane are known, it is possible to define constraints on the image of the circular points. At least 4 circular points are necessary to constrain the image of the absolute conic and to solve the calibration problem. Furthermore, Liebowitz and Zisserman systematically reported several scene and camera constraints on the image of the absolute conic, e.g. the above mentioned metric constraints on a plane, orthogonal vanishing points and the natural camera constraint. They discussed several cases of degeneracies of the absolute conic, whenever less than 5 constraints are available.

Beside the internal calibration of a camera its orientation to the ground plane and the distance of the optical center to a world reference point (external calibration) is also an interesting property. Liebowitz [2001] alludes that the orientation can be computed with known internal parameters and known vanishing line of the ground plane. The latter is the image of the intersection of the ground plane with the plane at infinity. The vanishing line can be found by at least two sets of parallel straight lines on the ground plane, e.g. two horizontal and orthogonal vanishing points. Criminisi et al. [2000] further showed that the camera position from a fixed world reference point can be estimated by knowing the vanishing line and the vanishing point perpendicular to the ground plane.

Every calibration approach discussed so far was formulated as a problem with a geometric interpretation and mostly with an algebraic solution. Deutscher et al. [2002] showed a probabilistic approach to calibrate a camera internally and to orientate the camera with respect to the ground plane. They used a likelihood of an image conditional on internal calibration [Coughlan and Yuille, 2000], supplied a prior distribution on the internal calibration and

gave an estimation of the posterior by iterated importance sampling. They demonstrated reasonable accuracy with the a priori knowledge of scene and camera priors. Problems of the estimation are mentioned with different lightning and faint images.

Instead of importance sampling that needs severe computational resources, Kosecka and Zhang [2002]; Pflugfelder et al. [2005] proposed the EM algorithm as a framework for vanishing point detection and successive calibration. The posterior of the vanishing points given the line segments is computed within the EM framework. Then, the internal parameters are estimated given the estimates of the vanishing points as discussed in the beginning. Schindler and Dellaert [2004] criticized the calibration as a postprocessing step. They argued with the fact that the vanishing points are of little interest and that the optimization should affect the parameters of the camera model directly. Hence, they proposed a top-down optimization of the actual parameters where the vanishing points are only an intermediate result that are directly derived from the parameters. Consequently, vanishing points are only used to group the line segments in the E-step.

2.1.4 Remarks

From the previous and related work we can remark the following important issues:

- The combination of the projective plane as space for the detection of vanishing points and the image plane as space to measure the error distance between line segments and these vanishing points fulfill properties 1 - 6 of a proper search space (Sec. 2.1.1).
- The mapping from the image plane to a sphere or from the image plane to Tuytelaars's subspaces does not preserve the distances. Hence, the cost function can be arbitrarily small and flat.
- The error distance of Liebowitz is the only error distance between a line segment and a vanishing point. It is independent of the location of vanishing points in the image plane. The resulting cost function will well behave, i.e. (i) will have no singularities, (ii) will be optimal under a Gaussian error mode, (iii) will be locally continuous to the second order and (iv) will be locally quadratic.
- The detection of vanishing points is a simultaneous clustering and estimation problem.
- The detection of vanishing points may be optimally solved in an EM framework using the error distance of Liebowitz.
- In contrast to all other detection methods, RANSAC works directly in the image. It is an elegant initialization method in combination with EM.

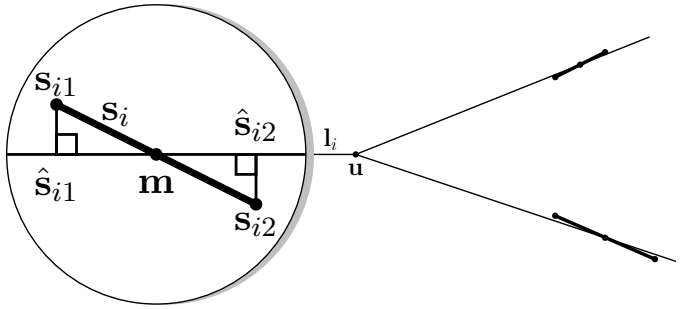


Figure 2.2: The geometry of the optimal intersection problem. The closest fit of a straight line to a line segment $\mathbf{s}_i$ passing exactly through $\mathbf{u}$ is $\mathbf{l}_i$. The endpoints of $\mathbf{s}_i$ are $\mathbf{s}_{i1}$ and $\mathbf{s}_{i2}$ respectively. $\mathbf{m}$ is the intersection point of $\mathbf{l}_i$ and $\mathbf{s}_i$. It is not necessarily the midpoint of the line segment. $\hat{\mathbf{s}}_{i1}$ and $\hat{\mathbf{s}}_{i2}$ are the normal projections of $\mathbf{s}_{i1}$ and $\mathbf{s}_{i2}$ onto $\mathbf{l}_i$.

- Vanishing points do not lie anywhere in the search space. In most cases knowledge about the scene and the camera is available. This knowledge can be utilized to guide the search.

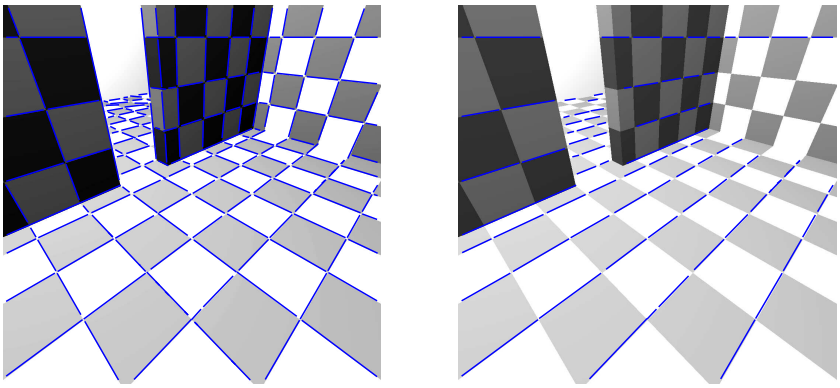
- Lens distortion is usually not negligible. A radial distortion model is sufficient with lenses of good quality. The estimation of the lens distortion can be part of the calibration method using the vanishing points.

- The estimation of the camera parameters should not be a postprocessing step after the detection of vanishing points. The camera parameters should rather be the objective of optimization. One can show that three orthogonal vanishing points are directly derivable from the camera parameters.

2.2 The optimal intersection of line segments

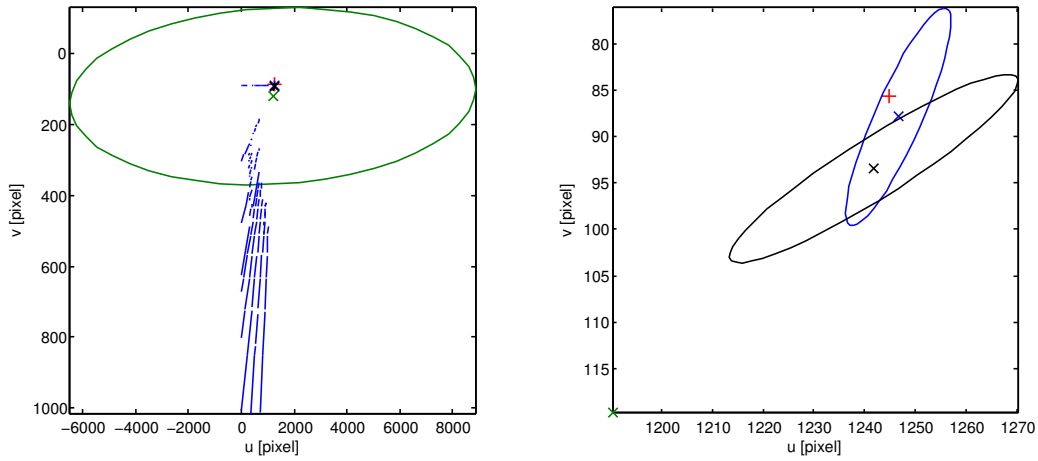
This section will describe the optimal intersection of many line segments. Fig. 2.2 illustrates schematically the geometry of the problem. Imagine the existence of N line segments $\mathbf{s}_1 = (\mathbf{s}_{11}\ \mathbf{s}_{12})^\top, \mathbf{s}_2 = (\mathbf{s}_{21}\ \mathbf{s}_{22})^\top, \ldots, \mathbf{s}_N = (\mathbf{s}_{N1}\ \mathbf{s}_{N2})^\top$ with endpoint measurements $\mathbf{s}_{i1} = (s_{i11}\ s_{i12})^\top$ and $\mathbf{s}_{i2} = (s_{i21}\ s_{i22})^\top$, $1 \leq i \leq N$, in an image; for example, as depicted in blue in Fig. 2.3a. For estimating $\mathbf{s}_{i1}$ and $\mathbf{s}_{i2}$ respectively, we decided to use the Canny operator to detect edges [Canny, 1986], followed by a straight line fitting [Guru et al., 2004].

Consider now a subset of line segments that meet theoretically in a single point, perhaps a vanishing point, as it is illustrated in Fig. 2.3b. In Sec. 2.1.1.3 the reader became acquainted with several approaches from the literature which are able to estimate this intersection point and we concluded in Sec. 2.1.4 that Liebowitz [2001] proposed an optimal method for this problem (Fig. 2.4). In the following we will describe his method in a nutshell.



(a) All detected line segments (b) Concurrent line segments

Figure 2.3: Line segments in a synthetic checkerboard world generated by POV-Ray (http://www.povray.org/; as at 21/06/2007).



(a) A large uncertainty in the mean (b) A closer zoom

Figure 2.4: Comparison between possible intersection points of the line segments in Fig. 2.3b. The true intersection is drawn as red cross. The ellipses depict 99% of the confidence interval. The mean of the intersection points between pairwise line segments (green cross, relative error 4.57% in u, 28.44% in v), the intersection point computed by SVD that has minimal normal Euclidean distance to all elongated line segments (black cross, relative error 0.24% in u, 8.30% in v) and finally the optimal intersection point computed by the Liebowitz's method (blue cross, relative error 0.15% in u, 2.47% in v). The mean has large uncertainty and is by fare the worst estimation. As expected the optimal intersection point gives the best result (more than three times better).

2.2.1 The error between a straight line and a line segment

Let the noise in $\mathbf{s}_{i1}$ and $\mathbf{s}_{i2}$ be Gaussian with zero mean and covariance matrix Λ and let $\mathbf{s}_{i1}$ and $\mathbf{s}_{i2}$ be independent from each other. Suppose an arbitrary straight line $\mathbf{l}_i = (l_{i1}\ l_{i2}\ l_{i3})^\top$ and let the normal projections of $\mathbf{s}_{i1}$ and $\mathbf{s}_{i2}$ onto $\mathbf{l}_i$ be $\hat{\mathbf{s}}_{i1} = (\hat{s}_{i11}\ \hat{s}_{i12})^\top$ and $\hat{\mathbf{s}}_{i2} = (\hat{s}_{i21}\ \hat{s}_{i22})^\top$

respectively, which is given by a projection

$$\mathrm{f}_p(\mathbf{l}_i, \mathbf{s}_i) = \begin{pmatrix} \hat{\mathbf{s}}_{i1} \\ \hat{\mathbf{s}}_{i2} \end{pmatrix} \tag{2.2}$$

with

$$\mathrm{f}_p(\mathbf{l}_i, \mathbf{s}_i) = \begin{pmatrix} \mathbf{s}_{i1} \\ \mathbf{s}_{i2} \end{pmatrix} - \begin{pmatrix} \bar{\mathbf{l}}_i^\top [\mathbf{s}_{i1}]_\mathrm{h} \bar{\mathbf{l}}_i \\ \bar{\mathbf{l}}_i^\top [\mathbf{s}_{i2}]_\mathrm{h} \bar{\mathbf{l}}_i \end{pmatrix} \tag{2.3}$$

and

$$\bar{\mathbf{l}}_i = \frac{1}{\sqrt{l_{i1}^2 + l_{i2}^2}} (l_{i1}\ l_{i2})^\top \tag{2.4}$$

the unit vector perpendicular to $\mathbf{l}_i$*.

One way to express the likelihood that a line segment supports a straight line or in other words that a straight line explains a line segment is

$$\begin{aligned} \mathrm{p}(\mathbf{l}_i|\mathbf{s}_i) &= \mathrm{p}(\begin{pmatrix} \hat{\mathbf{s}}_{i1} \\ \hat{\mathbf{s}}_{i2} \end{pmatrix} | \begin{pmatrix} \mathbf{s}_{i1} \\ \mathbf{s}_{i2} \end{pmatrix}) & (2.5) \\ &= \mathrm{p}(\hat{\mathbf{s}}_{i1}|\mathbf{s}_{i1})\mathrm{p}(\hat{\mathbf{s}}_{i2}|\mathbf{s}_{i2}). & (2.6) \end{aligned}$$

$\mathrm{p}(\hat{\mathbf{s}}_{i1}|\mathbf{s}_{i1})$ and $\mathrm{p}(\hat{\mathbf{s}}_{i2}|\mathbf{s}_{i2})$ are Gaussian probability density functions of each endpoint, thus, $\mathrm{p}(\mathbf{l}_i|\mathbf{s}_i)$ can be written as

$$\mathrm{p}(\mathbf{l}_i|\mathbf{s}_i) = \frac{1}{2\pi|\Lambda|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\mathrm{d}^2(\hat{\mathbf{s}}_{i1}, \mathbf{s}_{i1}) + \mathrm{d}^2(\hat{\mathbf{s}}_{i2}, \mathbf{s}_{i2}))}. \tag{2.7}$$

It can be seen that the error between a straight line $\mathbf{l}_i$ and a line segment $\mathbf{s}_i$ is $\mathrm{d}^2(\hat{\mathbf{s}}_{i1}, \mathbf{s}_{i1}) + \mathrm{d}^2(\hat{\mathbf{s}}_{i2}, \mathbf{s}_{i2})$ (Eqn. 2.7). The distance $\mathrm{d}^2(\cdot)$ between the endpoints and their projections is the Mahalanobis distance and is written as $\mathrm{d}^2(\hat{\mathbf{s}}_{i1}, \mathbf{s}_{i1}) = (\mathbf{s}_{i1} - \hat{\mathbf{s}}_{i1})^\top \Lambda^{-1} (\mathbf{s}_{i1} - \hat{\mathbf{s}}_{i1})$ and $\mathrm{d}^2(\hat{\mathbf{s}}_{i2}, \mathbf{s}_{i2}) = (\mathbf{s}_{i2} - \hat{\mathbf{s}}_{i2})^\top \Lambda^{-1} (\mathbf{s}_{i2} - \hat{\mathbf{s}}_{i2})$ respectively. Liebowitz [2001] assumed without severe restrictions an isotropic distribution with $\Lambda = \sigma I_2$; σ is the noise level of the image and I_2 is the 2×2 identity matrix. Kanazawa and Kanatani [2001] showed that this assumption is valid along image gradients with large magnitude which is the case at edges and corners.

2.2.2 The error in the intersection of concurrent line segments

We are not interested in the error between a line segment and an arbitrary line, however, we are interested in the error between a line segment and a straight line that fits the line segment best in a statistical sense and passes exactly through the potential intersection $\mathbf{u} = (u_1\ u_2\ u_3)^\dagger$. In other words, given a particular $\mathbf{s}_i$ we are looking for a straight line $\mathbf{l}_i$

*The operator $[\cdot]_\mathrm{h}$ homogenizes a vector, that is, when $\mathbf{a} = (a_1\ a_2)^\top$ then $[\mathbf{a}]_\mathrm{h} = (a_1\ a_2\ 1)^\top$.

†Note the notation with a homogeneous vector.

that maximizes the likelihood $\mathrm{p}(\mathbf{l}_i|\mathbf{s}_i)$ and satisfies $\mathbf{l}_i^\top \mathbf{u} = 0$.

This is the case when the distance between the endpoints and their normal projections is a minimum, i.e.

$$\mathrm{d}^2(\hat{\mathbf{s}}_{i1}, \mathbf{s}_{i1}) + \mathrm{d}^2(\hat{\mathbf{s}}_{i2}, \mathbf{s}_{i2}) \to 0 \tag{2.8}$$

The intersection $\mathbf{u}$ defines concurrent lines and exactly one line possesses this smallest error to a particular line segment. Liebowitz [2001] showed that this interrelation is expressed linearly by the function

$$\mathrm{f}_u(\mathbf{u}, \mathbf{s}_i) = \mathbf{l}_i \tag{2.9}$$

with

$$\begin{aligned} \mathrm{f}_u(\mathbf{u}, \mathbf{s}_i) &= [\mathbf{u}]_\times \mathbf{m} && (2.10)\\ &= [\mathbf{u}]_\times [\mathbf{s}_i]_\mathrm{h} \mathbf{v} && (2.11)\\ &= [\mathbf{u}]_\times \left([\mathbf{s}_{i1}]_\mathrm{h}\ [\mathbf{s}_{i2}]_\mathrm{h}\right) \mathbf{v}, && (2.12) \end{aligned}$$

where $\mathbf{m} = (m_1\ m_2\ m_3)$ is not necessarily the midpoint of the line segment. Remember that Rother [2003] explicitly assumed $\mathbf{m}$ to be the midpoint which is now obviously invalid. $\mathbf{m}$ is computed using $\mathbf{v}$ the eigenvector with respect to the larger eigenvalue of the 2×2 matrix

$$A = \frac{1}{2[\mathbf{u}^\top([\mathbf{s}_{i1}]_\mathrm{h} \times [\mathbf{s}_{i2}]_\mathrm{h})]^2} \left([\mathbf{s}_{i1}]_\mathrm{h}\ [\mathbf{s}_{i2}]_\mathrm{h}\right)^\top [\mathbf{u}]_\times^\top \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} [\mathbf{u}]_\times \left([\mathbf{s}_{i1}]_\mathrm{h}\ [\mathbf{s}_{i2}]_\mathrm{h}\right). \tag{2.13}$$

Details are in [Liebowitz, 2001]. Note that the eigenvalue decomposition of A has a closed-form solution and the evaluation of $\mathrm{f}_u(\cdot)$ (Eqn. 2.9) is straight forward [Strang, 2003]. This is important, because $\mathrm{f}_u(\cdot)$ is evaluated for hundreds of line segments and therefore the evaluation has to be efficient.

2.2.3 The optimal estimation

Following the idea to compute straight lines passing exactly through $\mathbf{u}$ and elongating optimally the line segments $\mathbf{s}_1, \ldots, \mathbf{s}_N$ in a statistical sense, the likelihood that line segments support the concurrent lines is expressed by

$$\mathrm{p}(\mathbf{u}|\mathbf{s}_1, \ldots, \mathbf{s}_N) = \prod_{i=1}^{N} \mathrm{p}(\mathrm{f}_u(\mathbf{u}, \mathbf{s}_i)|\mathbf{s}_i). \tag{2.14}$$

A maximum likelihood estimation of $\mathbf{u}$ with respect to the Gaussian error model maximizes the natural log-likelihood $\log(\mathrm{p}(\mathbf{u}|\mathbf{s}_1, \ldots, \mathbf{s}_N)$. If we substitute $\log(\mathrm{p}(\mathbf{u}|\mathbf{s}_1, \ldots, \mathbf{s}_N)$ by apply-

ing Eqn. 2.7 - 2.14 in reverse order we get

$$\begin{aligned}
\log(\mathrm{p}(\mathbf{u}|\mathbf{s}_1,\ldots,\mathbf{s}_N)) &= \log(\prod_{i=1}^{N}\mathrm{p}(\mathrm{f}_u(\mathbf{u},\mathbf{s}_i)|\mathbf{s}_i)) && (2.15)\\
&= \sum_{i=1}^{N}\log(\mathrm{p}(\mathrm{f}_u(\mathbf{u},\mathbf{s}_i)|\mathbf{s}_i)) && (2.16)\\
&= \sum_{i=1}^{N}\log(\mathrm{p}(\mathbf{l}_i|\mathbf{s}_i)) && (2.17)\\
&= -\frac{1}{4\pi|\Lambda|^{1/2}}\sum_{i=1}^{N}\mathrm{d}^2(\hat{\mathbf{s}}_{i1},\mathbf{s}_{i1})+\mathrm{d}^2(\hat{\mathbf{s}}_{i2},\mathbf{s}_{i2}) && (2.18)
\end{aligned}$$

Maximizing the log-likelihood (Eqn. 2.18) is equivalent to minimizing the geometric, non-linear, least-squares objective function

$$\mathrm{f}_d(\mathbf{u},\mathbf{s}_1,\ldots,\mathbf{s}_N)=\sum_{i=1}^{N}\mathrm{d}^2(\mathrm{f}_u(\mathbf{u},\mathbf{s}_{i1}),\mathbf{s}_{i1})+\mathrm{d}^2(\mathrm{f}_u(\mathbf{u},\mathbf{s}_{i2}),\mathbf{s}_{i2}). \quad (2.19)$$

$\mathrm{f}_d(\cdot)$ is the total error between all line segments and concurrent lines given $\mathbf{u}$. It enjoys all advantages discussed in Sec. 2.1.1.3. In practice, the minimization of $\mathrm{f}_d(\cdot)$ is implemented by an iterative, numerical method such as the popular Levenberg-Marquardt (LM) algorithm [Hartley and Zisserman, 2004; Lourakis and Argyros, 2005].

2.3 The robust intersection of line segments

Line segments that do not belong to a particular set of concurrent line segments are gross outliers. Such outliers cause severe errors in the intersection estimation and should be identified. The RANSAC approach is a very popular solution for the detection of gross outliers.

RANSAC chooses randomly without replacement two line segments from the set of all line segments and computes their intersection point. Two line segments is the minimal number that is needed to compute an intersection. Then, all inliers are identified by measuring the error between each line segment and the given intersection (Sec. 2.2.2). Following the RANSAC approach, a line segment $\mathbf{s}_i$ will be an inlier and thus part of the consensus set, if both Euclidean distances $\|\mathbf{s}_{i1}-\hat{\mathbf{s}}_{i1}\|$ and $\|\mathbf{s}_{i2}-\hat{\mathbf{s}}_{i2}\|$ are smaller than a threshold t, that is, $\|\mathbf{s}_{i1}-\hat{\mathbf{s}}_{i1}\|<t$ and $\|\mathbf{s}_{i2}-\hat{\mathbf{s}}_{i2}\|<t$ are satisfied. We set $t=\chi\sigma$ with χ be the α-quartile of a χ^2-distribution. The number of samples n and the proportion w of inliers to the total number of line segments N is determined adaptively [Hartley and Zisserman, 2004]. An early stopping criterion as it is described by Hartley and Zisserman is not used, because we simply do not know the number of inliers. However, the adaptive probing of the data via the consensus set

includes a statistically correct termination of the algorithm and works very well in practice when α is chosen conservatively, for example, $\alpha = .99$.

A remark should be given to the pixel lengths of line segments. The pixel length seems to be a valuable information to increase robustness [Rother, 2002b; Van den Heuvel, 1998]. On the one hand longer line segments are less uncertain in their position and orientation after the detection process and on the other hand longer line segments pass more reliable through a vanishing point. The former assumption can be validated by simple error propagation [Clarke, 1998]. However, we criticize the latter argument, because the length of line segments is completely independent from the orientation of the back-projected world line. Instead, the length of a line segment depends on the imaging process and on the scene structure. Rother [2002b] used the pixel length as weight during the voting stage, hence, he introduced in this way a bias in the intersection estimation. To avoid this bias, we propose that the pixel length should influence the sampling stage of our RANSAC method. Usually, the two line segments are sampled uniformly from the interval $[1, \ldots, N]$ by the adaptive algorithm. We sample from the same interval, but following a sample distribution given by the histogram of the pixel length. As we assume for a line segment no dependence between being an inlier and its pixel length, the standard adaptive algorithm is still applicable.

2.4 The robust grouping of line segments

A simple and self-evident idea to group concurrent line segments is the consecutive estimation of intersection points which was discussed in the last section. At the beginning, all line segments are used to estimate an intersection point. After the estimation, the consensus set is removed from the set of line segments. The set of outlier line segments is used in a consecutive step to estimate a further intersection point. This process can be repeated until all line segments are grouped into concurrent line segments or intersection points of a particular number M are identified.

To avoid equivalent groups of line segments a statistical test evaluates the pairwise coincidence between all intersection points. This is possible, because uncertainties are propagated in each computation. If an intersection point lies with a specific confidence within the uncertainty ellipse the intersection point will be rejected. Alg. 2.1 summarizes all necessary computations. Fig. 2.5 shows for a better understanding the grouping result in our synthetic world of Fig. 2.3a after the identification of three intersection points.

Threshold t during the RANSAC execution controls the grouping. If t is too small then the grouping for a particular intersection point will mistakenly exclude true line segments. Otherwise, if t is too large then the grouping for a particular intersection point will include some wrong line segments. However, Sec. 2.3 showed a clear interpretation of t from a

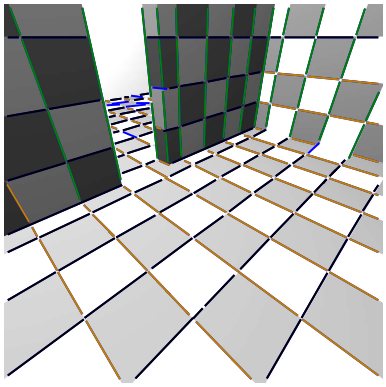


Figure 2.5: The intersection points of three ($M = 3$) concurrent line segments (black, brown, green) are identified. The remaining outlier line segments are drawn in blue. σ was set to 1 pixel. Obviously t is too large, because some line segments belong to the wrong concurrent lines or are spuriously identified as noise.

Algorithm 2.1 Group line segments to fit intersection points.

Let $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$ be N detected line segments in an image. Let χ be the α-quartile of a χ^2-distribution with typically $\alpha \geq 0.95$. Return M intersection points $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_M\}$.

1. Fit an intersection point $\mathbf{u}_i$ to $\mathcal{S}$ with RANSAC (Sec. 2.3), let the outlier line segments be the set $\mathcal{O}$ and the inliers be $\mathcal{I}$ with $\mathcal{S} = \mathcal{O} \cup \mathcal{I}$,
2. Fit $\mathbf{u}_i$ optimally to $\mathcal{I}$ (Sec. 2.2.3).
3. $\mathcal{S} \leftarrow \mathcal{O}$,
4. Test the coincidence of $\mathbf{u}_i$ to previously fitted intersection points $\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}$: If the Mahalanobis distance $(\mathbf{u}_i - \mathbf{u}_j)^\top \Lambda_j^{-1} (\mathbf{u}_i - \mathbf{u}_j)$ with $1 \leq j \leq i-1$ is smaller than χ then $\mathbf{u}_i$ will be rejected, otherwise $\mathcal{U} \leftarrow \mathcal{U} \cup \{\mathbf{u}_i\}$.
5. Go back to step 1 until M intersection points are fitted.

statistical point of view.

2.5 Robust calibration

The geometric relationship between on the one hand orthogonal vanishing points and the IAC [Hartley and Zisserman, 2004] and on the other hand the IAC and the interior orientation is the basis of our calibration method. The IAC is the image of the Absolute Conic [Semple and Kneebone, 1998], a conic which is present in every image and which is invariant to rotation and translation in the Euclidian plane. It is invariant for cameras with constant interior orientation. We write in the following

$$\omega = \begin{bmatrix} \omega_1 & \omega_2 & \omega_4 \\ \omega_2 & \omega_3 & \omega_5 \\ \omega_4 & \omega_5 & \omega_6 \end{bmatrix} \tag{2.20}$$

Algorithm 2.2 Method C-0: Calibrate a single camera.

Let $\mathbf{v}_1$, $\mathbf{v}_2$ and optionally $\mathbf{v}_3$ be orthogonal vanishing points computed by Alg. 2.1. Let $\mathbf{c}$ be the image center and $\Lambda_\mathbf{c}$ be the covariance matrix representing the uncertainty when assuming $\mathbf{p} = \mathbf{c}$ or $\mathbf{p}$ close to $\mathbf{c}$. Let t_∞ be a threshold on conditional numbers and t_ϕ be an angular threshold. Let χ be the α-quartile of a χ^2-distribution with typically $\alpha \geq 0.95$. Return the interior and exterior orientation matrix K and R respectively when the calibration is successful.

1. Test $\mathbf{v}_1$,$\mathbf{v}_2$ and when available $\mathbf{v}_3$ to be infinite points.
   Compute the condition number of the covariance matrix for each vanishing point. If two of the the condition numbers is larger than t_∞ terminate unsuccessfully.
2. Test when possible the angles of the triangle formed by finite $\mathbf{v}_1$,$\mathbf{v}_2$ and $\mathbf{v}_3$.
   Compute $\phi_1 = \angle(\mathbf{v}_1 \times \mathbf{v}_2, \mathbf{v}_1 \times \mathbf{v}_3)$, $\phi_2 = \angle(\mathbf{v}_1 \times \mathbf{v}_2, \mathbf{v}_2 \times \mathbf{v}_3)$ and $\phi_3 = \angle(\mathbf{v}_2 \times \mathbf{v}_3, \mathbf{v}_1 \times \mathbf{v}_3)$. If one of the inequalities $\frac{\pi}{2} - \phi_i < t_\phi$ for $i = 1, \ldots, 3$ is not fulfilled terminate unsuccessfully.
3. Test when $\mathbf{v}_1$, $\mathbf{v}_2$ are finite and $\mathbf{v}_3$ is infinite the orthogonality $(\mathbf{v}_1 \times \mathbf{v}_2) \perp \mathbf{v}_3$.
   Compute $\phi = \angle(\mathbf{v}_1 \times \mathbf{v}_2, \mathbf{v}_3)$. When $\frac{\pi}{2} - \phi < t_\phi$ is not fulfilled terminate unsuccessfully.
4. Constrain the principal point $\mathbf{p}$ when necessary (Sec. 2.5.4).
   - When only two finite $\mathbf{v}_1$, $\mathbf{v}_2$ are available, set $\mathbf{p}$ to $\mathbf{c}$.
   - When only one finite $\mathbf{v}_1$ and two infinite $\mathbf{v}_2$, $\mathbf{v}_3$ are available, set $\mathbf{p}$ to $\mathbf{c}$.
   - When $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ are available but only $\mathbf{v}_1$, $\mathbf{v}_2$ are finite, set $\mathbf{p}$ to a point with $\mathbf{p}^\top(\mathbf{v}_1 \times \mathbf{v}_2) = 0$ and where $\|\mathbf{p} - \mathbf{c}\| \to 0$.
   - When $\mathbf{v}_1$ is finite and $\mathbf{v}_2$ infinite, set $\mathbf{p}$ to a point with $\mathbf{p}^\top\mathbf{l} = \mathbf{v}_1^\top\mathbf{l} = 0$. The straight line $\mathbf{l}$ is further constrained by $\angle(\mathbf{l}, \mathbf{v}_3) = \frac{\pi}{2}$.
5. Calibrate the interior orientation $K \triangleq \{f, r, \mathbf{p}\}$ with $\mathbf{v}_1$, $\mathbf{v}_2$ and when available $\mathbf{v}_3$ (Sec. 2.5.5). Assume a natural camera, that is, $r = 1$, $s = 0$. Assume optionally the location of $\mathbf{p}$ given by the previous step 4.
6. Test sufficient conditions on the interior orientation.
   When some of the inequalities $f < 0$, $r < 0$ and $(\mathbf{p} - \mathbf{c})^\top \Lambda_\mathbf{c}^{-1} (\mathbf{p} - \mathbf{c}) > \chi^2$ are fulfilled terminate unsuccessfully.
7. Compute the third vanishing point $\mathbf{v}_3$ when necessary.
   Compute the straight line $\mathbf{l} = \mathbf{v}_1 \times \mathbf{v}_2$. $\mathbf{v}_3$ is then given by $\mathbf{v}_3 = KK^\top\mathbf{l}$.
8. Compute the exterior orientation R (Sec. 2.5.6).
   Construct $V = [\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3]$. Compute the directions $D = [\mathbf{d}_1\ \mathbf{d}_2\ \mathbf{d}_3] = K^{-1}V$. Normalize D to obtain the rotation matrix $R = [\mathbf{r}_1\ \mathbf{r}_2\ \mathbf{r}_3] = \left[\frac{\mathbf{d}_1}{\|\mathbf{d}_1\|}\ \frac{\mathbf{d}_2}{\|\mathbf{d}_2\|}\ \frac{\mathbf{d}_3}{\|\mathbf{d}_3\|}\right]$. When $\det(R) < 0$ set $\mathbf{r}_3 = -\mathbf{r}_3$.

for the IAC. As we can see, the IAC is a symmetric matrix. It is directly related to the interior orientation by scene and camera constraints, for example, the natural camera constraint or orthogonal vanishing points which form a linear equation system [Liebowitz and Zisserman, 1999] that is efficiently solved by SVD. Once the interior orientation and the three pairwise orthogonal vanishing points are known, the exterior orientation can be derived. A motivation for this approach with respect to other work was given in Sec. 2.1.3. The reader is asked to comprehend the method we call C-0 by Alg. 2.2. In the following the section discusses some details of the method.

2.5.1 Infinite intersection points

A simple test that locates an intersection point at infinity uses its propagated uncertainty. If the intersection point is infinite the supporting line segments are parallel or close to parallel in the image plane. The consequence is that the relation between the largest and the smallest eigenvalue of the covariance matrix is becoming infinite large, that is, the covariance matrix is singular for infinite vanishing points. In our empirical studies we found out that when this relation or in other words the condition number of the covariance matrix exceeds a value of 10^{10} the vanishing point has to be seen infinite*.

A short remark: On a first view a simple way to identify infinite points is to test the homogeneous coordinate to be zero. This is per definition a sufficient condition for infinite points. However, points on the projective plane are equal up to a scalar factor, that is, $(u\ v\ w)^\top \approx \lambda\,(u\ v\ w)^\top$. Hence, $\lambda \in \mathbb{R}$ makes w arbitrarily large and does not allow to set any threshold.

2.5.2 Identify orthogonal vanishing points

The next step in our calibration method is to identify vanishing points and then pairwise orthogonal vanishing points $\mathbf{v}_i = (v_{i1}\ v_{i2}\ v_{i3})^\top$ and $\mathbf{v}_j = (v_{j1}\ v_{j2}\ v_{j3})^\top$. Basically, each intersection point is a potential vanishing point. Unfortunately:

Result 2.1. *No test exists that decides, if two vanishing points are orthogonal, when the interior orientation is unknown.*

This result follows directly by the constraint on the IAC of two orthogonal vanishing points i and j [Liebowitz and Zisserman, 1999], that is,

$$\mathbf{v}_i^\top \omega \mathbf{v}_j = 0 \tag{2.21}$$

and by the direct relation between IAC and interior orientation

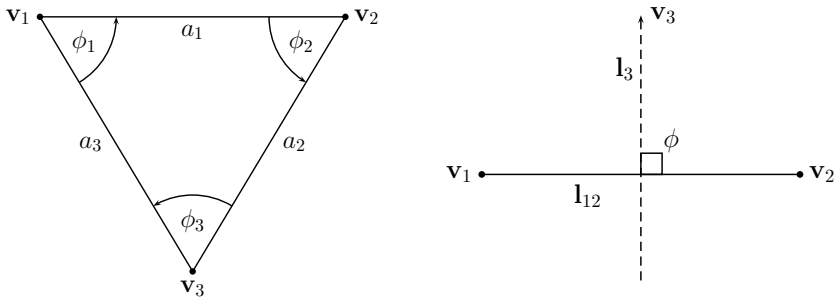
$$K = \begin{bmatrix} f & 0 & p_1 \\ 0 & rf & p_2 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.22}$$

which is

$$\omega^{-1} = KK^\top. \tag{2.23}$$

$\mathbf{p} = (p_1\ p_2)^\top$ is the principal point, f the focal length and r the aspect ratio. The IAC

*The rule of thumb in numerical mathematics is $\log \Lambda > p$ where Λ is the matrix and p the precision of matrix entries. For example, $p = 10$ means a loss of 10 digits in the results after a computation with this matrix.



(a) Ortho-triangle (b) Central line

Figure 2.6: Necessary conditions on vanishing points: (a) the ortho-triangle has $\phi_1 < \frac{\pi}{2}$, $\phi_2 < \frac{\pi}{2}$ and $\phi_3 < \frac{\pi}{2}$, (b) square pixel cameras fulfill $\mathbf{l}_3 \perp \mathbf{l}_{12}$ ($\phi = \frac{\pi}{2}$), whereas $\mathbf{l}_3$ is the central line with $\mathbf{l}_3^\top \mathbf{v}_3 = 0$ (dotted) and $\mathbf{l}_{12} = \mathbf{v}_1 \times \mathbf{v}_2$ is a vanishing line.

encodes the metric space, that is, the knowledge about angles and hence orthogonality. The dilemma of this result is that we cannot find any test that decides safely about orthogonality.

C-0 uses Alg. 2.1 to fetch at least two but ideally three intersection points. As RANSAC's nature is to find a model with the most support, our approach ensures that the number of line segments found in each step is monotonically decreasing. In a man-made world one can assume that the intersection points supported by the most line segments are the orthogonal vanishing points.

After the calibration we test the focal length and the principal point for plausible values. Plausibility tests were suggested by Kanatani [1996]. The focal length is always a positive value. As long as the image is not cropped, the principal point is usually close to the image center. Nevertheless, these tests are not sufficient to know that the interior calibration is correct, because nothing is known about the true focal length.

2.5.3 Necessary vanishing point conditions

Although no sufficient test for a vanishing point exists, two necessary conditions on the vanishing points can be tested. Imagine three intersection points $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ that are potential vanishing points. If $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ are vanishing points then they have to form an ortho-triangle [Caprile and Torre, 1990]. Fig. 2.6a illustrates the ortho-triangle of the image in Fig. 2.5.

Van den Heuvel [1998] recognized a fact he called "orthogonality criterion", that is, each angle between two sides of the ortho-triangle has to be smaller than $\frac{\pi}{2}$. The reader can convince himself that the orthogonality criterion is not sufficient, because one can easily find an example of two vanishing points including an angle smaller than $\frac{\pi}{2}$ and describing directions in the world that include an arbitrary angle.

How can we compute these angles? Let a_1 be the Euclidean distance between $[\mathbf{v}_1]_\mathrm{a}$* and $[\mathbf{v}_2]_\mathrm{a}$ in the image plane, a_2 between $[\mathbf{v}_2]_\mathrm{a}$ and $[\mathbf{v}_3]_\mathrm{a}$ and a_3 between $[\mathbf{v}_3]_\mathrm{a}$ and $[\mathbf{v}_1]_\mathrm{a}$. Let ϕ_1 be the included angle $\phi_1 = \angle(a_1, a_3)$, ϕ_2 be the included angle $\angle(a_1, a_2)$ and ϕ_3 be the included angle $\angle(a_2, a_3)$. The law of cosines tells us then

$$\phi_1 = \cos^{-1} \frac{a_1^2 + a_3^2 - a_2^2}{2a_1a_3} \tag{2.24}$$

$$\phi_2 = \cos^{-1} \frac{a_1^2 + a_2^2 - a_3^2}{2a_1a_2} \tag{2.25}$$

$$\phi_3 = \cos^{-1} \frac{a_2^2 + a_3^2 - a_1^2}{2a_2a_3}. \tag{2.26}$$

The orthogonality criterion is tested by checking $\phi_1 < \frac{\pi}{2}$, $\phi_2 < \frac{\pi}{2}$ and $\phi_3 < \frac{\pi}{2}$.

Another necessary vanishing point condition is illustrated in Fig. 2.6b. Consider two finite vanishing points $\mathbf{v}_1$ and $\mathbf{v}_2$ and one infinite vanishing point $\mathbf{v}_3$. Let $\mathbf{l}_{12}$ be the straight line $\mathbf{l}_{12} = \mathbf{v}_1 \times \mathbf{v}_2$ passing through $\mathbf{v}_1$ and $\mathbf{v}_2$. Let $\mathbf{l}_3$ be the central line [Gurdjos and Payrissat, 2000] meeting $\mathbf{v}_3$. We can observe the following result:

Result 2.2. *The angle ϕ between $\mathbf{l}_{12}$ and $\mathbf{l}_3$ is $\frac{\pi}{2}$, if and only if we assume a square pixel camera, that is, aspect ratio $r = 1$.*

This result was proposed by Gurdjos and Payrissat [2000] and is in conflict with Rother [2002b] who argued that for arbitrary aspect ratios, $\mathbf{l}_{12}$ and $\mathbf{l}_3$ are orthogonal. Rother's argument is incorrect which shows the following proof:

Proof. Affinize $\mathbf{v}'_i = [\mathbf{v}_i]_\mathrm{a}$. $\mathbf{l}_3$ has normal vector $\mathbf{n}_3 = \left(r^2(v'_{21} - v'_{11})\ (v'_{22} - v'_{12})\right)^\top$ [Gurdjos and Payrissat, 2000]. $\mathbf{n}_3$ is orthogonal to the normal vector $\mathbf{n}_{12} = \left((v'_{12} - v'_{22})\ (v'_{21} - v'_{11})\right)^\top$ of $\mathbf{l}_{12}$. Consequently,

$$\mathbf{n}_3^\top \mathbf{n}_{12} = \left(r^2(v'_{21} - v'_{11})\ (v'_{22} - v'_{12})\right) \begin{pmatrix} v'_{12} - v'_{22} \\ v'_{21} - v'_{11} \end{pmatrix} \tag{2.27}$$

$$= [(r^2 - 1)v'_{12} + (1 - r^2)v'_{22}](v'_{21} - v'_{11}). \tag{2.28}$$

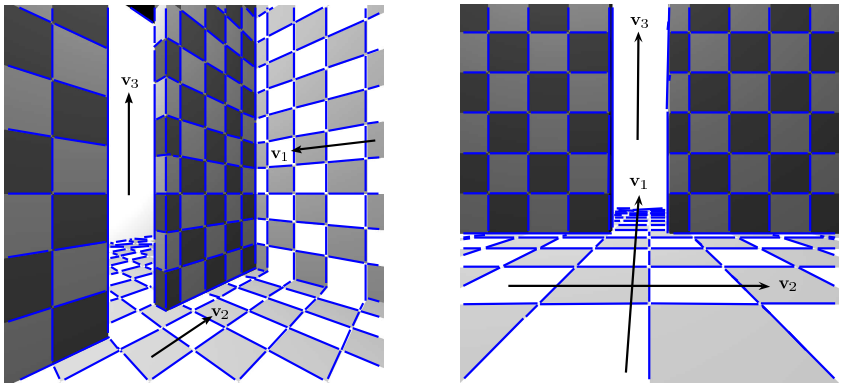
The reader can convince himself that for the non trivial case $v'_{11} \neq v'_{21}$ and $v'_{12} \neq v'_{22}$ the inner product $\mathbf{n}_3^\top \mathbf{n}_{12}$ is only for $r^2 = 1$ zero (Eqn. 2.28). □

Given this result, ϕ is then calculated by

$$\cos\phi = \mathbf{n}_3^\top \mathbf{n}_{12} \tag{2.29}$$

and can be tested to be close to $\frac{\pi}{2}$.

*The operator $[\cdot]_\mathrm{a}$ affinizes a vector, that is, when $\mathbf{a} = (a_1\ a_2\ a_3)^\top$ then $[\mathbf{a}]_\mathrm{a} = \left(\frac{a_1}{a_3}\ \frac{a_2}{a_3}\right)^\top$.



(a) An upright camera

(b) An upright camera and a parallel scene

Figure 2.7: Critical cases during camera calibration. The three orthogonal directions are illustrated. (a) vanishing point $\mathbf{v}_3$ is at infinity in the image plane, (b) vanishing point $\mathbf{v}_2$ and $\mathbf{v}_3$ are at infinity in the image plane.

2.5.4 Critical cases

We know that three pairwise orthogonal vanishing points with a finite location in the image plane provide three independent equations (Eqn. 2.21) which constrain the IAC. A natural camera provides further two equations, namely

$$(1\ 0\ 0)^\top \omega\, (0\ 1\ 0) = 0 \quad \text{(zero skew)} \tag{2.30}$$

$$(1\ r\ 0)^\top \omega\, (1\ -r\ 0) = 0 \quad \text{(known aspect ratio).} \tag{2.31}$$

Eqn. 2.21, 2.30 and 2.31 determine the remaining principal point and the focal length of the interior orientation uniquely, because the IAC has five degrees of freedom [Liebowitz, 2001]. In case of only two orthogonal vanishing points with a finite location in the image plane, the number of equations is three, hence, an ambiguity of the IAC is remaining. Assuming that the principal point is the image center resolves this ambiguity.

Liebowitz [2001] and Rother [2003] analyzed the critical cases that happen if one or more vanishing points are near or at infinity in the image plane. Such cases occur frequently in practice. The consequence of a critical case is an ambiguity in the interior orientation that prevents a successful calibration. In the following we will summarize their discussion:

Three vanishing points, two are finite: This critical case is shown in Fig. 2.7a. The camera is held upright so that the image plane is vertical to the ground plane. The two equations involving the infinite vanishing point $\mathbf{v}_3$ are not independent. Hence, the IAC is only constrained by four independent equations. Geometrically, the principle point can lie somewhere on the straight line passing through the two finite vanishing points $\mathbf{v}_1$ and $\mathbf{v}_2$. By choosing the principal point $\mathbf{p}$ as the point on the straight line with

minimal normal Euclidean distance to the image center $\mathbf{c} = (c_1\ c_2)^\top$, this ambiguity can be resolved and the focal length is uniquely determined. Fig. 2.8a shows this projection of $\mathbf{c}$ onto $\mathbf{l} = \mathbf{v}_1 \times \mathbf{v}_2$. More precisely,

$$\mathbf{p} = \mathbf{c} - \frac{1}{\sqrt{l_1^2 + l_2^2}} \mathbf{l}^\top [\mathbf{c}]_\mathrm{h} \begin{pmatrix} l_1 \\ l_2 \end{pmatrix}. \tag{2.32}$$

Three vanishing points, one is finite: This critical case is shown in Fig. 2.7b. The camera is held in an upright position and the wall in front of the camera is parallel to the image plane. Unfortunately, no calibration is possible, thus, an automatic detection of this case is important. The only action to resolve the ambiguity in the interior orientation is to rotate the camera. The reason for this is that all equations involving the vanishing points are dependent.

Only two finite vanishing points: The principal point is chosen as the image center. Then the IAC is constrained by five independent equations and the focal length is uniquely determined.

Two vanishing points, one is finite: The principal point is constrained to lie on a straight line $\mathbf{l}$ passing through the only finite vanishing point $\mathbf{v}_1$ (Fig. 2.8b). $\mathbf{l}$ is orthogonal to the direction $\mathbf{n}_3 = (v_{31}\ v_{32})^\top$ given by the infinite vanishing point $\mathbf{v}_3$. More precisely, $\mathbf{v}_1$ is projected onto an arbitrary straight line that has as normal vector $\mathbf{v}_3$. Say the projection is point $\mathbf{u}$, then

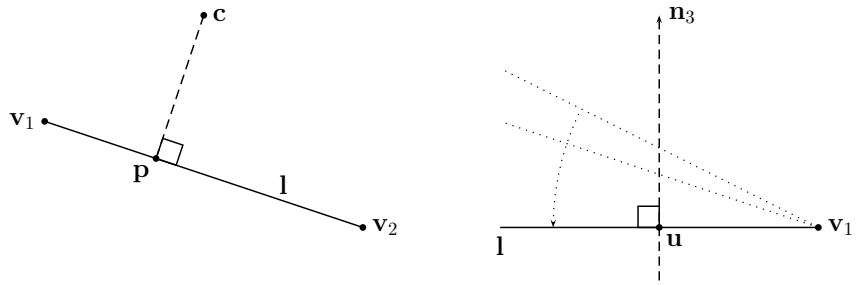
$$\mathbf{u} = \mathbf{v}_1 - \mathbf{v}_3^\top \mathbf{v}_1 \mathbf{n}_3^\top. \tag{2.33}$$

$\mathbf{l}$ is given by $\mathbf{u} \times \mathbf{v}_1$. Now, to get the principal point we perform the same projection as it is was previously shown with one infinite vanishing point (Eqn. 2.32). It is the point with the smallest Euclidean distance to the image center $\mathbf{c}$. The IAC is then fully constrained and the focal length is uniquely determined.

2.5.5 Calibrate the interior orientation

Vanishing points $\mathbf{v}_1, \ldots, \mathbf{v}_m$ give $\binom{m}{2}$ constraints on the IAC (Eqn. 2.21). The assumptions of zero skew and a constant aspect ratio yield two more constraints (Eqn. 2.30 and 2.31). A known principal point provides the last conceivable constraint, that is,

$$\omega[\mathbf{p}]_\mathrm{h} = (0\ 0\ 1)^\top. \tag{2.34}$$



(a) Projection of $\mathbf{c}$ onto $\mathbf{l}$ (b) Fit $\mathbf{l}$ that $\mathbf{l} \perp \mathbf{n}_3$

Figure 2.8: Constraints that are valid in critical cases: (a) the closest point on the finite vanishing line $\mathbf{l} = \mathbf{v}_1 \times \mathbf{v}_2$ to the image center $\mathbf{c}$ is the projection $\mathbf{p}$ (Eqn. 2.33). $\mathbf{p}$ is the principal point in case of one infinite vanishing point $\mathbf{v}_3$, (b) find $\mathbf{u}$ for which vanishing line $\mathbf{l} = \mathbf{u} \times \mathbf{v}_1$ is perpendicular to the direction $\mathbf{n}_3 = (v_{31}\ v_{32})^\top$.

Note that this constraint is only in critical cases necessary. Plugging Eqn. 2.20 into Eqn. 2.34 gives two independent equations in the vector's elements

$$p_1\omega_1 + p_2\omega_2 + \omega_4 = 0 \tag{2.35}$$

$$p_1\omega_2 + p_2\omega_3 + \omega_5 = 0. \tag{2.36}$$

Let us stack together the Eqn. 2.21, 2.30, 2.31 and 2.34 to form a homogeneous, linear equation system

$$A\mathbf{w} = 0, \tag{2.37}$$

where the unknowns are the elements of ω as a vector $\mathbf{w} = (\omega_1, \ldots, \omega_6)^\top$. The system matrix A contains rows in the same order resulting in

$$A = \begin{bmatrix} & & \mathbf{a}^\top & & & \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -r^2 & 0 & 0 & 0 \\ p_1 & p_2 & 0 & 1 & 0 & 0 \\ 0 & p_1 & p_2 & 0 & 1 & 0 \end{bmatrix}, \tag{2.38}$$

whereas $\mathbf{a} = (k_1\ (k_2 + k_4)\ k_5\ (k_3 + k_7)\ (k_6 + k_8)\ k_9)^\top$ with

$$\mathbf{k} = \mathbf{v}_i \otimes \mathbf{v}_j \tag{2.39}$$

$1 \leq i, j \leq m,\ i \neq j$*.

The IAC can be computed by SVD, that is, $\mathbf{w}$ is the null-vector of A, for $m > 1$, or if

*Operator $\otimes$ is the Kronecker product. It is a convenient operator that transforms bilinear equations to linear ones, that is, $\mathbf{a}^\top M \mathbf{b} \rightarrow (\mathbf{a}^\top \otimes \mathbf{b}^\top)\mathrm{vec}\,(M) = 0$, where $\mathrm{vec}\,(\cdot)$ is a vectorization of M [Heuel, 2004].

the principal point is not known for $m > 2$; for these conditions A has a rank of five. The computation of the interior orientation knowing ω is now straightforward by using Eqn. 2.23. K can be computed by a Cholesky decomposition of ω^{-1}.

A final remark should be given: A solution might be incorrect, because information about the orthogonality of world directions is lost during the imaging process. A simple way to test the plausibility of a solution of K is to test the relative error between f and the focal length of the lens f_0. If the relative error is smaller than a maximal error ϵ, typically 50 pixel,

$$\frac{|f - f_0|}{f_0} < \epsilon, \tag{2.40}$$

the solution of K will be accepted, otherwise no successful calibration is possible. By constraining $\mathbf{p}$ the only unknown is f, hence, a plausibility test makes sense. Note that the condition ω is positive definite and the case that a Cholesky decomposition is possible is necessary but not sufficient.

2.5.6 Calibrate the exterior orientation

The interior orientation and three orthogonal vanishing points encode the exterior orientation. K and $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ are related by

$$[\alpha\mathbf{v}_1 \ \beta\mathbf{v}_2 \ \gamma\mathbf{v}_3] = KR. \tag{2.41}$$

Although the coefficients α, β and γ are unknown, R can be computed using the fact that R is an orthonormal matrix. Consequently, each column of matrix D with

$$D = [\mathbf{d}_1 \ \mathbf{d}_2 \ \mathbf{d}_3] = K^{-1} [\mathbf{v}_1 \ \mathbf{v}_2 \ \mathbf{v}_3] . \tag{2.42}$$

is normalized to yield

$$R = [\mathbf{r}_1 \ \mathbf{r}_2 \ \mathbf{r}_3] = \left[\frac{\mathbf{d}_1}{\|\mathbf{d}_1\|} \ \frac{\mathbf{d}_2}{\|\mathbf{d}_2\|} \ \frac{\mathbf{d}_3}{\|\mathbf{d}_3\|} \right] \tag{2.43}$$

$\mathbf{r}_1$, $\mathbf{r}_2$ and $\mathbf{r}_3$ are normal vectors and they are canonical directions in the camera coordinate system. As it is usual to assume a left-handed coordinate system, the determinate of R is $+1$. If this is not the case, that is, it is -1, we simply flip the sign of one of the normal vectors; we define to flip the sign of $\mathbf{r}_3$ to $-\mathbf{r}_3$.

If only two vanishing points are available and the interior orientation is known, the third orthogonal vanishing point will be

$$\mathbf{v}_3 = KK^\top \mathbf{l} \tag{2.44}$$

with vanishing line $\mathbf{l} = \mathbf{v}_1 \times \mathbf{v}_2$.

2.6 Robust calibration with known interior orientation

This section treats the method C-1 that assumes the interior orientation to be known (Alg. 2.3). Most detection methods of vanishing points, for example the one that are using the Gaussian sphere, but also calibration methods like the one by Antone and Teller [2000] treat this case. Consider M intersection points $\mathbf{u}_1, \ldots, \mathbf{u}_M$. In principle, all line segments in an image could be grouped together to different concurrent lines. Knowing the interior orientation gives the necessary information to decide about pairwise, orthogonal vanishing points $(\mathbf{u}_i\ \mathbf{u}_j)$ with $i \neq j$ and $1 \leq i, j \leq M$ or triplets $(\mathbf{u}_i\ \mathbf{u}_j\ \mathbf{u}_k)$ with $i \neq j \neq k$ and $1 \leq i, j, k \leq M$. If two or three vanishing points are present, depends on the scene. Note that the important assumption for C-0 that most line segments intersect in three orthogonal vanishing points is no longer needed by C-1. The intersection points are back-projected to directions in the camera coordinate system, that is,

$$\mathbf{d}_{i|j|k} = K^{-1}\mathbf{u}_{i|j|k}. \tag{2.45}$$

One can decide about orthogonality by testing pairwise with

$$|\frac{\pi}{2} - \angle(\mathbf{d}_i, \mathbf{d}_j)| \quad < \quad t_\phi \tag{2.46}$$

$$|\frac{\pi}{2} - \angle(\mathbf{d}_i, \mathbf{d}_k)| \quad < \quad t_\phi \tag{2.47}$$

$$|\frac{\pi}{2} - \angle(\mathbf{d}_j, \mathbf{d}_k)| \quad < \quad t_\phi \tag{2.48}$$

against an angular threshold t_ϕ. Typically, t_ϕ is smaller than 3 deg.

When C-1 finds only two orthogonal directions $\mathbf{d}_1$ and $\mathbf{d}_2$, the third orthogonal one is simply computed by

$$\mathbf{d}_3 = \mathbf{d}_1 \times \mathbf{d}_2. \tag{2.49}$$

The rotation matrix is constructed as it is shown in Eqn. 2.43, however, $\det(R) = 1$ is not exactly fulfilled, because C-1 tests the orthogonality against a threshold and a small error remains. Thus, C-1 projects the invalid rotation R into the space of valid rotations. The closest valid rotation $\hat{R}$ which fulfills $\min_R \|R - \hat{R}\|_F \rightarrow 0$* is the solution. SVD of $R = USV^\top$ does the job, resulting in

$$\hat{R} = U \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \det\left(UV^\top\right) \end{bmatrix} V^\top. \tag{2.50}$$

*$\|\cdot\|_F$ is the Frobenius norm.

Algorithm 2.3 Method C-1: Calibrate a single camera with known interior orientation.

Detect M intersection points $\mathbf{u}_1, \ldots, \mathbf{u}_M$ by Alg. 2.1. Let K be the known interior orientation and let t_ϕ be an angular threshold. Return the exterior orientation R or terminate unsuccessful.

1. Back-project all intersection points to $\mathbf{d}_i = K^{-1}\mathbf{u}_i$ for $i = 1, \ldots, M$.
2. Normalize each direction by $\bar{\mathbf{d}}_i = \frac{\mathbf{d}_i}{\|\mathbf{d}_i\|}$ for $i = 1, \ldots, N$.
3. When three orthogonal vanishing points are present, test triplewise the orthogonality of the directions, that is, $\|\frac{\pi}{2} - \angle(\bar{\mathbf{d}}_i, \bar{\mathbf{d}}_j)\| < t_\phi$, $\|\frac{\pi}{2} - \angle(\bar{\mathbf{d}}_i, \bar{\mathbf{d}}_k)\| < t_\phi$ and $\|\frac{\pi}{2} - \angle(\bar{\mathbf{d}}_j, \bar{\mathbf{d}}_k)\| < t_\phi$ for $i \neq j \neq k$ and $1 \leq i, j, k \leq M$.
4. In case of two present and orthogonal vanishing points, test pairwise the orthogonality of the intersection points, that is, $\|\frac{\pi}{2} - \angle(\bar{\mathbf{d}}_i, \bar{\mathbf{d}}_j)\| < t_\phi$ for $i \neq j$ and $1 \leq i, j \leq M$.
   If $\bar{\mathbf{d}}_i \perp \bar{\mathbf{d}}_j$, then compute the third unknown direction by $\bar{\mathbf{d}}_k = \bar{\mathbf{d}}_i \times \bar{\mathbf{d}}_j$.
5. If three orthogonal directions are found, construct rotation matrix $R = \left[\bar{\mathbf{d}}_i\ \bar{\mathbf{d}}_j\ \bar{\mathbf{d}}_k\right]$, otherwise terminate unsuccessful.
6. Enforce $\det(R) = 1$ (Eqn. 2.50).

2.7 Optimal calibration

This section treats a calibration method C-2 that is optimal with respect to the noise in the line segment's endpoints. The noise is assumed to be Gaussian. C-2 is an encapsulation of C-0 and C-1 that tries to overcome the problem with RANSAC's threshold t after the optimization (Step 2, Alg. 2.1). Although, Sec. 2.4 gave a clear understanding of t from a statistical point of view, RANSAC may fail to assign some line segments to the consensus set that would be inliers after optimization. Hartley and Zisserman [2004] discussed this problem in their Golden standard type algorithms under the term "guided matching". Guided matching extends continually the consensus set by these new inliers and re-optimizes alternately until no more line segments are within t. However, the case that line segments might become outliers during this process is not considered. Lens distortion is another reason why it is difficult to find an appropriate t, because it violates the assumption of Gaussian noise.

In contrast to Hartley and Zisserman [2004], we propose in this section C-3 as the grouping of line segments and C-0's and C-1's estimation of vanishing points within an EM framework. Sec. 2.2 showed an approach for the estimation of vanishing points that is optimal with respect to Gaussian noise in the line segment's endpoints. Under the same assumption, Dempster et al. [1977] showed that EM is a MLE with respect to the grouping of line segments. Thus, we conclude that C-3 is optimal with respect to Gaussian noise in the line segment's endpoints. However, this conclusion is only valid, because we care about lens distortion; otherwise the calibration is not optimal.

Alg. 2.4 goes through each step of C-2 which can be used with C-0 that delivers initial values for K and R or with C-1 that delivers initial values for R alone. C-2 follows a recommendation to combine RANSAC-type initializations with EM [Forsyth and Ponce, 2002]. The following

Algorithm 2.4 Method C-2: Calibrate a single camera optimally.

Let $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$ be N line segments in an image. Let $\mathbf{c}$ be the image center. Let K^0 be an initial interior orientation and R^0 be an initial exterior orientation. Let k^0 be the initial distortion coefficient and $\mathbf{c}_k^0$ be the initial radial center. Let $\Pr(\mathbf{v}_1)$, $\Pr(\mathbf{v}_2)$ and $\Pr(\mathbf{v}_3)$ be the priors of the vanishing points and $\Pr(\text{noise})$ be the prior of the noisy line segments. Set $\delta = 0$. Return K^δ, R^δ, k^δ and $\mathbf{c}_k^\delta$ after $\delta_{\min}$ iterations.

1. Normalize $\mathcal{S}$ by computing $\mathrm{T}(\lambda, \mathbf{o})$ (Sec. 2.7.1)

$$\mathcal{S} \leftarrow \mathrm{T} \circ \mathcal{S} = \{\mathrm{T} \circ \mathbf{s}_1, \ldots, \mathrm{T} \circ \mathbf{s}_N\} = \{\begin{pmatrix} \mathrm{T}[\mathbf{s}_{11}]_\mathrm{h} \\ \mathrm{T}[\mathbf{s}_{12}]_\mathrm{h} \end{pmatrix}, \ldots, \begin{pmatrix} \mathrm{T}[\mathbf{s}_{N1}]_\mathrm{h} \\ \mathrm{T}[\mathbf{s}_{N2}]_\mathrm{h} \end{pmatrix}\}.$$

   Let $\mathrm{T}_\mathrm{h} = [\mathrm{T}]_\mathrm{h}$. Normalize

$$\delta_{\max} \leftarrow \lambda\delta_{\max},\ \sigma \leftarrow \lambda\sigma,\ \mathbf{c} \leftarrow \mathrm{T}[\mathbf{c}]_\mathrm{h},\ K^0 \leftarrow \mathrm{T}_\mathrm{h}K^0,\ k^0 \leftarrow \frac{\|\mathbf{c}\|^2}{\|\mathrm{T}[\mathbf{c}]_\mathrm{h}\|^2}k,\ \mathbf{c}_k^0 \leftarrow \mathrm{T}[\mathbf{c}_k^0]_\mathrm{h}.$$

2. Compute vanishing points $\mathbf{v}_i = K^\delta \mathbf{r}_i^\delta$, $1 \le i \le 3$ with $R^\delta = \begin{bmatrix} \mathbf{r}_1^\delta & \mathbf{r}_2^\delta & \mathbf{r}_3^\delta \end{bmatrix}$.
3. Undo the lens distortion $\mathcal{S}' = \mathrm{L}_{k^\delta, \mathbf{c}_k^\delta, \mathbf{c}} \circ \mathcal{S}$ (Sec. 2.7.2).
4. Perform E-step (Sec. 2.7.4)

$$q_{ki} = \Pr(\mathbf{v}_k, k^\delta, \mathbf{c}_k^\delta | \mathbf{s}_i).$$

5. Perform M-step (Sec. 2.7.5)
   - Compute

$$\Pr(\mathbf{v}_k) = \frac{1}{N}\sum_{i=1}^{N} q_{ki} \text{ for } k \in \{1,2,3\}, \qquad \Pr(\text{noise}) = 1 - \sum_{k=1}^{3}\Pr(\mathbf{v}_k).$$

   - Find k^δ and $\mathbf{c}_k^\delta$ with $\sum_{k=1}^{3}\sum_{i=1}^{N} q_{ki}\mathrm{f}_d(\mathbf{v}_k, L_{k^\delta, \mathbf{c}_k^\delta, \mathbf{c}}(\mathbf{s}_i)) \to 0$.
   - Find K^δ and R^δ with $\sum_{k=1}^{3}\sum_{i=1}^{N} q_{ki}\mathrm{f}_d(K^\delta \mathbf{r}_k^\delta, \mathbf{s}_i') \to 0$ and $\mathbf{s}_i' \in \mathcal{S}'$.
6. $\delta \leftarrow \delta + 1$. If $\delta \le \delta_{\min}$, go back to step 3.
7. Denormalization (Sec. 2.7.1)

$$K^\delta \leftarrow \mathrm{T}_\mathrm{h}^{-1}K^\delta,\ k^\delta \leftarrow \frac{\|\mathrm{T}\,(\mathbf{c}\ 1)^\top\|^2}{\|\mathbf{c}\|^2}k^\delta,\ \mathbf{c}_k^\delta \leftarrow [\mathrm{T}_\mathrm{h}^{-1}[\mathbf{c}_k^\delta]_\mathrm{h}]_\mathrm{a}.$$

sections highlight important parts of C-2.

2.7.1 Normalization and denormalization

Normalization of the line segment's endpoints is beneficial during the optimization. Hartley [1995] brought arguments in the context of the 8-point algorithm. Normalization expresses the endpoints within the same range of the coordinate frame. Large differences between the endpoints can degrade the condition of the measurement matrices which are usually decomposed by SVD. In contrast, Triggs et al. [2000] argues that LM is not harmed by such numerical instabilities, however, the normalization prevents the Jacobian from becoming

singular.

The normalization works as follows: It computes in a first step the centroid

$$\mathbf{o} = \frac{1}{2N} \sum_{i=1}^{N} (\mathbf{s}_{i1} + \mathbf{s}_{i2}) \tag{2.51}$$

of all endpoints given all line segments. Then, these endpoints are transformed into a new coordinate system by

$$\mathbf{s}'_{ij} = \mathbf{s}_{ij} - \mathbf{o}, \tag{2.52}$$

with $i = 1, \ldots, N$ and $j \in \{1, 2\}$. Each endpoint's coordinate should lie in an interval $[-k, k]$. Hence the $\mathbf{s}'_{ij}$ are multiplied by

$$\lambda = \frac{2\sqrt{2}kN}{\sum_{i=1}^{N} \|\mathbf{s}'_{ij} - \mathbf{o}\|}, \tag{2.53}$$

resulting in

$$\mathbf{s}'_{ij} = \lambda(\mathbf{s}_{ij} - \mathbf{o}). \tag{2.54}$$

The normalization is isotropic, because it is equally done in each direction with the consequence that each endpoint has an average distance of $\sqrt{2}k$ to the centroid $\mathbf{o}$. A typical value for k is one [Hartley, 1995]. Another way to write the the linear relationship between $\mathbf{s}'_{ij}$ and $\mathbf{s}_{ij}$ (Eqn. 2.54) is by using the similarity transformation

$$\mathrm{T} = \begin{pmatrix} \lambda & 0 & -\lambda o_1 \\ 0 & \lambda & -\lambda o_2 \end{pmatrix}. \tag{2.55}$$

with

$$\mathbf{s}'_{ij} = \mathrm{T}[\mathbf{s}_{ij}]_{\mathrm{h}}. \tag{2.56}$$

The inverse transformation $\mathrm{T}_{\mathrm{h}}^{-1}$ is called denormalization, whereas

$$\mathrm{T}_{\mathrm{h}} = \begin{bmatrix} & \mathrm{T} & \\ 0 & 0 & 1 \end{bmatrix}. \tag{2.57}$$

The denormalization of $\mathbf{s}'_{ij}$ is then

$$\mathbf{s}_{ij} = [T_h^{-1}[\mathbf{s}'_{ij}]_{\mathrm{h}}]_{\mathrm{a}}. \tag{2.58}$$

All other parameters concerning C-2 like the interior orientation, the rotation, the parameters of the lens distortion and the thresholds are normalized or denormalized with the same T. The according transformations are in Alg. 2.4, Step 1 and 7.

2.7.2 Undo the radial lens distortion

We found out empirically that lens distortion has a significant influence in the calibration. This fact is not astonishing, because every kind of nonlinear distortion will impede the localization of concurrent lines. Hence, this method considers a first-order radial lens distortion model. It includes the distortion coefficient k and the radial center $\mathbf{c}_k = (c_{k1}\ c_{k2})^\top$. Positive values of k indicate pincushion, negative values indicate barrel distortion. The larger these values are, the larger the distortion will be. The distortion increases isotropically with the distance from the radial center. The reader is refereed to [Devernay and Faugeras, 2001] for more detail.

First-order radial models are polynomial models. For example, consider a distorted point $\mathbf{x} = (x_1\ x_2)^\top$ in the image plane. In the following these points are the line segment's endpoints. The undistorted point $\mathbf{x}'$ is given by

$$\mathbf{x}' = \mathrm{L}(x, k, \mathbf{c}_k, \mathbf{c}) \tag{2.59}$$

with the undistortion function

$$\mathrm{L}(x, k, \mathbf{c}_k, \mathbf{c}) = \mathbf{x} + k(\mathbf{x} - \mathbf{c}_k)\frac{\|\mathbf{x} - \mathbf{c}_k\|^2}{\mathbf{c}^\top\mathbf{c}}. \tag{2.60}$$

$\mathbf{c}_k$ is for real lenses usually neither the camera's principal point $\mathbf{p}$ nor the image center $\mathbf{c}$ [Hartley and Kang, 2005]. We omit a modeling of the tangential distortion, because the radial model is applicable for a large variety of real lenses. Furthermore, more complicated models are even more vulnerable to noisy line segments, hence, we try to keep the model simple. This decision is supported by observations of Ahmed and Farag [2005], who reported an instable calibration when including the coefficients of the tangential distortion and the radial center.

2.7.3 An EM framework

The optimization problem is as follows: We want to maximize the posterior probability density function of the unknown interior orientation, the rotation and the distortion parameters, given the line segments, while marginalizing over the unknown grouping of line segments J with respect to the three orthogonal vanishing points. Mathematically, we write

$$\underset{K,R,k,\mathbf{c}_k}{\operatorname{argmax}} \sum_{J\in\mathcal{J}} \mathrm{p}(K, R, k, \mathbf{c}_k, J|\mathbf{s}_1, \ldots, \mathbf{s}_N). \tag{2.61}$$

$\mathcal{J}$ are all combinations of possible labels that can be assigned to the line segments. Labels are either a specific vanishing point or noise. The problem we want to solve here is the well-known

chicken and egg problem, that is, a known grouping solves for the parameter's optimization and knowing the parameters delivers straightforward the grouping of the line segments. EM offers the grouping of line segments (E-step) and the calibration (M-step) in an alternating fashion. It was shown that EM converges to an optimal solution with respect to a Gaussian error model. More details about the EM algorithm are in [Bishop, 2006; Dellaert, 2002; Duda et al., 2001; Forsyth and Ponce, 2002; McLachlan and Krishnan, 2008].

Let the probability density function for a line segment $\mathbf{s}_i$ supporting one of the m vanishing points $\mathbf{v}_k$, $1 \leq k \leq m$ be proportional to a mixture of m conditional Gaussian densities

$$\mathrm{p}(\mathbf{s}_i|\mathbf{v}_1, \ldots, \mathbf{v}_m) = \sum_{k=1}^{m} \mathrm{p}(\mathbf{l}_i|\mathbf{v}_k) \Pr(\mathbf{v}_k), \tag{2.62}$$

whereas $\mathbf{l}_i = \mathrm{f_u}(\mathbf{v}_k, \mathbf{s}_i)$ (Eqn. 2.9) is the line segment that passes through $\mathbf{v}_k$ and huddles optimally against $\mathbf{s}_i$ (Sec. 2.2). $\Pr(\mathbf{v}_k)$ are the priors of the vanishing points which express their dominance in the image. Thus, a prior $\Pr(\mathbf{v}_k)$ is initially set to the number of line segments supporting $\mathbf{v}_k$ divide by the total number N of line segments.

If a line segment does not support any vanishing point, it will be a noisy line segment. More precisely, when for all $k = 1, \ldots, m$ the error function $\mathrm{f_d}(\mathbf{v}_k, \mathbf{s}_i)$ (Sec. 2.7) is larger than 2σ we can statistically say the line segment is noisy. To implement noise as a special component into our Gaussian mixture model, we evaluate the one-sided Gaussian density function at 2σ which gives the value

$$\mathrm{p}(\text{noise}) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \mathrm{e}^{(-\frac{1}{\sigma})}. \tag{2.63}$$

Remember σ as the noise level of the image. The prior probability of the noisy line segment's occurrence is

$$\Pr(\text{noise}) = 1 - \sum_{k=1}^{m} \Pr(\mathbf{v}_k), \tag{2.64}$$

because the prior probabilities must sum to one.

In the work of Antone and Teller [2000]; Kosecka and Zhang [2002], EM's unknown parameters to estimate are the vanishing points itself. Thus, the calibration of the interior orientation and the rotation to the scene is a post-processing step. Schindler and Dellaert [2004] criticized this approach, because the actual unknown parameters are K and R. In fact, K and R simply derive the three orthogonal vanishing points (Eqn. 2.41). We follow Schindler et al. by our problem formulation (Eqn. 2.61) and replace the mixture model based on vanishing points (Eqn. 2.62) by a mixture model based on the actual parameters to optimize, that is,

$$\mathrm{p}(\mathbf{s}_i|K, R, k, \mathbf{c}_k) = \sum_{k=1}^{3} \mathrm{p}(\mathbf{l}_i|K\mathbf{r}_k, k, \mathbf{c}_k) \Pr(K\mathbf{r}_k), \tag{2.65}$$

with $R = [\mathbf{r}_1\ \mathbf{r}_2\ \mathbf{r}_3]$.

2.7.4 E-step

Eqn. 2.61 is directly inapplicable, because J is unknown. However, when we would assume for a moment the labeling k of a $\mathbf{s}_i$, we are able to compute its posterior membership probability by

$$\begin{aligned} q_{ki} &= \Pr(K, R, k, \mathbf{c}_k, J = (k\ i)\,|\mathbf{s}_i) && (2.66)\\ &= \Pr(K\mathbf{r}_k, k, \mathbf{c}_k|\mathbf{s}_i). && (2.67)\end{aligned}$$

This idea is the essence of EM, unfortunately it assumes a good initial approximation of the parameters. Using Bayes law, the data likelihood of mixture component k in Eqn. 2.65 and the noise model (Eqn. 2.63 and Eqn. 2.64) the membership probability changes into

$$q_{ki} = \frac{1}{\mathrm{p}(\mathbf{l}_i)} \left(\mathrm{p}(\mathbf{l}_i|K\mathbf{r}_k, k, \mathbf{c}_k) \Pr(K\mathbf{r}_k) + \mathrm{p}(\text{noise}) \Pr(\text{noise})\right). \tag{2.68}$$

$\mathrm{p}(\mathbf{l}_i)$ is a normalization factor to ensure Eqn. 2.64 and it is written by the sum of the complete evidence in the data and in the noise which gives

$$\mathrm{p}(\mathbf{l}_i) = \mathrm{p}(\text{noise}) \Pr(\text{noise}) + \sum_{k=1}^{3} \mathrm{p}(\mathbf{l}_i|K\mathbf{r}_k, k, \mathbf{c}_k) \Pr(K\mathbf{r}_k). \tag{2.69}$$

2.7.5 M-step

Two consecutive optimizations happen in the M-step. Based on the expected grouping q_{ki} which is previously done in the E-step, first the distortion parameters and then K and R are optimized. The former optimization problem is written by

$$\underset{k,\mathbf{c}_k}{\operatorname{argmin}} \sum_{k=1}^{3} \sum_{i=1}^{N} q_{ki} \mathrm{f}_d(\mathbf{v}_k, \mathrm{L}_{k,\mathbf{c}_k,\mathbf{c}}(\mathbf{s}_i)). \tag{2.70}$$

As K and R are kept fixed, the orthogonal vanishing points are pre-computed. Then, a nonlinear optimization in the parameters of $\mathrm{L}(\cdot)$ takes place. This kind of optimization is well known under the term "plumb line calibration" (Sec. 2.1.2). As the vanishing points are directly derived by K and R, they establish a very elegant relationship between the interior orientation, the rotation and the lens distortion.

Contrary, the optimization of the interior orientation and the rotation is given by

$$\underset{K,R}{\operatorname{argmin}} \sum_{k=1}^{3} \sum_{i=1}^{N} q_{ki} \mathrm{f}_d(K\mathbf{r}_k, \mathrm{L}_{k,\mathbf{c}_k,\mathbf{c}}(\mathbf{s}_i)). \tag{2.71}$$

Here the undistortion of the line segments can be done in a pre-processing. Then, the optimization is similar as for the optimization of the intersection point given a number of line segments (Sec. 2.2). However, we optimize in the parameter values directly and the computation of the vanishing points is only an intermediate step to measure the error to the line segments by $\mathrm{f_d}(\cdot)$. Note that a proper R's parametrization is important to avoid problems in the objective function like singularity [Triggs et al., 2000]. The optimization is for both cases done by the LM algorithm.

2.8 Incremental calibration

This last section of the chapter describes an incremental version of Alg. 2.4. This method, we call it C-3, is able to run either on-line on a video stream or off-line on a sequence of images. The outline of C-3 is shown by Alg. 2.5. Before C-3 can run, initial values of the interior orientation and the rotation matrix are required. C-0 (Alg. 2.2) or when we know K, C-1 (Alg. 2.3) do this initialization. Next, initial values of the lens distortion parameters are needed. When the distortion is moderate, an adequate initialization would be $\mathbf{c}_k \leftarrow \mathbf{c}$ and $k \leftarrow 0$. However, in some cases the lens distortion is severe, thus, other approaches like the method of Devernay and Faugeras [2001] are thinkable. Unfortunately, we observed empirically that especially his method fails frequently when the camera is in critical cases or when the line segments are short.

C-3 now reads multiple, consecutive images of a static camera, that is, the edges that are present in the image should be the same. Noise will disturb this expectation, however, the underlying random process is the same for all images. Illumination changes and moving objects will also introduce new line segments perhaps in different areas of the image. Exactly this makes the advantage of C-3 against C-2, because new information is used to perhaps improve the calibration.

After reading an image at time instant t and detecting all line segments $\mathcal{S}$, C-3 calls C-2 with the current estimates K, R, k, $\mathbf{c}_k$. One could suppose to run the EM in C-2 until convergence, however, as a sequence of images is available and to be more efficient we propose to run EM only once ($\delta_{\max} = 1$) for each image. As C-3 calls C-2 for consecutive images, the EM is used as a generalized EM which will also converge like the original EM, because the expected log-likelihood is never increased [McLachlan and Krishnan, 2008]. C-2 returns values K', R', k', $\mathbf{c}'_k$ for a possible update.

Algorithm 2.5 Method C-3: Calibrate a single camera incrementally.

Use either C-0 or C-1 to initially compute the interior orientation K, the rotation matrix R and the distortion parameters k and $\mathbf{c}_k$ respectively. Let τ_g be a global level of the uncertainty that is initialized by $\text{trace}(\Lambda_K) + \text{trace}(\Lambda_R) + \sigma_k^2 + \text{trace}(\Lambda_{\mathbf{c}_k})$. Set $t \leftarrow 1$. Let ϵ be the precision of the computer arithmetics. Let μ be a learning rate. Return K, R, k and $\mathbf{c}_k$ after processing $t_{\max}$ image frames of a video.

1. Detect line segments $\mathcal{S}$ in image t.
2. Perform C-2 with $\mathcal{S}$, K, R, k, $\mathbf{c}_k$, $\delta_{\min} = 1$. Store the updated values in K', R', k' and $\mathbf{c}'_k$.
3. Compute the uncertainties
$$\tau' \leftarrow \text{trace}(\Lambda_{K'}) + \text{trace}(\Lambda_{R'}) + \sigma_{k'}^2 + \text{trace}\left(\Lambda_{\mathbf{c}'_k}\right)$$
$$\tau \leftarrow \text{trace}(\Lambda_K) + \text{trace}(\Lambda_R) + \sigma_k^2 + \text{trace}(\Lambda_{\mathbf{c}_k}).$$
4. Update

   Regular update: If $\tau' < \tau$, then
   $K \leftarrow K'$, $R \leftarrow R'$, $k \leftarrow k'$, $\mathbf{c}_k \leftarrow \mathbf{c}'_k$. $\tau \leftarrow \tau'$.

   Random update: Let $\xi \sim [0, 1]$. If $\tau' \geq \tau$ and $\tau_g > \tau'$ and $(\xi + \epsilon) < \mathrm{e}^{\frac{\tau - \tau'}{\tau_g - \tau' + \epsilon}}$, then
   $K \leftarrow K'$, $R \leftarrow R'$, $k \leftarrow k'$, $\mathbf{c}_k \leftarrow \mathbf{c}'_k$. $\tau \leftarrow \tau'$.
5. Adaptation
   $\tau_g \leftarrow \mu\tau + (1 - \mu)\tau_g$.
6. Go back to step 1, until $t > t_{\max}$.

Uncertainties τ and τ' of on the one hand K, R, k, $\mathbf{c}_k$ and on the other hand K', R', k', $\mathbf{c}'_k$ decide about an update. We propose to use the trace of the related covariance matrices, that is,

$$\tau \leftarrow \text{trace}(\Lambda_K) + \text{trace}(\Lambda_R) + \sigma_k^2 + \text{trace}(\Lambda_{\mathbf{c}_k}) \tag{2.72}$$
$$\tau' \leftarrow \text{trace}(\Lambda_{K'}) + \text{trace}(\Lambda_{R'}) + \sigma_{k'}^2 + \text{trace}\left(\Lambda_{\mathbf{c}'_k}\right). \tag{2.73}$$

The covariance matrices will be smaller, if more and longer line segments are found. If the angle between the lines of concurrent line segments is small, then the position of the corresponding vanishing point will be highly uncertain. Note that this case happens frequently. As K and R are linked to the vanishing points, the covariance matrices will also have larger elements - especially along the diagonals. The contrary happens when the angle between the lines is large. Look at Liebowitz [Liebowitz, 2001, 3.6, p.72-77] for a detailed discussion of the uncertainty of intersection points.

Let τ_g be a global limit that cannot be exceeded by neither τ nor τ'. When the uncertainty τ' of the possible update is smaller than the uncertainty τ of the current estimate, then the update should effectively occur, that is, the line segments in the current image yield to less uncertainty in the estimates than at every time instant before. Otherwise, no update should

happen, however, in some scenes the argument that less line segments or line segments in better position give always a better estimate is not true. For example, some line segments do not lie exactly on orthogonal real world edges. This often happens when illumination changes and shades produce irritating edges. Our solution to this problem is that C-3 has the chance for an update even when $\tau' > \tau$ by using Simulated Annealing (SA) [Duda et al., 2001]. The so-called random update is only allowed when τ' is smaller than the absolut limit τ_g and inequality

$$(\xi + \epsilon) < \mathrm{e}^{\frac{\tau - \tau'}{\tau_g - \tau' + \epsilon}} \tag{2.74}$$

is fulfilled. ξ is a evenly distributed random variable $\xi \sim [0, 1]$. The chance for a random update is small in case τ_g is very close to τ' or when the difference between τ and τ' is too large.

Regardless an update happens or not, τ_g will be adapted in every step t towards the current uncertainty τ by using

$$\tau_g = \mu\tau + (1 - \mu)\tau_g \tag{2.75}$$

with a learning rate μ that controls the speed of adaptation. This decreasing upper limit guarantees that the current estimate with perhaps a low uncertainty is not lost. Without τ_g C-3 could loose a good estimate when suddenly line segments are lost due to illumination changes, occlusion or the like.

Chapter 3

Localization of distant cameras

Imagine two cameras $P_1 = K_1 R_1 \bar{R} [I_3 \ -\mathbf{C_1}]$ and $P_2 = K_2 R_2 [I_3 \ -\mathbf{C_2}]$. The last chapter came up with the methods C-0 till C-3 that allow a calibration of the interior orientations K_1, K_2 and the orientation to the orthogonal world R_1, R_2. This chapter describes the methods L-0, L-1 and L-2 that will estimate the missing camera centers $\mathbf{C_1}$ and $\mathbf{C_2}$ and the final rotational ambiguity $\bar{R}$ between the two cameras under different conditions and assumptions.

The emphasis is especially laid on distant cameras with slightly overlapping views (Sec. 3.2) and with non-overlapping views (Sec. 3.3), where our approach of successive calibration and then localization has significant advantages compared to comparable methods in the literature.

The novelties of L-0, L-1 and L-2 are as follows: (i) The localization is possible with at least three point matches, (ii) all these points can be collinear and (iii) the influence of the point distribution within the images with respect to the accuracy of the calibration is reduced. (iv) L-2 can simultaneously handle cameras with overlapping and non-overlapping views, (v) it is the first method that does not assume objects moving on a common plane in case of cameras with non-overlapping views and finally (vi) L-2 is computationally attractive, because it computes the solution in closed-form by using SVD.

3.1 Previous and related work

Localization is of concern in Photogrammetry, Computer Vision and Robotics. Although the geometry of the problem is well understood since the last century [Semple and Kneebone, 1998], global, robust and efficient computational methods are not that old [Hartley and Zisserman, 2004]. Unfortunately, scalable methods localizing large networks with hundreds of cameras and exploiting the camera's information all at once are in the distance, because the similar "Structure From Motion" problem (SFM) is NP-hard [Nister et al., 2007]. Sec. 3.1.1

gives the reader an overview of the research on localization in overlapping views. Disjoint views are an especially hard problem with only one encouraging approach, therefore, this topic is separately discussed in Sec. 3.1.2. Related fields like SFM or "Simultaneous Localization and Mapping" (SLAM) are also touched by these two sections. Finally, Sec. 3.1.3 outlines the difficulties of the current research where some of them are then tackled by the rest of this chapter.

3.1.1 Localization with overlap

Sanden [1908] formulated already in 1908 a bilinear equation system for the localization of pre-calibrated cameras, where a matrix analogous to the Essential matrix appears which, however, was introduced decades later to the Computer Vision community by Longuet-Higgins [1981]. The Essential matrix E encapsulates the former mentioned parameters $\bar{R}$, $\mathbf{C_1}$ and $\mathbf{C_2}$ between two cameras. From that time instant, many people in the Computer Vision community worked on efficient, robust and automatic estimators using image features. Hartley and Zisserman [2004] showed that the bilinear relation between E and the point correspondences can be written as a linear equation system in the unknowns of the Essential matrix and that at least eight point correspondences lead to a unique, global solution by using SVD. Their method is further robust when it is embedded into a RANSAC framework and it is efficient with respect to the SVD, because more than the minimal eight point correspondences make no sense in a hypothesis-and-test architecture as RANSAC is. Unfortunately this optimization is over-constrained, because eight points are more points than necessary to solve the underlying geometric problem, that is, the degree of freedom of the geometry is smaller than the number of constraints given by the eight point correspondences. The reason for this problem is that Hartley and Zisserman's method does not constrain the Essential matrix properly; an Essential matrix E is singular as the more general Fundamental matrix is and it fulfills additionally the Demazure constraint [Demazure, 1981], that is, the two non-zero singular values are equal. Therefore, people looked for closed-form optimizers that additionally care about these constraints and work with a minimal number of point correspondences.

Kruppa [1913] showed in 1913 that this minimal number of point correspondences for a proper localization is five and that eleven complex and real solutions can be formulated as the intersections of two sextic curves. As the whole optimization is iterative and rather inefficient, the people hoped to find a way to translate the problem into a simpler algebraic representation. Later, Faugeras and Maybank [1990] formulate the problem as a polynomial and they showed that at most ten instead of eleven solutions exist. After six further years, Philip [1996], Triggs [2000b], and then Nister [2004] developed upon this polynomial representation non-iterative, global optimizers that solve the problem for the first time efficiently and by ransacing also robustly. A summary is given in [Henrik Stewenius and Nister, 2006].

For the sake of completeness, we allude also the early work of Sturm [1869] who used directly the singularity constraint and therefore needed at least seven point correspondences, resulting in three possible solutions of E. Although six point correspondences suffice together with the Demazure constraint, the resulting optimizer fails when all points are coplanar or near coplanar [Philip, 1998]. Oscar Pizzaro and Singh [2003] found a solution in picking only four out of the nine possible equations form the Demazure constraint that make the algorithm invulnerable to coplanar points.

Instead of localization alone we might think about localization and a reconstruction of the points that give the measurements. The situation of a moving camera and static points is called SFM and has been intensively studied in Photogrammetry and Computer Vision over the last decades. The gold standard is to use Bundle Adjustment. Excellent reviews are found in McGlone [2004] and Triggs et al. [2000]; particular the least-squares optimization and the right parametrization of $\bar{R}$ is reviewed by Horn [1990]. The awkward problem in Bundle Adjustment is the necessary good initialization and not primarily the efficiency or robustness [Engels et al., 2006]. Fortunately, closed-form optimizers similar to the aforementioned ones deliver reasonably good initializations.

If one can assume a camera's affine model instead of the pinhole model, an optimal solution of SFM with respect to a Gaussian assumption about the noise is possible by using a closed-form factorization instead of Bundle Adjustment [Tomasi and Kanade, 1992]. The heart of the optimizer is a SVD of a constructed measurement matrix. Unfortunately, the factorization is not applicable with a pinhole model, however, a bunch of research has created some useful methods [Sturm and Triggs, 1996]. Measurements of a point in all views are needed and this is by far the greatest disadvantage of this approach, although, ideas to fill missing elements of the measurement matrix show certain success [Martinec and Pajdla, 2002].

Knowing $\bar{R}$ overcomes the problem of missing measurements, because instead of having rows in the measurement matrix formed by measurements among all cameras, each row is now formed by a measurement of a single camera. In other words, the interior orientation and $\bar{R}$ define the infinite homography which is pictured in the first three columns of the camera matrix. Rother [2003] concluded that any finite plane-induced homography has the same effect, hence, this approach is called planar reconstruction or sometimes plane+parallax method. Several people discovered this simplification [Heyden and Aström, 1995; Shashua and Navab, 1994; Triggs, 2000a], however, it was Rother [2003] who used this insight for the reconstruction with their novel Direct Reference Plane (DRP) method, whereas two points in each of the two views are sufficient to solve the problem. A similar approach that localizes solely the cameras and neglects the reconstruction of points was developed by Kaucic et al. [2001]. Although his method is more efficient, collinear points are a critical configuration for the algorithm.

All methods discussed so far are off-line. An alternative, probabilistic on-line method was proposed by Funiak et al. [2006] who used a Kalman filter for the estimation. In that sense all previously discussed methods belong to the area of optimization instead to the area of estimation. As the common objective function to optimize is non-linear, they formulated a state-space where the parametrization of the rotation is linear. The method is efficient and better scalable as the aforementioned ones, because it uses an approximation of the current belief state, but neither a guarantee that the equilibrium is a global optimum is possible, nor care about gross outliers is taken. Funiak's method is similar to the work of Taylor et al. [2006]. Both are numbered among the SLAT approach which is discussed in the next section.

3.1.2 Localization without overlap

Work concerning the pure localization with non-overlapping views is rare and their research started approximately six years ago by the work of Caspi and Irani [2002]. Measuring a point in several cameras is impossible, thus, other sources of information must be found. An obvious source is the dynamic behavior of objects, for example moving people, because the extrapolation of the trajectories in the unseen areas hallucinates the missing point correspondences.

The literature mentions two different ways to use an object's dynamics. One is to extrapolate the trajectory into the unseen regions. Javed et al. [2003] and Fisher [2002] followed this approach by tracking objects on a ground plane. When, for example, an object leaves with constant speed the visible area of camera one, the trajectory is linearly elongated by the last observed object's velocity, until the person enters camera two. Javed et al. were uninterested in 3D localization, but to find the projected borders of a camera's field of view with the aim to use these lines to match people. Cameras are uncalibrated and the extrapolation happens in the image plane with all the disadvantages of perspective projection. Fisher, however, knew the interior orientation which provides together with an orientation step using stars as vanishing points the homographies induces by the ground plane. An attempt to relax the linear unseen trajectory to a quadratic one was recently made by Anjum et al. [2007]. They transform all the measurements into the ground plane. To suppress noise a Kalman filter is used. A polynomial is fitted to the tracklets in each view into the unseen area on the ground plane. The given measurements and the hallucinated ones are then reconciled to find the position and orientation.

The other way is to localize the cameras and to track simultaneously the moving objects. This approach breaks into the field "Simultaneous Localization and Tracking" (SLAT) which was first mentioned by Rahimi et al. [2004]. They showed that except SLAT, the separate localization and tracking by EM type methods without further assumptions is impossible.

SLAT is in a sense dual to SLAM [Thrun et al., 2005], where a moving robot registers markers in an unknown environment, builds up a global map and localizes itself. In SLAT the cameras are static and the objects move, in SLAM the cameras move and the markers are fixed. The remarkable property of SLAT is that smoothness of motion is sufficient to assume. Rahimi et al. [2004] formulated SLAT as Bayesian problem and proposed a MAP estimation that is under Gaussian assumptions a least-squares optimization of a well known objective function. However, their Newton-Raphson optimizer extravagates quickly the computational tractability, because the number of unknowns in the objective function increases rapidly by adding further trajectories. This disadvantage is recently tackled by Rudoy and Rohrs [2006]. They split the trajectory estimation into an independent estimation while the object is visible and while it is outside the view. Treating the measurement terms as equality constraints realizes further improvements.

A known correspondence of tracklets in case of multiple moving objects is usually assumed. Little work exist that tries to relax this assumption. Rahimi et al. [2004] mentioned the idea that EM could solve this problem by optimizing the reconstruction error over all camera and trajectory configurations while marginalizing over the object labels [Dellaert et al., 2000]. Sheikh et al. [2007] proofed this approach to be applicable in aerial surveillance, but Li and Hartley [2007] stated rightly that EM cannot guarantee a correct solution, because it can stuck in local minima. To reach global optimality, they propose a simultaneous estimation of correspondences and pose of the cameras by exhaustively search over all configurations using Lipschitz optimization.

3.1.3 Remarks

We remark the following difficulties for overlapping views:

- Generally, at least measurements of five points in all views are necessary and sufficient to localize the cameras and reconstruct the 3D points. However, five points are in many practical situations not visible. As our approach envisages the calibration and the localization as separate steps, only two points instead of five are needed with overlapping views.
- The spatial configuration of points influences significantly the accuracy. Usually, evenly distributed measurements are necessary for a precise reconstruction. As it is always in Computer Vision, prior knowledge, for example, about the rotation will help to mitigate the error.

For non-overlapping views:

- Current research assumes in all cases coplanar trajectories. In many practical cases

this assumption is difficult to achieve, for example with head tracking. Instead of using a plane, we propose the 3D space for fusion. However, the computation of the points is more complicated, because it cannot be done separately for each camera. Instead triangulation with both cameras is done.

- The projection of measurements onto the ground plane is sensible to noise, that is, relatively small errors in the image plane cause large errors in the ground plane. The higher the perspective distortion the larger the problem which usually happens on the borders of the image. Again, the 3D space is a better place for a projection, because the intersection of two rays respond less sensible to changes in the measurements.
- The methods except the SLAT methods assume constant, linear velocity which is a strong assumption.
- None of the non-overlapping view methods can guarantee a global optimum, because the objective function is non-linear.
- Although Rahimi's approach is promising, it is inefficient, because the number of unknowns grows quickly.

Furthermore, no method work with overlapping and non-overlapping views. Sec. 3.1.2 presents a method L-3 that is capable in doing with both scenarios and tackles all the difficulties. Unfortunately, L-3 is still inefficient with 50 or more trajectories.

3.2 Two cameras with slightly overlapping views

The last section presented well studied methods that compute the Essential matrix in overlapping scenarios. Evenly distributed measurements are the requirement for an accurate estimation. Unfortunately, the slight overlaps violate this requirement. Sec. 4.3.1 shows especially the estimation of the rotation as sensitive. As methods C-0 till C-3 already compute the rotation to the orthogonal scene, only a finite number of rotational ambiguities for $\bar{R}$ come into question.

In the following we discuss a method L-0 that uses this a priori known rotations together with the DRP-method to compute $\mathbf{C_1}$, $\mathbf{C_2}$ and $\bar{R}$. Let the vanishing points in the image of a camera be $\mathbf{v_1}$, $\mathbf{v_2}$ and $\mathbf{v_3}$. Consider the pairwise, orthogonal rays

$$\mathbf{d_i} = K^{-1}\mathbf{v_i}, \tag{3.1}$$

$1 \leq i \leq 3$. For any two rays $\mathbf{d_i}$ and $\mathbf{d_j}$, the third, orthogonal ray $\mathbf{d_k}$ is determined by the cross product

$$\mathbf{d_k} = \mathbf{d_i} \times \mathbf{d_j} \tag{3.2}$$

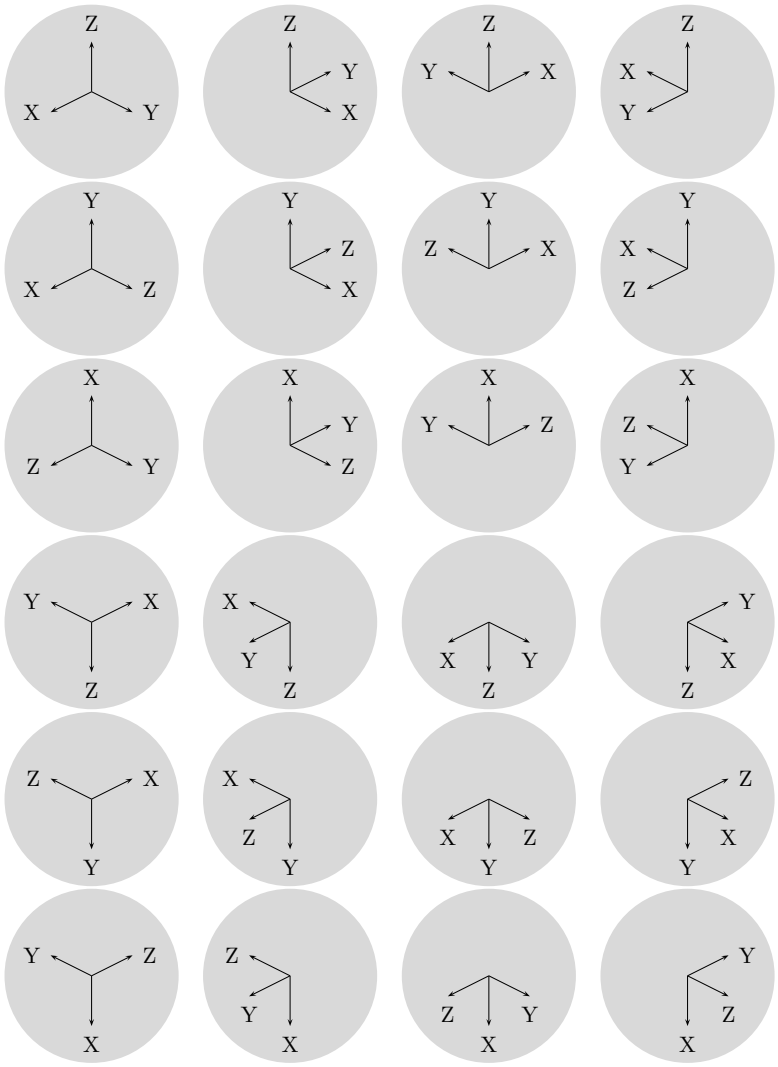


Figure 3.1: This illustration shows the 24 possible rotation ambiguities. Each column depicts a 90 deg rotation around the vertical axis. There are $3! = 6$ different combinations (rows) to label the axes by X, Y and Z.

with $1 \leq k, i, j \leq 3, k \neq i \neq j$. Each ray $\mathbf{d_i}$ passes exactly through the vanishing point $\mathbf{v_i}$ and, obviously, each ray represents a coordinate axis of the camera coordinate system. The rotation matrix R that transforms the canonical world coordinate system into the camera coordinate system is given by

$$R = \left(\pm \frac{\mathbf{d_i}}{\|\mathbf{d_i}\|} \pm \frac{\mathbf{d_j}}{\|\mathbf{d_j}\|} \pm \frac{\mathbf{d_k}}{\|\mathbf{d_k}\|} \right) \tag{3.3}$$

with $1 \leq i, j, k \leq 3$ and $i \neq j \neq k$. Note that each column of R is normalized to 1 and $\det(R) = 1$, because rotation matrices are per definition orthonormal.

3.2.1 Orientation ambiguity

Unfortunately, the construction of R (Eqn. 3.3) is ambiguous [Antone and Teller, 2000; Rother, 2003]. Hence, the question is how many geometrically valid rotation matrices exist. Any two columns of R can be occupied by two arbitrary $\mathbf{d_i}$ and $\mathbf{d_j}$. The remaining third column is always determined by $\mathbf{d_k}$ (Eqn. 3.2). Consequently, $3! = 6$ combinations of $\mathbf{d_k}$, $\mathbf{d_i}$ and $\mathbf{d_j}$ exist. $\mathbf{d_i}$ and $\mathbf{d_j}$ can further have arbitrary direction, that is, the sign of $\mathbf{d_i}$ and $\mathbf{d_j}$ is undefined. However, the direction of $\mathbf{d_k}$ is also determined by the cross product. Here, $2^2 = 4$ combinations are possible. Together, $6 \cdot 4 = 24$ geometrically valid rotation matrices exist. Fig. 3.1 shows an illustration of these possible rotations.

Let us now discuss the case of two cameras. Each view is orientated to the environment by 24 possible rotation matrices R_1^{1-24} and R_2^{1-24}. As the orientation of the environment to the world is the same for both cameras, we may chose one of the 24 rotation matrices for Camera 1 as R_1. This fixes the overall rotation of the metric space. For Camera 2, all R_2^{1-24} are still possible, but which of them are geometrically correct? Rother proposed to compute the fundamental matrix F_{12} between Camera 1 and Camera 2 and then to check the skew-symmetry of matrix

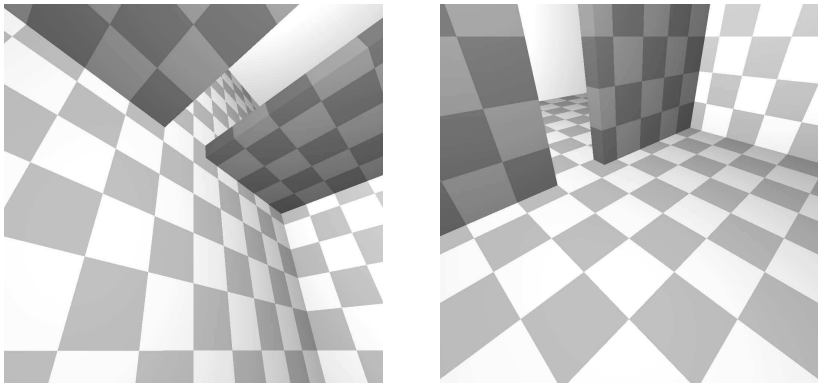
$$[\mathbf{e}]_\times = (K_2 R_2)^\top F_{12} K_1 R_1 \tag{3.4}$$

for every R_2 out of R_2^{1-24} [Rother, 2003]. However, this approach assumes the computation of F_{12} from point correspondences which is difficult with distant cameras and certainly impossible with small overlapping field of views [Pflugfelder and Bischof, 2006a]. Furthermore, the evaluation of the skew-symmetry of $[\mathbf{e}]_\times$ is numerically unstable and is on top of everything not sufficient to identify the correct R_2. Hartley and Zisserman [2004] showed that the Essential matrix encapsulates two geometrically correct rotation matrices between two views. The same problem is here. Without the test that all reconstructed 3D points of all point correspondences lie in front of both cameras, the physically correct R_2 is not identifiable.

Our approach assumes cameras in natural pose (Fig. 3.2). Under this assumption, the vanishing point that represents the vertical direction in the world can be separately identified in each camera (Fig. 3.3).

Result 3.1. *Let* $\mathbf{l_k} = \mathbf{v_i} \times \mathbf{v_j}$ *be the polar vanishing line of* $\mathbf{v_k}$. $\mathbf{v_k}$ *is the vanishing point that represents the vertical direction in the world if the enclosed angle between* $\mathbf{l_k}$ *and the abscissa of the image coordinate system is smaller than a threshold* t.

Usually, t is chosen between 3 deg and 5 deg. The identification of the vertical direction reduces R_2^{1-24} to $2! \cdot 4 = 8$ possible rotation matrices R_2^{1-8}, because the last column of R_2 is now a priori determined (Fig. 3.5). However, the two geometrically correct rotation matrices



(a) Camera in arbitrary pose (b) Camera in natural pose

Figure 3.2: Contrary to an arbitrary pose (a), a camera is in natural pose when it is in an upright position (b). The vanishing line of the ground plane is parallel to the image's abscissa. Natural pose is the usual pose of a surveillance camera.

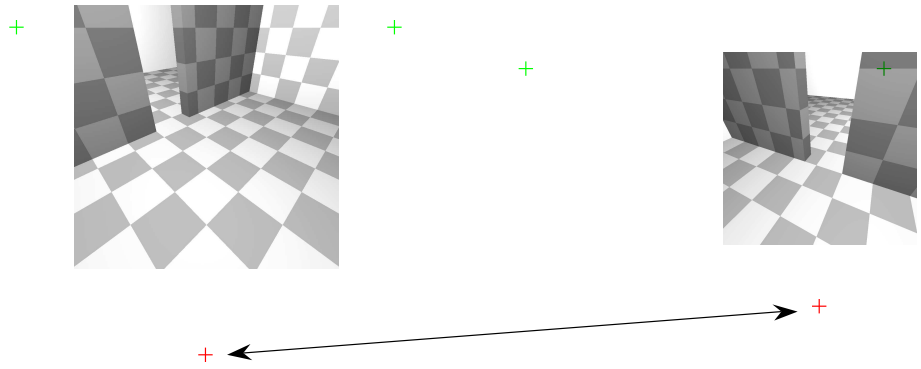


Figure 3.3: Matching the vertical vanishing points (red) among the others (green) by using the idea of Result 3.1.

are still included in R_2^{1-8}. They only differ by the arbitrary choice of the direction of the vertical coordinate axis. Fortunately, this problem can be fixed by assuming that the vertical axis is always pointing into the opposite direction of the image ordinate.

Now, we do the following: Choose any rotation matrix $R_2 \in \{R_2^1, \ldots, R_2^8\}$ and construct a camera matrix $P_2 = (K_2 R_2 \ \mathbf{p_2})$ with arbitrary last column, for example, let the last column be the principal point $\mathbf{p_2}$ which is the last column of K_2. Next, determine if the view of the camera is down to earth or up to the sky. The former view will be present, if $\mathbf{v_k}$ is below its polar $\mathbf{l_k}$ in the image plane (Fig. 3.4). Below means in this context in the direction of the image ordinate. Conversely, the latter view will be present, if $\mathbf{v_k}$ is above $\mathbf{l_k}$ in the image plane. Furthermore, let point $\mathbf{a}$ be the projection of any finite world point along the vertical coordinate axis that corresponds to $\mathbf{v_k}$.

Result 3.2. *If* $\|\mathbf{v_k} - \mathbf{a}\| > \|\mathbf{v_k} - \mathbf{p_2}\|$ *and the camera is looking up or if* $\|\mathbf{v_k} - \mathbf{a}\| < \|\mathbf{v_k} - \mathbf{p_2}\|$ *and the camera is looking down,* $\mathbf{d_k}$ *will be replaced by* $-\mathbf{d_k}$.

Now, R_2^{1-8} reduces to $2! \cdot 2 = 4$ possible rotation matrices R_2^{1-4} (Fig. 3.6). Interestingly, the

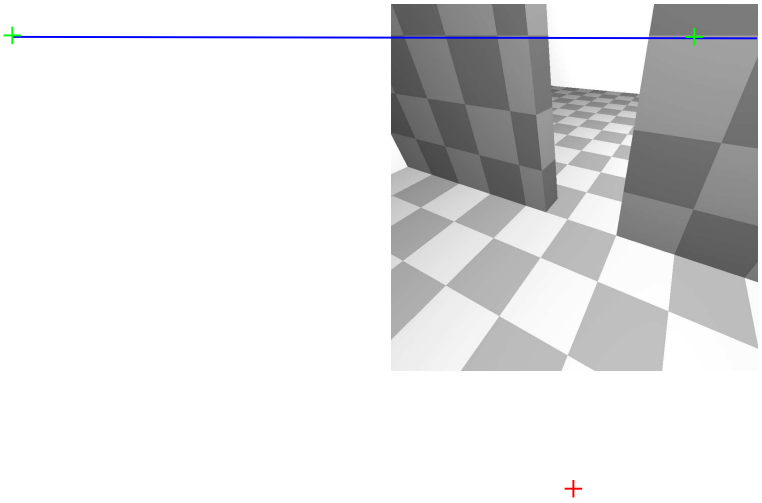


Figure 3.4: After the vertical vanishing point (red) is identified, the other two vanishing points (green) form the horizon (blue). The camera is looking up, because the vertical vanishing point is below the horizon.

only geometrically correct rotation matrix of R_2^{1-4} is also the only physically correct rotation matrix. L-0 uses this fact in the next section. L-1 relaxes the assumption that $\bar{R}$ is one of the R_2^{1-4}. It only assumes for cameras in natural pose that the scenes for each view are orthogonal and that the rotation between the cameras is an arbitrary rotation around the common vertical axis. Fig. 3.7 illustrates this rotation and Sec. 3.2.3 explains L-1 in detail.

3.2.2 Method L-0

Without loss of generality, we can assign the origin of the world coordinate system to the center of Camera 1, that is, $R_1 = I_3$ and $\mathbf{C_1} = 0$. We know from the previous section that one of the R_2^{1-4} is R_2. As R_1 is the identity matrix, R_2 is the unknown ambiguity

$$\bar{R} = R_2 R_1^{-1} \tag{3.5}$$

and the last unknown translation between Camera 2 and Camera 1 is $\mathbf{C_2}$. Assume for a moment that we would know R_2. In that case we are able to compute $\mathbf{C_2}$ with at least two point correspondences by using a plane and parallax method such as the DRP-method or Kaucic's method. We decided for the DRP-method, because it also works with collinear points. Sec. 3.3.1 explains the DRP-method in detail.

As only one rotation matrix out of the R_2^{1-4} is physically correct, the error between the point measurements and the re-projection of the reconstructed points is minimal for the correct $\bar{R}$. The reconstruction of the points happens either implicitly by the DRP-method or by triangulation. This approach to search for the correct rotation is current research, for example, Li and Hartley [2007] proposed a global method which is based on a similar idea.

Algorithm 3.1 Method L-0: Localize two cameras using the DRP-method.

Let $\mathbf{x_{ij}}$ be the image measurements of N points ($1 \leq i \leq N$) in two cameras ($j \in \{1, 2\}$). Let K_j and R_j be the camera parameters which are obtained by C-0 till C-3 (Chap. 2). Return the final rotation ambiguity $\bar{R}$ and the camera center $\mathbf{C_2}$. $\mathbf{C_1}$ is $\mathbf{0}$, R_1 is I_3. Let i_{max} in the beginning be zero.

1. Go through all rotation matrices R_2^i with $i \in \{1, \ldots, 4\}$ and assign the current rotation matrix R_2^i to R_2.
2. All point correspondences $\mathbf{x_{ij}}$ are mapped to the plane-at-infinity by the inverse, infinite homographies $H_{\infty j}^{-1} = (K_j R_j)^{-1}$. The points are then $\mathbf{d_{ij}} = H_{\infty j}^{-1} \mathbf{x_{ij}}$.
3. Normalize $\mathbf{d_{ij}}$ spherically, that is, $\bar{\mathbf{d}}_{\mathbf{ij}} = \frac{\mathbf{d_{ij}}}{\|\mathbf{d_{ij}}\|}$.
4. Two randomly chosen point correspondences out of $\{(\bar{\mathbf{d}}_{\mathbf{11}}\ \bar{\mathbf{d}}_{\mathbf{12}}), \ldots, (\bar{\mathbf{d}}_{\mathbf{N1}}\ \bar{\mathbf{d}}_{\mathbf{N2}})\}$ are used to compute the camera center of the second camera $\mathbf{C}$ by the DRP method within RANSAC.
5. The consensus set $\mathcal{I}$ is given by a number of point correspondences that possess after triangulation and re-projection an error smaller than a threshold. Usually, this threshold is between one and three pixel.
6. If $|\mathcal{I}| > i_{max}$ then $i_{max} = |\mathcal{I}|$ and $\bar{R} = R_2$ and $\mathbf{C_2} = \mathbf{C}$.
7. Go back to 1. until $i = 4$.
8. Select only the inliers and re-compute $\mathbf{C_2}$ with the final $\bar{R}$ by using once more the DRP-method.

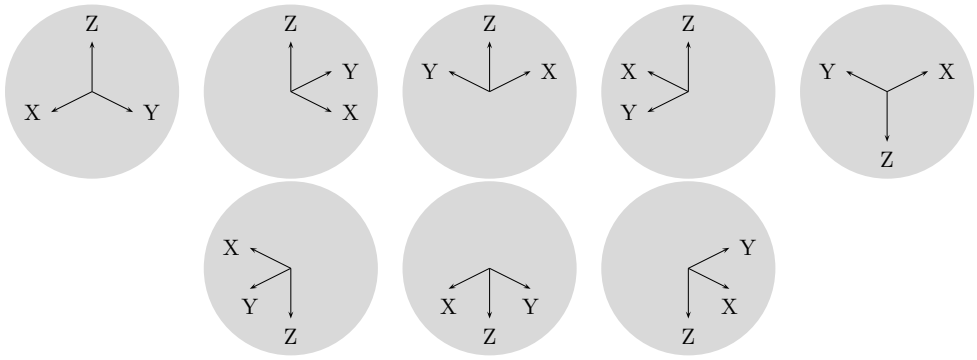


Figure 3.5: Eight rotation ambiguities from the 24 possible remain under natural pose; more precisely row one and four of Fig. 3.1.

They use a branch and bound method and search in an oct-tree which guarantees to find the global optimum. In our case only four possibilities exist, thus, L-0 computes the re-projection error for all R_2^{1-4} separately and takes the rotation with the smallest error. The DRP method alone is not robust, especially, in the case when the point correspondences are generated by the head tracks of single people walking in the field of view of the cameras. Hence, $\bar{R}$ and $\mathbf{C_2}$ are estimated within a RANSAC framework. Alg. 3.1 summarizes the steps of L-0.

3.2.3 Method L-1

Consider the case when $\bar{R}$ can take any possible rotation around the vertical Z axis. The 3×3 matrix $\bar{R}$ has the property that the second element in the first column is zero. Furthermore,

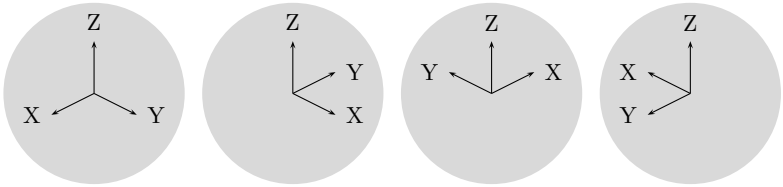


Figure 3.6: Rotating one camera in natural pose into the other is four-fold ambiguous. The possible rotations are from left to right: 0 rad, $\frac{\pi}{2}$ rad, π rad and $-\frac{\pi}{2}$ rad around the vertical coordinate axis Z.

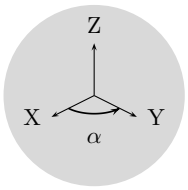


Figure 3.7: Instead of the four possible rotations in Fig. 3.6, the relaxed case allows an arbitrary rotation α around the Z axis.

the unknowns $\mathbf{C_1}$, $\mathbf{C_2}$ and $\bar{R}$ are captured in the Essential matrix

$$E = [\mathbf{t}]_\times \bar{R} \tag{3.6}$$

with $\mathbf{C_1} = \mathbf{0}$ and $\mathbf{t} = -\bar{R}\mathbf{C_2}$. Known point correspondences constrain bilinearly the Essential matrix. At least five such correspondences are necessary (Sec. 3.1.1) and as $\bar{R}$ has the aforementioned property this number reduces to two, because three correspondences are known by definition, that is,

$$\mathbf{d_{1j}} = (1\ 0\ 0)^\top \tag{3.7}$$

$$\mathbf{d_{2j}} = (0\ 1\ 0)^\top \tag{3.8}$$

$$\mathbf{d_{3j}} = (0\ 0\ 1)^\top \tag{3.9}$$

for cameras $j \in \{1, 2\}$. L-1 uses Stewenius's implementation [Henrik Stewenius and Nister, 2006] which is an improvement of Nister's 5-point algorithm [Nister, 2004]. L-1 does not work when all points and all cameras are coplanar, however, this case should not happen in practice. As in L-0, L-1 is formulated in a RANSAC framework to add robustness.

One remark has to be given: Nister's algorithm delivers at most ten real solutions of the Essential matrix wich are all equal up to a scalar whereas only one is the true Essential matrix. The true Essential matrix can be decomposed into four combinations of possible rotations and translations, however, only one is physically correct, that is, all points are in front of the two cameras. A test for the right rotation and translation given the Essential matrix is given by Hartley and Zisserman [2004]. The question how the correct Essential matrix out of the possible ones is simply answered. L-1 chooses randomly one Essential matrix candidate

Algorithm 3.2 Method L-1: Localize two cameras by using Nister's method.

Let $\mathbf{x_{ij}}$ be the image measurements of N points ($1 \leq i \leq N$) in two cameras ($j \in \{1, 2\}$). Let K_j and R_j be the camera parameters which are obtained by C-0 till C-3 (Chap. 2). Return the final rotation ambiguity $\bar{R}$ and the camera center $\mathbf{C_2}$. $\mathbf{C_1}$ is $\mathbf{0}$, R_1 is I_3. Let i_{max} in the beginning be zero.

1. All point correspondences $\mathbf{x_{ij}}$ are mapped to the plane-at-infinity by the inverse, infinite homographies $H_{\infty j}^{-1} = (K_j R_j)^{-1}$. The points are then $\mathbf{d_{ij}} = H_{\infty j}^{-1} \mathbf{x_{ij}}$.
2. Normalize $\mathbf{d_{ij}}$ spherically, that is, $\bar{\mathbf{d}}_{\mathbf{ij}} = \frac{\mathbf{d_{ij}}}{\|\mathbf{d_{ij}}\|}$.
3. Two randomly chosen point correspondences out of $\{(\bar{\mathbf{d}}_{\mathbf{11}}\ \bar{\mathbf{d}}_{\mathbf{12}}), \ldots, (\bar{\mathbf{d}}_{\mathbf{N1}}\ \bar{\mathbf{d}}_{\mathbf{N2}})\}$ are used to compute the camera center of the second camera $\mathbf{C}$ by Stewenius et al.'s method within RANSAC. Add the known point correspondences given by the orthogonal world to the $\bar{\mathbf{d}}_{\mathbf{ij}}$ to have the necessary five point correspondences (Eqn. 3.9). During the model building stage of RANSAC, test the necessary properties of a randomly chosen Essential matrix candidate as it is described in the text. The consensus set is given by a number of point correspondences that possess after triangulation and re-projection an error smaller than a threshold. Usually, this threshold is between one and three pixel.
4. Select only the inlier point correspondences and re-compute $\bar{R}$ and $\mathbf{C_2}$ by using once more the Stewenius et al.'s method.

and tries to decompose it correctly. If this is not possible the next candidate is chosen. The correct $\bar{R}$ and $\mathbf{C_2}$ is found when the re-projection error of the given point correspondences is a minimum. Alg. 3.2 summarizes the steps of L-1.

3.3 Two cameras without overlapping views

Sec. 3.1 explained possible localization approaches for cameras without overlapping views. Only Rahimi's work is able to avoid a constant velocity assumption, however, he was unable neither to relax the coplanarity assumption of all positions nor to overcome the multiple local minima of his objective function. This section presents a new way to use Rother's DRP-method for the same task, where the disadvantages of Rahimi's approach disappear, that is, on the one hand the trajectory's positions and the camera centers are simultaneously reconstructed within the 3D-space and on the other hand the localization guarantees to find a global solution without the difficulty of local minima. This guarantee is currently bought by assumptions about the possible ambiguous rotations between the cameras, however, Sec. 3.2.1 showed that a general solution without any restrictions is possible. Furthermore, our extended version of the DRP-method is computationally attractive, because it computes the solution in closed-form. Rahimi's method is in contrast iterative. Our results in the next Chap. 4 show that the extended DRP-method serves as a powerful method in overlapping and non-overlapping views; to our knowledge no such method currently exists.

3.3.1 The Direct Reference Plane method

The 3×4 homogeneous matrix P expresses a pinhole camera's projection of a real world point $\mathbf{X}$ to a point $\mathbf{x}$ on the image plane with homogeneous vector. Formally, we write

$$\lambda \mathbf{x} = P \left(\mathbf{X}\ 1\right)^{\top}, \tag{3.10}$$

where λ is the unknown projective depth of $\mathbf{x}$. A geometrically meaningful decomposition of P is the block matrix

$$P = \left[M\ \mathbf{t}\right], \tag{3.11}$$

whereas the 3×3 homogeneous matrix M is a homography induced by some reference plane and vector $\mathbf{t}$ is the translation between the camera center and the origin within the camera coordinate system.

One may assume without loss of generality that the plane-at-infinity Π_∞ is this reference plane. Then, Π_∞ induces the infinite homography $M = H_\infty$ that changes the projection matrix (Eqn. 3.11) to

$$P = H_\infty \left[I_3\ -\mathbf{C}\right]. \tag{3.12}$$

I_3 is the 3×3 identity matrix. The camera center $\mathbf{C}$ is as well as $\mathbf{t}$ a translation, but it rests now in the world coordinate system.

Plug P's decomposition (Eqn. 3.12) into the projection equation (Eqn. 3.10), that is,

$$\lambda \mathbf{x} = H_\infty \left[I_3\ -\mathbf{C}\right] \left(\mathbf{X}\ 1\right)^{\top} \tag{3.13}$$

and then multiply both sides of the equation by the inverse of H_∞. This step results in

$$\lambda \mathbf{d} = \mathbf{X} - \mathbf{C}. \tag{3.14}$$

$\mathbf{d} = H_\infty^{-1}\mathbf{x}$ is the world direction of the ray that passes through $\mathbf{C}$ and pierces the image plane at $\mathbf{x}$. Rother and Carlsson called this rectification of $\mathbf{x}$ under a known H_∞ "stabilization". The bilinear relationship between $\mathbf{X}$ and $\mathbf{C}$ (Eqn. 3.13) turns into a linear relationship during stabilization which is obvious in Eqn. 3.14. This linear relationship is the basis of the DRP-method that finds the camera centers and the world points at once in a closed-form solution [Rother, 2003].

We know since Chap. 2 that H_∞ encodes the camera's intrinsics K and the rotation R between the camera and the world. This relationship is explained by $H_\infty = KR$ and Chap. 2 gave methods to compute K and R by the methods C-0 till C-3. The reader also knows that R has an ambiguity, but let us assume at this point that this ambiguity is removed. The next Sec. 3.3.5 will show a solution how to identify the correct rotation.

Let the coordinates of the vectors in Eqn. 3.14 be

$$\lambda \left(d_1 \; d_2 \; d_3\right)^\top = \left(X_1 \; X_2 \; X_3\right)^\top - \left(C_1 \; C_2 \; C_3\right)^\top . \tag{3.15}$$

To understand the DRP-method, we express Eqn. 3.15 in an equivalent form by a homogeneous, linear equation system of the vector product

$$\lambda \mathbf{d} \times (\mathbf{X} - \mathbf{C}) = \mathbf{0} \tag{3.16}$$

with the advantage that the unknown projective depth diminishes. The equation system of the coordinates (Eqn. 3.16) is rewritten in the form

$$\begin{pmatrix} d_2(X_3 - C_3) - d_3(X_2 - C_2) \\ d_3(X_1 - C_1) - d_1(X_3 - C_3) \\ d_1(X_2 - C_2) - d_2(X_1 - C_1) \end{pmatrix} = \mathbf{0} \tag{3.17}$$

or more compactly

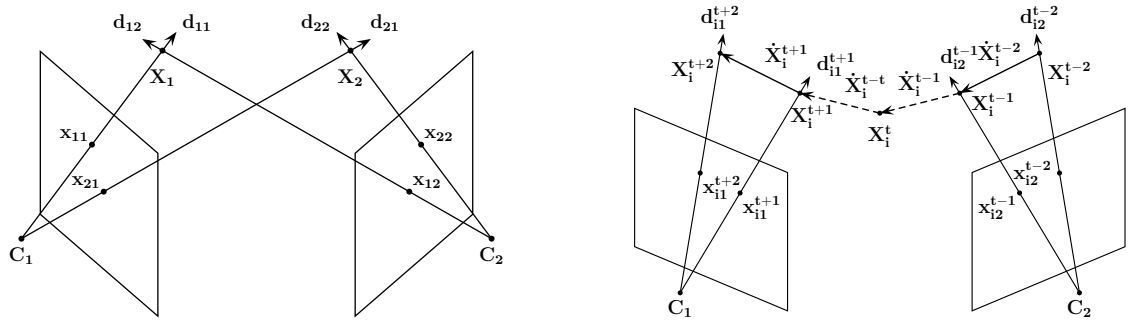
$$\left(S \;\; -S\right) \begin{pmatrix} \mathbf{X} \\ \mathbf{C} \end{pmatrix} = \mathbf{0}, \tag{3.18}$$

whereas matrix S is skew-symmetric with $S = [\mathbf{d}]_\times$. Matrix $(S \;\; -S)$ is entirely formed by the known directions and all elements of vector $(\mathbf{X} \; \mathbf{C})^\top$ are the unknown parameters.

One camera and one world point make no sense, therefore, we consider two cameras and N world points $\mathbf{X_1}, \ldots, \mathbf{X_N}$ which let us reformulate the projection equations (Eqn. 3.10) by

$$\begin{aligned} \lambda_{i1}\mathbf{x_{i1}} &= P_1 \left(\mathbf{X_i} \; 1\right)^\top && (3.19) \\ &= H_{\infty 1} \left[I_3 \;\; -\mathbf{C_1}\right] \left(\mathbf{X_i} \; 1\right)^\top && (3.20) \\ \lambda_{i2}\mathbf{x_{i2}} &= P_2 \left(\mathbf{X_i} \; 1\right)^\top && (3.21) \\ &= H_{\infty 2} \left[I_3 \;\; -\mathbf{C_2}\right] \left(\mathbf{X_i} \; 1\right)^\top && (3.22) \end{aligned}$$

with $1 \leq i \leq N$. After stabilization with the known intrinsics K_1, K_2 and rotation matrices R_1, R_2 and thus known infinite homographies $H_1^\infty = K_1 R_1$, $H_2^\infty = K_2 R_2$, the linear equation



(a) Overlap (b) Without overlap

Figure 3.8: A graphical illustration of the geometry in cameras (a) with overlapping views and (b) without overlapping views. The minimal configuration in the overlapping case are two static points visible in two views. In the non-overlapping case the minimal configuration are four successive positions of a dynamic point that moves through the views of two cameras, whereas two consecutive positions are visible in one camera and the other two positions are visible in the other camera. Note that the number of unseen positions is arbitrary for the computation, however, the more points are unseen the worse the reconstruction will be.

system (Eqn. 3.18) takes the generalized form

$$A\mathbf{h} = \begin{pmatrix} S_{11} & 0 & \cdots & 0 & 0 \\ S_{12} & 0 & \cdots & 0 & -S_{12} \\ 0 & S_{21} & \cdots & 0 & 0 \\ 0 & S_{22} & \cdots & 0 & -S_{22} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & S_{N1} & 0 \\ 0 & 0 & \cdots & S_{N2} & -S_{N2} \end{pmatrix} \begin{pmatrix} \mathbf{X_1} \\ \vdots \\ \mathbf{X_N} \\ \mathbf{C_2} \end{pmatrix} = \mathbf{0}. \tag{3.23}$$

Fig. 3.8a illustrates the geometry for the two point case. To fixe the gauge, we choose without any restrictions $\mathbf{C_1}$ as the origin of the world coordinate system, that is, $\mathbf{C_1} = \mathbf{0}$ and let $R_1 = I_3$, hence, $H_{\infty 1} = K_1$. The matrices S_{ij}, $1 \leq i \leq N$, $j \in \{1, 2\}$, are given by $S_{ij} = [\mathbf{d_{ij}}]_\times$ with stabilized image points $\mathbf{d_{ij}}$.

The camera center $\mathbf{C_2}$ and the world points are the unknowns and they form the solution vector $\mathbf{h}$ which gives the smallest residuum $\|A\mathbf{h}\|$ under the condition $\|\mathbf{h}\| = 1$. Remember that we usually have more equations than unknowns, hence, an exact solution is inexistent; the optimal solution is the one that fulfills $\min_\mathbf{h} \|A\mathbf{h}\|$. This optimization is common in Computer Vision and and is usually solved by SVD [Golub and van Loan, 1996] of matrix $A = USV^\top$. It is the vector of the right nullspace of A with the smallest associated singular value, that is, the last column of the orthogonal matrix V. Note that $A_{m\times n}$ has theoretically rank $r = \dim(\mathbf{h}) - 1$, however, noisy stabilized image points cause A to be of full rank. The rank reduction has its reason in the homogeneous representation of the geometry. A family of

solution vectors $\lambda\mathbf{h}$ with $\lambda \in \mathbb{R}$ remains; therefore the further condition $\|\mathbf{h}\| = 1$. The reader might argue that an efficient computation is quickly intractable for many and long trajectories. Unfortunately, we are not able to debilitate this argument at the moment, however, some actions can be done to solve this problem. A full computation of U, S and V is expensive. It has time complexity of $\mathrm{O}(mnr)$ and space complexity of $\mathrm{O}((m+n)r)$ [Brand, 2002], hence, we used a more efficient implementation of SVD that computes only the last column of V; see Matlab's `svds`. With 200 trajectories each 20 positions long, the computation reduces then to a couple of minutes. A final solution might be an incremental approach as it is proposed by Brand [Brand, 2002].

A great advantage of the DRP-method an in contrast to Projective Factorization [Triggs, 2000a] is its natural handling of occluded or invisible image points, that is, not all images of the world points must exist. Rother showed that two distinct world points in two cameras are necessary and sufficient for a non-trivial solution of Eqn. 3.23, because only two of the three equations given by $\mathbf{d_{ij}}$ are independent; $\mathbf{x_{ij}}$ is a homogeneous vector. More precisely, a unique solution will exist if at least the number of independent equations is equal to the number of unknowns minus the overall scale, that is,

$$4N \geq 3(N+1) - 1. \tag{3.24}$$

This inequality is fulfilled for $N \geq 2$.

Furthermore, these points are not allowed to be coplanar with the camera centers which is pointed out as critical configuration by Rother. Compared to features that are tracked on people, the height of the mounted cameras is usually different, hence, a critical configuration should never happen in practice. Note that the difficult separation of points on and off the reference plane no longer has any effect when Π is the reference plane, because all measured points on the image plane will be images of finite points in the world. Much more details about the DRP-method are in [Rother, 2003].

3.3.2 Dynamic instead of static points

The previous section showed that at least two points in two views are necessary and sufficient to solve the reconstruction problem. Unfortunately, this result becomes invalid when the views of two cameras are non-overlapping.

Result 3.3. *The reconstruction problem for two cameras with non-overlapping views is unsolvable by using the DRP-method.*

Proof. Intuitively, we know that a solution is inexistent, because at no time instant image point correspondences between the two views are available, that is, a world point delivers at

most two instead of four independent equations. The left side of Inequality 3.24 changes to $2N$, resulting in

$$\begin{aligned} 2N &\geq 3(N+1)-1 && (3.25)\\ N &\leq -4. && (3.26) \end{aligned}$$

The result would be false, if $N \leq -4$ is satisfied by an N with $N \geq 0$. As no such N exists, the result must be true. □

Nevertheless, we will show until the end of this section that the DRP-method works in non-overlapping views, however, the idea is to assume dynamic points instead of static ones. Fig. 3.8b shows an illustration of the problem.

Similar to Rahimi et al. [2004], we model the dynamic behavior as a linear Gaussian random walk. The motion model of a point at time instant t is a state $\mathbf{s^t} = \left(\mathbf{X^t}\ \dot{\mathbf{X}}^\mathbf{t}\right)^\top$ with the point's 3D position $\mathbf{X^t}$ and the point's velocity $\dot{\mathbf{X}}^\mathbf{t}$ in all three directions. The state's random evolution from $\mathbf{s^t}$ to $\mathbf{s^{t+1}}$ is given by

$$\mathbf{s^{t+1}} = T_s\mathbf{s^t} + \eta_\mathbf{t}. \tag{3.27}$$

Matrix

$$T_s = \begin{bmatrix} I_3 & I_3 \\ 0 & I_3 \end{bmatrix} \tag{3.28}$$

is a linear transformation that adds for the new position the current velocity to the current position and keeps the velocity unchanged. At each time instant the motion model mimics a Markov process by adding a Gaussian noise to the new position and to the velocity. The time-varying random variable $\eta_\mathbf{t}$ is therefore zero-mean Gaussian with covariance $\sigma_\eta^\mathbf{2} I_6$. We consider a covariance matrix with zero off-diagonal elements and assume constant variances for the position and the velocity.

In the following we will show that this additional motion model will allow us to add a sufficient number of equations to the system in Eqn. 3.23, so that the number of equations becomes equal to the number of unknowns and the system is solvable. As we consider now the trajectories $\mathbf{X_i^1}, \ldots, \mathbf{X_i^{T_i}}$ $(1 \leq i \leq N)$ of N moving points, the static points of solution vector $\mathbf{h}$ are replaced by the positions of the trajectory, that is,

$$\mathbf{h} = \left(\mathbf{X_1^1}\ \ldots\ \mathbf{X_1^{T_1}}\ \cdots\ \mathbf{X_N^1}\ \ldots\ \mathbf{X_N^{T_N}}\ \mathbf{C_2}\right)^\top. \tag{3.29}$$

For each position, $\mathbf{h}$ is further extended by the velocities $\dot{\mathbf{X}}_\mathbf{i}^\mathbf{t}$ to

$$\mathbf{h} = \left(\mathbf{X_1^1}\ \ldots\ \mathbf{X_1^{T_1}}\ \cdots\ \mathbf{X_N^1}\ \ldots\ \mathbf{X_N^{T_N}}\ \dot{\mathbf{X}}_\mathbf{1}^\mathbf{1}\ \ldots\ \dot{\mathbf{X}}_\mathbf{1}^{\mathbf{T_{N-1}}}\ \cdots\ \dot{\mathbf{X}}_\mathbf{N}^\mathbf{1}\ \ldots\ \dot{\mathbf{X}}_\mathbf{N}^{\mathbf{T_{N-1}}}\ \mathbf{C_2}\right)^\top. \tag{3.30}$$

These velocities are slack variables that depend only on $\mathbf{X_i^t}$ and $\mathbf{X_{i+1}^t}$ by the difference

$$\dot{\mathbf{X}}_\mathbf{i}^\mathbf{t} = \mathbf{X}_\mathbf{i}^\mathbf{t+1} - \mathbf{X}_\mathbf{i}^\mathbf{t}. \tag{3.31}$$

Although, these slack variables give no further information, it is now possible to constrain the velocities by adding further equations. On the one hand, this linear dependency (Eqn. 3.31) gives three independent equations for all $N-1$ velocities of a trajectory i, that is, $3(N-1)$ equations, that are collected by matrix

$$C_i = \begin{bmatrix}
_1 & 0 & & 0 & I_3 & -I_3 & 0 & \cdots & 0 & 0 & & 0 & I_3 & 0 & \cdots & 0 & 0 & & 0 & 0 \\
_2 & 0 & & 0 & 0 & I_3 & -I_3 & \cdots & 0 & 0 & & 0 & 0 & I_3 & \cdots & 0 & 0 & & 0 & 0 \\
 & \vdots & \overset{c_1}{\cdots} & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \overset{c_2}{\cdots} & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \overset{c_3}{\cdots} & \vdots & \vdots \\
_{T_i} & 0 & & 0 & 0 & 0 & \cdots & I_3 & -I_3 & 0 & & 0 & 0 & 0 & \cdots & I_3 & 0 & & 0 & 0
\end{bmatrix}. \tag{3.32}$$

All elements of C_i that constrain variables concerning other trajectories except i are zero, that is, zeros between the block matrices with $c_1 = \sum_{j=1}^{i-1} T_j$, $c_2 = \sum_{j=i+1}^{N} T_j + \sum_{j=1}^{i-1}(T_j - 1)$ and $c_3 = \sum_{j=i+1}^{N}(T_j - 1)$. On the other hand, and this is the novelty in contrast to Rother's DRP-method, the Markov assumption

$$\dot{\mathbf{X}}_\mathbf{i}^\mathbf{t+1} = \dot{\mathbf{X}}_\mathbf{i}^\mathbf{t} + \eta_\mathbf{t} \tag{3.33}$$

between two consecutive velocities allows to add further three independent equations that are stacked together in the same way as for C_i to matrix

$$D_i = \begin{bmatrix}
_1 & 0 & & 0 & I_3 & -I_3 & 0 & \cdots & 0 & 0 & & 0 & 0 \\
_2 & 0 & & 0 & 0 & I_3 & -I_3 & \cdots & 0 & 0 & & 0 & 0 \\
 & \vdots & \overset{d_1}{\cdots} & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \overset{d_2}{\cdots} & \vdots & \vdots \\
_{T_i} & 0 & & 0 & 0 & 0 & \cdots & I_3 & -I_3 & 0 & & 0 & 0
\end{bmatrix} \tag{3.34}$$

with $d_1 = \sum_{j=1}^{N} T_j + \sum_{j=1}^{i-1}(T_j - 1)$ and $d_2 = \sum_{j=i+1}^{N}(T_j - 1)$. With these constraints we assume similarly to Rahimi at most smooth trajectories and not constant and equal magnitudes of the velocities as it is in the work of Javed et al. [2003]. It is clear that the Markov assumption as it is formulated in Eqn. 3.31 is not exactly satisfiable, because velocities will change from one position to the next. Smoothness, however, means formally that the Markov assumption should be fulfilled as best as possible, that is, $\|\dot{\mathbf{X}}_\mathbf{i}^\mathbf{t+1} - \dot{\mathbf{X}}_\mathbf{i}^\mathbf{t}\| \rightarrow \min$ and this is what the DRP-method does when more equations than unknowns are available.

Finally, the images of points in one of the two cameras deliver the last necessary equations

which form matrix

$$B_i = \begin{bmatrix} {\scriptstyle 1} & 0 & & 0 & S_{i1}^1 & 0 & \cdots & 0 & 0 & & 0 & 0 \\ {\scriptstyle 2} & 0 & & 0 & S_{i2}^1 & 0 & \cdots & 0 & 0 & & 0 & -S_{i2}^1 \\ {\scriptstyle 3} & 0 & & 0 & 0 & S_{i1}^2 & \cdots & 0 & 0 & & 0 & 0 \\ {\scriptstyle 4} & 0 & & 0 & 0 & S_{i2}^2 & \cdots & 0 & 0 & & 0 & -S_{i2}^2 \\ & \vdots & \overset{a_1}{\cdots} & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \overset{a_2}{\cdots} & \vdots & \vdots \\ {\scriptstyle 2T_i-1} & 0 & & 0 & 0 & 0 & \cdots & S_{i1}^{T_i} & 0 & & 0 & 0 \\ {\scriptstyle 2T_i} & 0 & & 0 & 0 & 0 & \cdots & S_{i2}^{T_i} & 0 & & 0 & -S_{i2}^{T_i} \end{bmatrix}. \tag{3.35}$$

with $a_1 = \sum_{j=1}^{i-1} T_j$ and $a_2 = \sum_{j=i+1}^{N} T_j + \sum_{j=1}^{N}(T_j - 1)$. S_{ij}^t are skew-symmetric matrices formed by the stabilized image positions d_{ij}^t.

The trajectory's matrices B_i, C_i and D_i $(1 \leq i \leq N)$ compose a new system matrix

$$A' = \begin{bmatrix} B_1 & \cdots & B_N & C_1 & \cdots & C_N & D_1 & \cdots & D_N \end{bmatrix}^\top \tag{3.36}$$

that replaces A and substitutes Eqn. 3.23 with the extended solution vector (Eqn. 3.30) by

$$A' \left(\mathbf{X_1^1} \ \ldots \ \mathbf{X_1^{T_1}} \ \cdots \ \mathbf{X_N^1} \ \ldots \ \mathbf{X_N^{T_N}} \ \dot{\mathbf{X}}_\mathbf{1}^\mathbf{1} \ \ldots \ \dot{\mathbf{X}}_\mathbf{1}^{\mathbf{T_{N-1}}} \ \cdots \ \dot{\mathbf{X}}_\mathbf{N}^\mathbf{1} \ \ldots \ \dot{\mathbf{X}}_\mathbf{N}^{\mathbf{T_{N-1}}} \ \mathbf{C_2} \right)^\top = \mathbf{0}. \tag{3.37}$$

Again, the idea is to find a solution that gives minimal residuum $\min_{\mathbf{h}} \|A'\mathbf{h}\|$. Solving Eqn. 3.37 is passing a smooth trajectory into the 3D-space where each position and the camera centers are constrained by the rays that pierce the corresponding image points. The more positions are visible, the more rays are present, the better the reconstruction problem is constrained. At least two rays of two consecutive positions in each view must be present, because otherwise the velocity between these positions is unconstrained. All other unseen positions are sufficiently constrained by these two velocity vectors and the Markov assumption. An exact solution for the minimal configuration exists only for trajectories that have constant and equal velocities. As we assume not more than smoothness, such an exact solution will never exists. The reader should imagine SVD as minimizer that on the one hand bends and moves the reconstructed trajectories and on the other hand localizes the cameras until an optimal configuration is reached given the imaged positions and the Markov assumption. In the next section we will further investigate how such a minimal configuration looks like.

3.3.3 Minimal and critical configurations

Matrix A encapsulates all information about trajectories and image points, however, it is still unclear how long a trajectory at least has to be and how many positions of a particular trajectory have at least to be visible in each of the two views. We answer these questions by

presenting the minimal configuration of points that are necessary and sufficient for a unique reconstruction. The minimal configuration is the minimal number of points and the minimal number of their visible images in both views that produce a sufficient number of independent rows in A. To simplify the search for a minimal configuration, we consider only one trajectory $\mathbf{X^1}, \ldots, \mathbf{X^T}$ instead of N trajectories which is possible without restricting the reconstruction, because the temporal coherence among several trajectories has no influence.

The trajectory's positions are invisible in both views or at most visible in one of the views which makes the assessment if a particular configuration of points will allow a unique reconstruction, a difficult problem. In order to ease and to formalize the problem, Rother introduced a binary visibility matrix V, where each column is a point and each row is a camera. An element V_{ij} is true, if the image of point $\mathbf{X_j}$ is visible in camera i, otherwise V_{ij} is false.

The number of unknowns in $\mathbf{h}$ with a single trajectory (Eqn. 3.30) is

$$N_u = 3(T + (T - 1) + 1). \tag{3.38}$$

$\mathbf{h}$ consists of $3T$ positions, $3(T-1)$ velocity unknowns and three variables for the second camera center. All these unknowns are coordinates in the world coordinate system. The number of rows (equations) in A is

$$N_e = 2|V_{ij}| + 3(T - 1) + 3(T - 2). \tag{3.39}$$

Each image point i, imaged in camera j, sets element V_{ij} true and gives exactly two independent equations in B. Hence, two times the cardinality of V_{ij} is exactly the number of rows in B. Similar to B, the three independent equations formed by each of the $T-1$ velocities produce all rows in C and the $3(T-2)$ independent equations between two temporal consecutive velocities given by the Markov assumption produce exactly the number of rows in D.

Now, we want to identify the minimal configuration which is certainly the configuration with the smallest T and $|V_{ij}|$. A lower bound is given by Rother's result for overlapping views, that is, two points ($T \geq 2$) with their two images in two views ($|V_{ij}| \geq 4$) are a minimal configuration. Hence, we are looking for a minimal assignment of T and $|V_{ij}|$ which still satisfies inequality

$$N_e \geq N_u - 1. \tag{3.40}$$

The overall scale causes the reduction of the degrees of freedom (N_u) by one (right side of Eqn. 3.40), that is, as all point images are homogeneous vectors, the non-trivial solution in the right nullspace of $A\mathbf{h} = \mathbf{0}$ is a family of vectors $\lambda\mathbf{h}$ with $\lambda \in \mathbb{R}$. This leads to the following result:

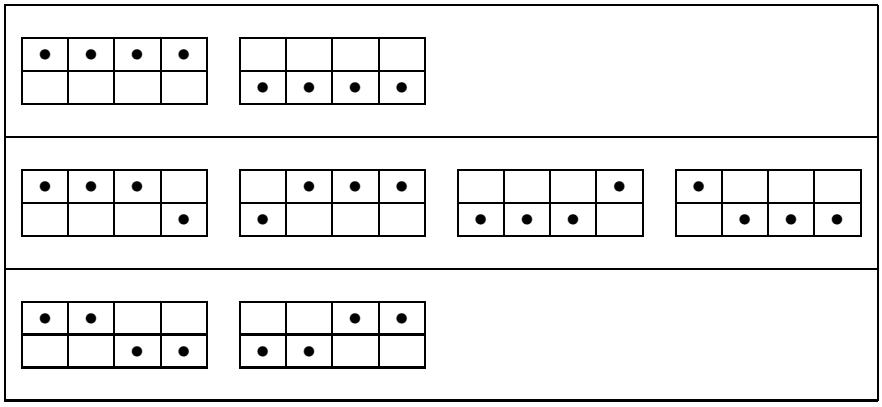


Figure 3.9: Potential minimal configurations. The first row illustrates all observations only in one view which yield no motion parallax. The second row give all combinations of imaged positions where in one view only one observation is available. These configurations all collapse to a meaningless geometry. The third row shows the only remaining configuration which is the correct minimal configuration; two observations in two views.

Result 3.4. *The conditions $T \geq 4$ and $|V_{ij}| \geq 4$ are necessary that the reconstruction problem for two cameras with non-overlapping views is solvable by using our extended DRP-method.*

Proof. We substitute N_e and N_u in Eqn. 3.40 by the Eqn. 3.39 and Eqn. 3.38 and simplify, that is,

$$\begin{aligned} 2|V_{ij}| + 3(T-1) + 3(T-2) &\geq 3(T + (T-1) + 1) - 1 && (3.41)\\ 2|V_{ij}| + 6T - 9 &\geq 6T - 1 && (3.42)\\ 2|V_{ij}| &\geq 8 && (3.43)\\ |V_{ij}| &\geq 4. && (3.44) \end{aligned}$$

The minimal value for $|V_{ij}|$ is four under condition $|V_{ij}| \geq 4$. As each point has at most one image, T is also at least four. □

Inequality 3.44 shows obviously the dependency between the velocities and the image points in that a velocity adds exactly three independent equations. Each new perhaps unobserved position is thus constrained by the Markov equations. Hence, the length of a trajectory as long as it is larger or equal to four is insignificant. This is an important consequence of result 3.4. The same is valid for the number of unobserved positions. As long as four positions are projected into one of the camera's images, a valid, non-trivial solution for Eqn. 3.37 exists. However, only one configuration of four possible, minimal configurations with $|V_{ij}| = 4$ makes geometrically sense which has the following consequence:

Result 3.5. *The conditions $V_{1j} \geq 2$, $V_{2j} \geq 2$ are sufficient that the reconstruction problem for two cameras with non-overlapping views is solvable by using our extended DRP-method.*

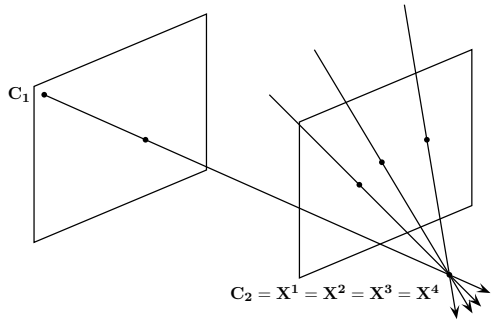


Figure 3.10: Illustration of the configurations that collapse to a meaningless geometry. The reader can see that $\mathbf{C_2}$ and all reconstructed positions of the trajectory are a single point. This is possible, because only one observation of a position in Camera 1 is available.

Proof. Fig. 3.9 illustrates the three possible configurations. All images in only one view prohibit a motion parallax, hence, it is an undesirable configuration. Exactly one image in one of the views is correct, however, the solution is geometrically meaningless. All positions collapse in a single point which is $\mathbf{C_2}$. Fig. 3.10 illustrates this case. As only one potential configuration remains (the last row of Fig. 3.9), it is the one that is desirable. □

Apart from the minimal configuration the question about critical configurations arises which are minimal configurations with linear dependencies among the rows of A'. One way to detect critical configurations is rank deficiency in A', that is,

$$\begin{aligned} \operatorname{rank}\left(A'\right) &< N_u^{1^{st}\text{trajectory}} + \ldots + N_u^{N^{th}\text{trajectory}} \\ &< 3\sum_{j=1}^{N} T_j + 3\sum_{j=1}^{N}(T_j - 1) + 3 - 1. \end{aligned} \tag{3.45}$$

Rother showed that critical configurations exist even when visible points are sufficiently available. He identified two cases for two views. On the one hand, when the two point lie on the reference plane and on the other hand, when the camera centers and the two points are all coplanar. The first case is irrelevant, because the reference plane is the plane-at-infinity and all measured points are assumed to be finite. The second case of a critical configuration still remains without overlapping views, because the coplanar dependencies between points and cameras will as in the overlapping case force all sub-determinants of $B_1, \ldots, B_N$ to be zero. The detailed proof is by Rother [2003].

3.3.4 Geometric analysis of the objective function

Rother mentioned that the objective function $\|A\mathbf{h}\|$ is an algebraic error and this disadvantage remains also for the extended DRP-method. The objective function $\|A'\mathbf{h}\|$ that is minimized

by using SVD is expressible by

$$\|A'\mathbf{h}\| = \sum_{i=1}^{N}\sum_{t=1}^{T_i} \underbrace{\|\mathbf{d_{i1}^t} \times (\mathbf{d_{i2}^t} - \mathbf{C_2})\|}_{\text{stab. image points}} + \underbrace{\|\mathbf{X_i^t} - \mathbf{X_i^{t-1}} - \dot{\mathbf{X}}_\mathbf{i}^{\mathbf{t-1}}\|}_{\text{velocities}} + \underbrace{\|\dot{\mathbf{X}}_\mathbf{i}^\mathbf{t} - \dot{\mathbf{X}}_\mathbf{i}^{\mathbf{t-1}}\|}_{\text{smoothness}} + \mathbf{C_1}. \quad (3.46)$$

We see that the error term given by the image points has no physical meaning, although it has a geometric meaning, because the norm of the vector product is clearly the length of the vector that is orthogonal to the rays passing through $\mathbf{d_{i1}^t}$ and $\mathbf{d_{i2}^t} - \mathbf{C_2}$. An alternative could be the term

$$\arccos \frac{\|\mathbf{d_{i1}^t} \times (\mathbf{d_{i2}^t} - \mathbf{C_2})\|}{\|\mathbf{d_{i1}^t}\|\|\mathbf{d_{i2}^t} - \mathbf{C_2}\|}, \quad (3.47)$$

which represents the angle between the rays, but needs further research. Rahimi et al. [2004] used instead the re-projection error which is a non-linear function, thus, he was not able to give a closed-form solution but used an iterative Newton-Raphson optimizer. The other two terms representing the velocities and the Markov assumption between two successive velocities are physically meaningful and are directly comparable to the motion terms in Rahimi's objective function.

3.3.5 Rotation estimation

Sec. 3.2 described a coordinate ascent strategy that accomplishes a camera localization and a simultaneous rotation estimation. The same strategy is also valid for cameras with non-overlapping views. One of the four possible rotation matrices $\bar{R}_1, \ldots, \bar{R}_4$ between the two cameras is kept as current estimation, for example matrix $\bar{R}_1$, whereas the world coordinate system lies in Camera 1. Then, all stabilized image points $\mathbf{d_{i1}^t}$ in Camera 1 are rectified by $\bar{R}_1^{-1}\mathbf{d_{i1}^t}$ and the camera localization as it is described in the previous section is performed.

Once again, the correct rotation matrix is the one that gives a minimal re-projection error of all visible trajectory's positions. More precisely, the correct $\bar{R}$ is given by

$$\bar{R} = \underset{\bar{R}\in\{\bar{R}_1,\ldots,\bar{R}_4\}}{\operatorname{argmin}} \sum_{i\in\mathcal{N}_1}\sum_{t\in\mathcal{T}_1^i} \|\mathbf{x_{i1}^t} - K_1 R_1 \bar{R}\,[I_3 \; -\mathbf{C_1}]\left(\mathbf{X_i^t}\;1\right)^\top\|^2 + \\ + \sum_{i\in\mathcal{N}_2}\sum_{t\in\mathcal{T}_2^i} \|\mathbf{x_{i2}^t} - P_2\left(\mathbf{X_i^t}\;1\right)^\top\|^2, \quad (3.48)$$

whereas $\mathcal{N}_1$ and $\mathcal{N}_2$ are index sets containing indices of all trajectories that are at least partly visible in Camera 1 and Camera 2. Similarly, $\mathcal{T}_1^i$ and $\mathcal{T}_2^i$ contain the indices of the visible positions within a particular trajectory i.

The camera matrices $P_1 = K_1 R_1 \bar{R}\,[I_3 \; -\mathbf{C_1}]$ and $P_2 = K_2 R_2\,[I_3 \; -\mathbf{C_2}]$ are the result of the camera localization and they are necessary for the re-projection of the reconstructed

positions. We can do this process step by step for each rotation matrix and evaluate in the end the re-projection error or we implement the estimation of the rotation matrix directly within the RANSAC framework which is discussed in the next section. K_1, R_1, K_2 and R_2 are the result of camera calibration which is discussed in Chap. 2.

Note that the re-projection error is computed in the image plane and not on the plane-at-infinity, that is, the Euclidian distance between the re-projected point and the stabilized point might be computed. However, even for visible positions infinite large distances are possible which makes this space inconvenient for distance measures.

3.3.6 Robust localization

The solution of the linear equation system (Eqn. 3.37) is sensible to random changes in the entries of the system matrix. Outliers of the underlying feature detectors will cause such changes, thus, the accuracy of the solution depends heavily on a correct detection of these outliers. One way would be to weight the equations in B_i (Eqn. 3.35) by the relative reliability of the image positions that generate the particular equations. High weights are assigned to measurements that are with high probability inliers, while low weights indicate outliers. B_i would then change to a weighted matrix $B_i^\top W B_i$. Matrix W is most common a diagonal matrix with the diagonal elements equal to the corresponding weights. However, the question that arises immediately is how one should determine these weights. Although this approach is well suited to handle noisy elements that follow a probabilistic model, it is improper to handle gross outliers. Therefore, we did not follow this approach. Instead we choose RANSAC which is in many other Computer Vision problems the most successful way to detect outliers given a mathematical model, in our case the perspective projection of the cameras and the motion model of the moving points.

The outline of the robust localization is as follows: Given N trajectories of a moving object, in each iteration RANSAC chooses randomly exactly one trajectory. This trajectory is used to form the system matrix A' and the DRP-method computes a reconstruction for the positions of this trajectory and for the camera centers. Then RANSAC determines the consensus set by testing each single trajectory. Testing means that the re-projection error of visible positions must be smaller that a specific threshold t. Usually, the people use $t = 3\,\text{pixel}$ as a realistic threshold. As the reconstruction of positions is only done for one trajectory, a reconstruction of the positions of all other $N - 1$ trajectories has to be done, before a re-projection can happen. This situation leads to an interesting triangulation problem with cameras having non-overlapping views.

The problem formulation is as follows: Let $\mathbf{x_1^i}$ with $i \in \mathcal{T}_1$ and $\mathbf{x_2^j}$ with $j \in \mathcal{T}_2$ be the known point images in both views. Let P_1 and P_2 be the known camera matrices. The question is

how to compute the unknown positions $\mathbf{X^1}, \ldots, \mathbf{X^T}$ with this information in non-overlapping views. Metaphorically spoken, triangulation in overlapping views traces back the rays passing through the measurements and estimates their eventual intersection points in space. This approach is not possible in non-overlapping views, because no pair of rays exist that could intersect. We escape this dilemma again by using the motion model, consequently, the DRP-method serves also as a solution for this triangulation problem.

Our intension is to construct a solution vector $\mathbf{h}$ that contains only the unknown positions and their dynamics, that is, we eliminate $\mathbf{C_2}$ from the original $\mathbf{h}$ (Eqn. 3.30) which gives a new solution vector

$$\mathbf{h} = \left(\mathbf{X_1^1} \ \ldots \ \mathbf{X_1^{T_1}} \ \cdots \ \mathbf{X_N^1} \ \ldots \ \mathbf{X_N^{T_N}} \ \dot{\mathbf{X}}_\mathbf{1}^\mathbf{1} \ \ldots \ \dot{\mathbf{X}}_\mathbf{1}^{\mathbf{T_{N-1}}} \ \cdots \ \dot{\mathbf{X}}_\mathbf{N}^\mathbf{1} \ \ldots \ \dot{\mathbf{X}}_\mathbf{N}^{\mathbf{T_{N-1}}}\right)^\top. \tag{3.49}$$

As $\mathbf{C_2}$ is already known, we can summarize the last three columns of A' to a vector

$$\mathbf{b} = A' \left(0 \ \cdots \ 0 \ \mathbf{C_2}^\top\right)^\top. \tag{3.50}$$

The homogeneous equation system (Eqn. 3.37) turns now to an inhomogeneous equation system

$$A''\mathbf{h} = -\mathbf{b}, \tag{3.51}$$

whereas A'' is A' reduced by the last three columns. Two measurements in each view are sufficient (Result 3.5) that A'' will have at least full rank minus one but mostly full rank according to the noise, thus, a simple solution is

$$\mathbf{h} = -A''^{+}\mathbf{b} \tag{3.52}$$

with the pseudo-inverse A''^{+} of A'' [Hartley and Zisserman, 2004]. After knowing $\mathbf{h}$ it is simple to extract the reconstructed positions and to re-project them to evaluate the trajectory as potential outlier trajectory.

When more than one object is present in the tracking data, then the unsolved correspondence of objects between the non-overlapping views prohibits a straightforward use of the DRP-method. However, similar to the trivial matches (Sec. 3.2), we can construct trivial trajectories where it is guaranteed that after consecutive measurements of an object in the first view and no measurements in the second a time without any measurements in both views decays. After this time the order is vice versa, that is, consecutive measurements in the second camera and no measurements in the first. The reader can imagine the process of constructing such trivial trajectories by simply sweeping a window of fixed size, for example 2×20, over each view's measurements which are delivered by the synchronized detectors and which are correctly timely ordered. If this window is too small then we could miss

Algorithm 3.3 Method L-2: Localize two cameras without overlapping views.

Let $\mathbf{x_{ij}^t}$ be the visible image positions at time instant $t \in \mathcal{T}_i$ of N trajectories ($1 \leq i \leq N$) in two cameras ($j \in \{1, 2\}$). $\mathcal{T}_j$ is the trajectory i's index set of successive visible positions. Let K_j and R_j be the camera parameters which are obtained by C-0 till C-3 (Chap. 2). Return the final rotation ambiguity $\bar{R}$ and the camera centers $\mathbf{C_j}$.

1. Extract the trivial trajectories $\mathbf{x_{kj}^t}$, $k \in \mathcal{K}$ from the N trajectories (Sec. 3.3.6). $\mathcal{K}$ is an index subset of all $1, \ldots, N$ trajectories.
2. Stabilize the $\mathbf{x_{kj}^t}$ with the inverse, infinite homographies $H_{\infty j}^{-1} = (K_j R_j)^{-1}$. The stabilized image positions are then $\mathbf{d_{kj}^t} = H_{\infty j}^{-1} \mathbf{x_{kj}^t}$.
3. Normalize $\mathbf{d_{kj}^t}$ spherically, that is, $\bar{\mathbf{d}}_{\mathbf{kj}}^{\mathbf{t}} = \frac{\mathbf{d_{kj}^t}}{\|\mathbf{d_{kj}^t}\|}$.
4. Perform the extended DRP-method (Sec. 3.3.2) in a RANSAC framework (Sec. 3.3.6) with the $\bar{\mathbf{d}}_{\mathbf{kj}}^{\mathbf{t}}$. One $\bar{\mathbf{d}}_{\mathbf{kj}}^{\mathbf{t}}$ of the available $|\mathcal{K}|$ trajectories and $\bar{R}$ out of $\bar{R}_1, \ldots, \bar{R}_4$ are randomly chosen in the model selection step. The reconstruction is then computed by solving Eqn. 3.37. The reprojection error is computed by using the camera matrices $P_1 = K_1 R_1 \bar{R} [I_3 \ -\mathbf{C_1}]$ and $P_2 = K_2 R_2 [I_3 \ -\mathbf{C_2}]$, the current estimates $\bar{R}$, $\mathbf{C_j}$ and the reconstructed positions $\mathbf{X_k^t}$. The result are the final rotation ambiguity $\bar{R}$, the camera centers $\mathbf{C_j}$ and the inlier trajectories $\mathbf{x_{lj}^t}$, $l \in \mathcal{L} \subset \mathcal{K}$.
5. Re-stabilize the inliers in Camera 1 $\mathbf{x_{l1}^t}$ with the updated inverse, infinite homography $H_{\infty 1}^{-1} = (K_1 R_1 \bar{R})^{-1}$, that is, $\mathbf{d_{l1}^t} = H_{\infty 1}^{-1} \mathbf{x_{l1}^t}$.
6. Re-normalize $\mathbf{d_{lj}^t}$ spherically, that is, $\bar{\mathbf{d}}_{\mathbf{lj}}^{\mathbf{t}} = \frac{\mathbf{d_{lj}^t}}{\|\mathbf{d_{lj}^t}\|}$.
7. Re-compute the camera centers by using the extended DRP-method with $\bar{\mathbf{d}}_{\mathbf{lj}}^{\mathbf{t}}$.

some trivial trajectories and if it is too large, perhaps, no trajectory can be found. So this parameter is delicate and depends on the part of the trajectory that is unseen from both views. This length is given by the framerate, the speed of the objects and the distance of the gap between the cameras. This structure of a trivial trajectory is necessary for a correct correspondence, but of course it is hypothetical. We argue that if the trajectory is caused by a wrong correspondence then it will be detected by RANSAC as an outlier.

To obtain a robust camera localization, the camera centers are recomputed by the DRP-method with the consensus set, that is, A' is build solely from stabilized image positions of trajectories that are in the consensus set.

Alg. 3.3 summarizes the steps of the final camera localization.

Chapter 4

Experimental results

This chapter discusses the achieved accuracy of the approach under realistic conditions by experiments with synthetic and real image data. We acquired two challenging, real image datasets and use the well known PETS 2006 dataset for further comparison; more details are presented in Sec. 4.1. Sec. 4.2 presents the results of the first step - the single camera calibration and the next two sections treat the second step - the localization of two cameras with overlapping fields of view (Sec. 4.3) and additionally non-overlapping fields of view (Sec. 4.4). During all the experiments we had people tracking as application in mind which follows in the next Chapter. Thus, we expected the worst case reconstruction error of a person smaller than the average volume a person occupies in space which is in width approximately between half a meter and a meter.

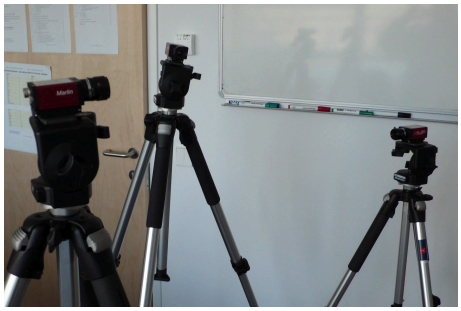
4.1 Image datasets

The image acquisition system consists of three Marlin MF-046C cameras from Allied Vision Technologies*, one PC and a pulse generator for camera synchronization. Details are in Fig. 4.1. To be able to operate the acquisition at a framerate of 10 Hz, the raw Bayer pattern image frames (300 kB each) with a resolution of 780×582 pixel are simultaneously stored on three separate hard disks - one for each camera. The cameras are attached to the computer via a single IEEE 1394a adapter card which delivers a maximal bandwidth of 400 Mbit/s. To acquire image sequences we use the software provided by the camera manufacturer.

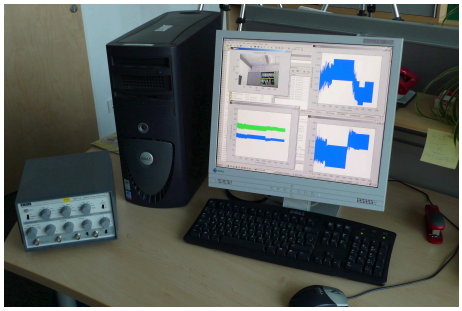
4.1.1 The Seminar room dataset

We acquired the first dataset in a Seminar room of our Institution. Only two of the four corners in the room are formed by orthogonal walls that makes neither method L-0 nor L-2

*http://www.alliedvisiontec.com; as at 30/11/2007.



(a) Camera 0 (left), 1 (middle), 2 (right)



(b) Pulse generator (left) and PC (right)

Figure 4.1: The image acquisition system. (a) shows the three Marlin MF-046C. Each camera is equipped with a low-cost Tamron 219-HB/8 8mm lens. The cameras are mounted on tripods and are connected to the acquisition PC via IEEE 1394a. (b) The PC is a DELL Precision 370 with an Adaptec FireConnect 4300 PCI card. The 10 MHz pulse generator is from Thurlby Thandar Instruments and delivers a constant 10 Hz TTL-signal for camera synchronization.

applicable. The day was cloudy, so the amount of light in the room changed rapidly from one time instant to the next, because two of the four walls are large glass facades. Sometimes, walking people mirrored themselves in the windows. All cameras were with almost zero pan in upright position which is critical for an accurate calibration. Fig. 4.2 shows sample images of two sequences. We see only a few edges on the rectangular windows and on the ceiling which are all distorted by the lens. This is a further challenge. The views of both cameras in Sequence 1 overlap in a small area (1 m across the views). We moved then Camera 2 along the windows until the camera was exactly imaged onto the image border of Camera 1. This non-overlapping situation was acquired as Sequence 2. Each image sequence contains 5,000 image frames.

Camera 1 and Camera 2 are used in the experiments that will follow in the next sections. Camera 0 served only for Thomas Svoboda's Multi-Camera Self-Calibration [Svoboda et al., 2005] which is an appropriate reference method. The method needs a point light source - for example a LED light - to generate corresponding points between at least three cameras. Fig. 4.3a depicts our battery operated LED light. The LED is from Signal-Construct*. This method estimates the interior orientation, the exterior orientation and the lens distortion all at once with sufficient accuracy and is applicable with lenses that have their focus at infinity which was the case in our camera setting. In contrast, calibration methods in the spirit of Tsai [1987] like the well known Matlab calibration toolbox† are useless with such lenses, because chessboard patterns appear blurry at a reasonable distance to the camera.

Two further sensors were used to estimate the exterior orientation in the non-overlapping case and to have the metric distance between the cameras. One was a laser sensor that

*http://www.signal-construct.de; as at 30/11/2007.
†http://www.vision.caltech.edu/bouguetj/calib_doc/, as at 30/11/2007.

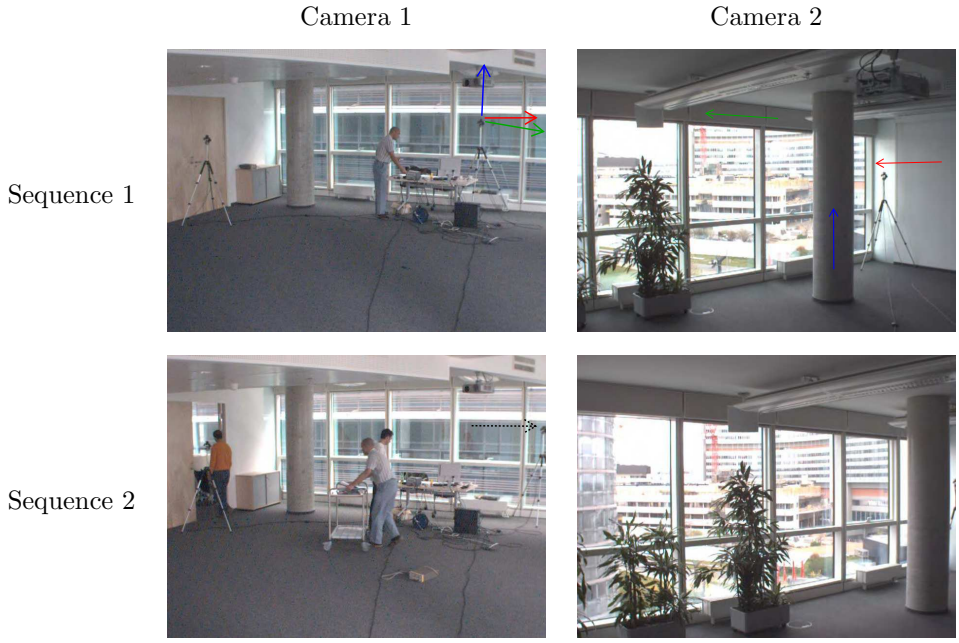


Figure 4.2: Synchronous sample image frames of the Seminar room dataset. Sequence 1 shows the cameras with overlapping views, Sequence 2 shows the cameras without overlapping views. Therefore, Camera 2 was moved 1 m along the window (black arrow). Camera 2 is visible in the image of Camera 1 and vice versa. The origin of the world coordinate system lies in Camera 2. The axes are shown by red (X), green (Y) and blue (Z) arrows. This definition is valid for both sequences.

(a) MWCE 55 (b) Distance sensor (c) Inertial sensor

Figure 4.3: Sensors that measure the ground truth. (a) LED light as point light source. The LED MWCE 5571 is extremely bright (460 mcd) and has an angle of reflected beam of 150 deg. (b) A commercial laser sensor from Bosch. The accuracy is ±0.5 mm for a distance of 30 m. (c) The inertial sensor. The accuracy in the Euler angles is ±0.5 deg.

measures the distance between two points in space (Fig. 4.3b) and an inertial sensor from Xens* that is capable of measuring the absolute orientation of itself to the Earth magnetic field (Fig. 4.3c). Tab. 4.1 summarizes the results of the calibration with the LED light and with the two sensors. The camera center of Camera 2 is in the origin of the world coordinate system. C_Z of Camera 1 is in all cases zero, because we used tripods with the same height for both cameras.

*http://www.xens.com; as at 30/11/2007.

Seq.	Cam.	f [pixel]	$\mathbf{p}$ [pixel]	$\mathbf{c}$ [pixel]	k [·]	ϕ [deg]	θ [deg]	ψ [deg]	C_X [m]	C_Y [m]
1	1	972.92	$\begin{pmatrix}376.58\\300.11\end{pmatrix}$	$\begin{pmatrix}379.62\\303.72\end{pmatrix}$	−.1606	0.25	8.53	−32.44	5.822	−7.739
	2	974.17	$\begin{pmatrix}393.96\\309.33\end{pmatrix}$	$\begin{pmatrix}398.98\\313.25\end{pmatrix}$	−.1625	0.19	2.12	−33.16	0.000	0.000
2	1	no difference to Sequence 1				0.25	8.53	−32.44	5.003	−8.372
	2					0.11	2.40	−11.13	0.000	0.000

Table 4.1: The ground truth of the Seminar room dataset. C_Z is zero in all cases.

4.1.2 The TechGate dataset

The distance between the people and the cameras is between half a meter and ten meters in the Seminar room which is a typical indoor scenario. In contrast to the Seminar room, we acquired a second dataset but at this time in the far-viewed foyer of our building; the distance between the people and the cameras are several meters. We captured two sequences with two cameras, Sequence 2 without overlapping views and Sequence 3 with substantial overlap. In Sequence 3, Camera 2, the two glass facades in the background are not orthogonal (62 deg), hence, only L-1 is applicable. The gap between the cameras in Sequence 2 is 2 m across their views which is a big challenge. Both sequences are 15 min long and contain 9,000 image frames. Both cameras were positioned on a gallery approximately ten meters above the foyer's floor. Only a few people walked through the scene during the acquisition, most of them between the elevators in the back and the exit of the building which is left of Camera 1. Fig. 4.4 shows sample images of the dataset. The illumination conditions varied rapidly during the acquisition, because the day had broken clouds. The effects were especially prominent on the floor where the changing shades let the tile's structure alternately appear and disappear.

The interior orientation of the cameras is the same as in the Seminar room. The inertial sensor delivered measurements of the orientation between the cameras. Unfortunately, we have not measured the exact coordinates of the translation between the two cameras in metric units. Nevertheless, the distance sensor gave us an absolute distance between the cameras. Tab. 4.2 shows a summary of the ground truth.

4.1.3 The PETS 2006 S3 dataset

The PETS 2006 dataset* contains multiple scenarios of left-luggage at Victoria train station, London. The image data was acquired as a benchmark for the PETS (Performance Evaluation of Tracking and Surveillance) workshop in 2006. Four hand-held cameras were placed at a

*http://www.pets2006.net, as at 30/11/2007.

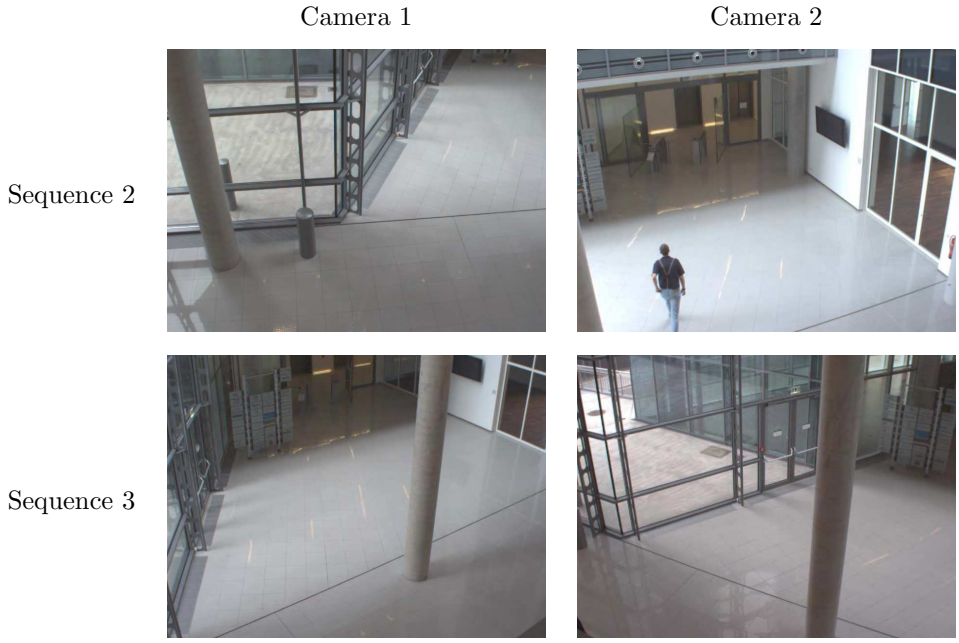
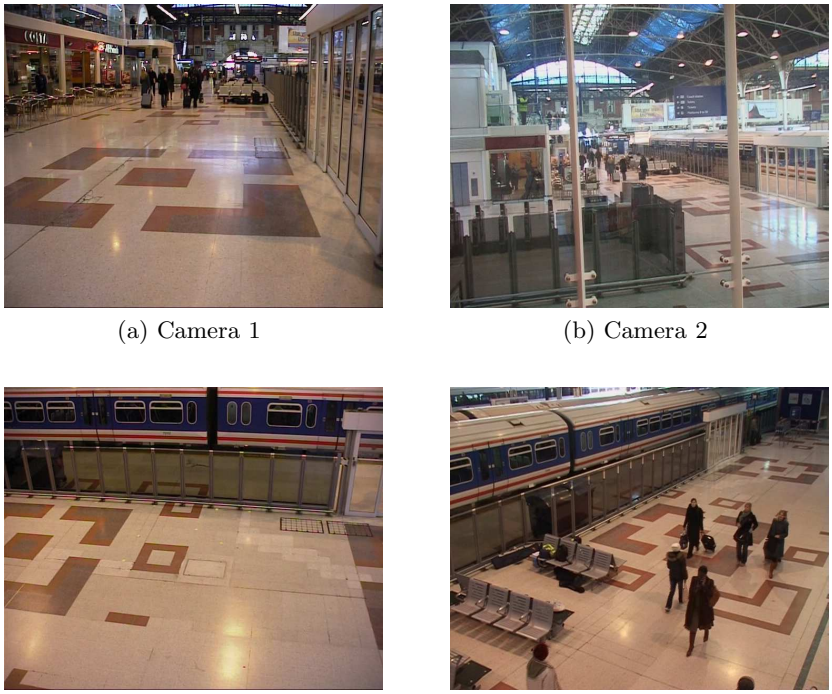


Figure 4.4: Synchronous sample image frames of the TechGate dataset. Sequence 2 is the non-overlapping case, Sequence 3 the overlapping one. Although small parts of the background overlap in Sequence 2, walking people are never visible in both cameras at the same time instant. The reader should also note that a non-overlapping situation can also arise with an overlapping background, because the detector will not fire as long as the people are not fully visible.

Seq.	Cam.	f [pixel]	$\mathbf{p}$ [pixel]	$\mathbf{c}$ [pixel]	k [·]	ϕ [deg]	θ [deg]	ψ [deg]	$\|C\|_2$ [m]
2	1	972.92	$\begin{pmatrix}376.58\\300.11\end{pmatrix}$	$\begin{pmatrix}379.62\\303.72\end{pmatrix}$	−.1606	0.91	42.03	−55.20	6.22
	2	974.17	$\begin{pmatrix}393.96\\309.33\end{pmatrix}$	$\begin{pmatrix}398.98\\313.25\end{pmatrix}$	−.1625	−0.28	25.41	−48.55	0.00
3	1	no difference to Sequence 2				1.01	30.44	−84.37	7.72
	2					1.69	31.43	9.23	0.00

Table 4.2: The ground truth of the TechGate dataset.

particular spot heading towards a nearby platform. Fig. 4.5 illustrates sample images of scenario S3. We choose this scenario, because of its low crowd density. It enfolds 2,370 image frames (720×576, JPEG, quality factor 90%) at a framerate of 25 Hz for each camera. Camera synchronization was achieved manually. The separate image sequences were subsequently aligned by dint of a flash light seen in all four cameras at the same time instant. Thus, the jitter between consecutive image frames has an upper bound of 40 ms. If we reasonably assume that a walking person moves a meter in a second the possible spatial error is bounded to 4 cm.



(a) Camera 1 (b) Camera 2

(c) Camera 3 (d) Camera 4

Figure 4.5: Synchronous sample image frames (S3-T7-A.00000.jpg) of the PETS dataset. Although the camera's views overlap in large parts of the world, the imaged areas that are visible are small and distant from the cameras, thus suffering under severe perspective distortion with low resolution and less texture due to the floor tiles. For example, note the small visible overlap of Camera 1 and Camera 3, although, both views overlap substantially in the world.

The authors of the dataset performed a manual calibration of all four cameras using known coplanar world points on the floor and their images [Thirde et al., 2006]. The first-order lens coefficient is not comparable to the k's in our datasets, because Tsai [1987] uses a slightly different lens distortion model. In the following sections we will transform the k's of our estimation, to make them comparable. We omit to show the camera centers, because due to a lack of space the dataset is not considered in the experiments of Sec. 4.3. Tab. 4.3 collects the results. We mistrust the accuracy of these parameters, because values of $(258\ 204)^\top$ for the principal point in uncropped images with image center $(360\ 288)^\top$ are unrealistic. Usually the difference is between 10 pixel and 20 pixel. Problems with the estimated lens parameters are also reported by Auvinet et al. [2006]. A possible reason for the substantial errors are the high perspective distortion which makes a manual selection of the imaged points with the necessary accuracy hard to achieve. More details are in Sec. 4.2.

Cam.	f	$\mathbf{p}$	$\mathbf{c}$	k	ϕ	θ	ψ
	[pixel]	[pixel]	[pixel]	[·]	[deg]	[deg]	[deg]
1	779.32	$(258.00\ 204.00)^\top$	$(258.00\ 204.00)^\top$	−0.5088	−0.98	14.33	3.98
2	899.66	$(258.00\ 204.00)^\top$	$(258.00\ 204.00)^\top$	0.3431	6.53	2.05	−6.66
3	832.16	$(258.00\ 204.00)^\top$	$(258.00\ 204.00)^\top$	0.3070	4.98	26.33	17.19
4	913.87	$(258.00\ 204.00)^\top$	$(258.00\ 204.00)^\top$	−0.2308	−0.67	17.01	29.52

Table 4.3: The ground truth of the PETS dataset.

4.2 Experiments with single cameras

The first experiment confirms our hypothesis that many images instead of a single image improve the accuracy of the calibration. C-0 in combination with C-3 or shortly C-0+3 runs over 3,000 image frames which are taken with a Logitech QuickCam Pro 4000. Sample images at the time instants 1, 1001 and 2001 are embedded into Fig. 4.6. They have a resolution of 320×240 pixel. The images show an office with the ceiling lightning switched on except between the frames 1001 and 2000. During this time some line segments appear in the upper right corner of the images, because more scene structure on the ceiling is now visible. Another change is the closed room door. Line segments appear in an area on the bookshelf which was previously covered by the door. Line segments on the closed door are also in alignment with the room which is not the case as it is open.

The grouping of line segments is correct except for some short line segments on the bookshelf, because the noise level is too large for a correct grouping. This is the classical Bias-Variance problem [Duda et al., 2001] and the only solution to reduce this problem is to provide more line segments.

The underlaid graph in Fig. 4.6 shows the adaptation behavior of C-3. What we see is a decrease in the uncertainty of the data after image frame 1000. During this time, C-3 improved the interior orientation by repeated adaptations. No adaptation happens after image frame 2000. Thus, C-3 is able to keep the good estimate when the uncertainty in the data arise.

To measure the accuracy in the estimation, we compared the results at different time instants to the result of Bouguet's reference calibration with a chessboard pattern*. This method achieved to be a standard in calibration with calibration patterns. Tab. 4.4 summarizes these results. C-3 was able to improve the focal length's relative error by a factor of 3.37. The location of the principal point was slightly improved. The two plots at the bottom of Fig. 4.6 give further insight at which time instant the improvement of the focal length happens. We

*http://www.vision.caltech.edu/bouguetj/; as at 30/11/2007.

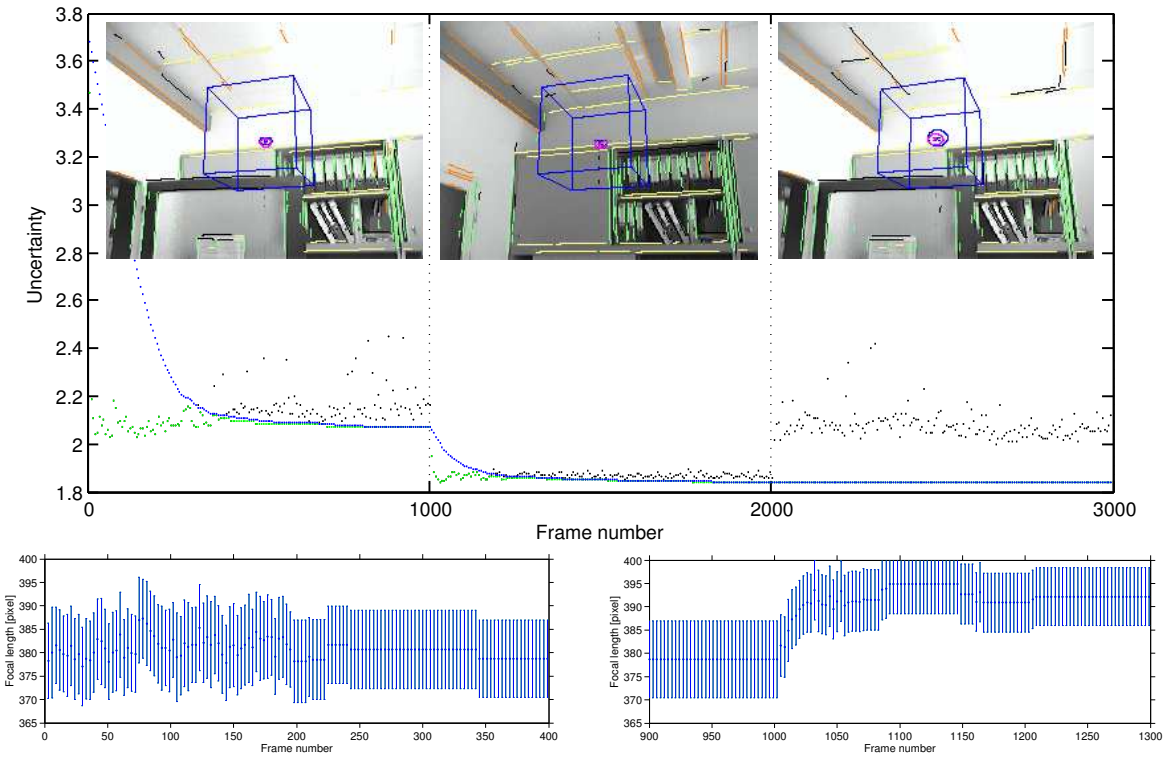


Figure 4.6: Underlaid graph: The behavior of the adaption in C-3. Each green point in the graph represents an image frame that gives an estimate of the interior orientation with lower uncertainty than the current estimate, because this image has more line segments and/or line segments in a better orientation. In this case, an adaptation step happens which is shown by the blue monotonically decreasing curve. No adaptation occurs in an image frame represented by a black point. The lightning change at image frame 1001 triggered abruptly the adaptation, while no adaptation occurs properly after image frame 2000. Sample images: The colors of the line segments encode the vanishing point except black that shows a noisy line segment. Note especially the line segments on the door that are associated with the right vanishing point. The proper estimate of the rotation is visualized by the blue cube. The principal point (blue) and the radial center (purple) are also drawn. The ellipses show their uncertainties. Bottom plots: The focal length's variation over time. No substantial improvement happens between image frame 1 and 400. Between image frame 900 and 1300 an abrupt improvement occurs. At the same time instant (1001) the ceiling lightning is switched off. Between image frame 2000 and 2400 C-3 was able to keep the good estimate.

identify image frame 1001 which is obviously the same time instant where the ceiling lightning was switched off.

Unfortunately, we have no ground truth of the rotation. So we cannot comment on that. With a k close to zero, C-0+3 fails to estimate the lens distortion. We believe the reason for that lies in the small image size and the low number of line segments.

The second experiment evaluates the accuracy of our approach in realistic surveillance environments. Thus, we let the calibration C-0 and combinations with C-2 (C-0+2) and C-3 (C-0+3) run over the PETS dataset and we let the calibration C-1 and combinations with

Frame	f	δf	$\mathbf{p}$	$\delta\mathbf{p}$
	[pixel]	[%]	[pixel]	[%]
1	378.25 ± 3.20	5.62	$(159.00 \pm 3.27\ 119.00 \pm 2.56)^\top$	$(2.09\ 1.47)^\top$
3000	392.94 ± 3.17	1.67	$(160.17 \pm 3.32\ 122.02 \pm 2.38)^\top$	$(1.35\ 1.04)^\top$

Table 4.4: The results of C-0+3 before (first image frame) and after adaptation (image frame 3000). Absolute values, their uncertainty and their relative error to the reference calibration are given. C-3 improved significantly the interior orientation by using many images.

Seq.	Cam.	Met.	δf	$\delta\mathbf{p}$	$\delta\mathbf{c}$	δk	k
			[%]	[%]	[%]	[%]	[·]
2	1	C-1	0.94	$(3.19\ 3.49)^\top$	$(4.21\ 1.35)^\top$	9.93	$-.1783 \pm .0068$
		C-1+2	0.94	$(3.19\ 3.49)^\top$	$(0.68\ 1.86)^\top$	4.12	$-.1675 \pm .0056$
		C-1+3	0.11	$(0.10\ 3.34)^\top$	$(0.94\ 1.04)^\top$	15.65	$-.1904 \pm .0074$
	2	C-1	1.06	$(1.28\ 6.67)^\top$	$(6.17\ 8.61)^\top$	19.22	$-.1363 \pm .0041$
		C-1+2	1.06	$(1.28\ 6.67)^\top$	$(8.85\ 7.67)^\top$	3.96	$-.1692 \pm .0032$
		C-1+3	0.08	$(3.43\ 12.88)^\top$	$(9.31\ 2.63)^\top$	14.36	$-.1421 \pm .0042$
3	1	C-1	0.94	$(3.19\ 3.49)^\top$	$(2.50\ 7.42)^\top$	66.25	$-.0966 \pm .0024$
		C-1+2	no difference to C-1				
		C-1+3	2.37	$(10.18\ 1.30)^\top$	$(2.62\ 5.48)^\top$	386.67	$-.0330 \pm .0008$
	2	C-1	1.06	$(1.28\ 6.67)^\top$	$(5.76\ 8.96)^\top$	11.23	$-.1461 \pm .0056$
		C-1+2	1.06	$(1.28\ 6.67)^\top$	$(3.28\ 8.58)^\top$	76.06	$-.0923 \pm .0028$
		C-1+3	0.61	$(1.14\ 7.63)^\top$	$(1.53\ 8.88)^\top$	84.03	$-.0883 \pm .0036$

Table 4.5: The errors of the proposed calibration with the TechGate dataset.

C-2 (C-1+2) and C-3 (C-1+3) run over the self-acquired datasets. In the latter case we knew the lens characteristics. Appendix A contains Tab. A.1 (PETS), Tab. A.2 (Seminar room) and Tab. A.3 (TechGate) that summarize all estimated parameter values. These results are compared with the ground truth; see the previous section. The relative errors except for the PETS dataset are given by Tab. 4.5 and Tab. 4.6. The last column in these Tables shows the converted values of k which are compared to the ground truth values.

We observe a large deviation to the ground truth for the parameters in the PETS dataset. Auvinet et al. [2006] also report problems with the ground truth. We confirm their observation with a qualitative evaluation in Fig. 4.7. We conclude that the Tsai calibration by Reg Wilson* [Tsai, 1987] fails with the manually defined points, because some of the points lie close together.

A first surprise was the with 1 % small focal length error of the lens. In the best case (TechGate, Sequence 2, Camera 2), C-1+3 was able to reduce this error by a factor of ten,

*http://www.comp.leeds.ac.uk/chrisn/Tsai/index.html; as at 30/11/2007.

Seq.	Cam.	Met.	δf [%]	$\delta\mathbf{p}$ [%]	$\delta\mathbf{c}$ [%]	δk [%]	k [·]
1	1	C-1	0.94	$(3.19\ 3.49)^\top$	$(5.11\ 6.07)^\top$	118.80	$-.0734 \pm .0059$
		C-1+2	0.94	$(3.19\ 3.49)^\top$	$(3.24\ 3.16)^\top$	8.44	$-.1481 \pm .0083$
		C-1+3	3.38	$(7.67\ 8.96)^\top$	$(5.68\ 5.39)^\top$	111.59	$-.0759 \pm .0048$
	2	C-1	1.06	$(1.28\ 6.67)^\top$	$(12.00\ 1.80)^\top$	34.32	$-.2474 \pm .0130$
		C-1+2	1.06	$(1.28\ 6.67)^\top$	$(4.24\ 6.05)^\top$	19.63	$-.2022 \pm .0075$
		C-1+3	3.16	$(12.40\ 4.18)^\top$	$(5.68\ 4.78)^\top$	0.99	$-.1609 \pm .0070$
2	1	C-1	0.94	$(3.19\ 3.49)^\top$	$(10.58\ 18.81)^\top$	77.68	$-.7192 \pm .0542$
		C-1+2	0.94	$(3.19\ 3.49)^\top$	$(2.73\ 10.24)^\top$	9.10	$-.1472 \pm .0083$
		C-1+3	2.51	$(4.67\ 4.25)^\top$	$(0.60\ 13.50)^\top$	38.40	$-.2607 \pm .0156$
	2	C-1	1.06	$(1.28\ 6.67)^\top$	$(0.23\ 3.04)^\top$	19.55	$-.2020 \pm .0074$
		C-1+2	1.06	$(1.28\ 6.67)^\top$	$(1.45\ 5.22)^\top$	24.28	$-.2146 \pm .0073$
		C-1+3	2.63	$(26.20\ 4.10)^\top$	$(10.74\ 4.01)^\top$	13.75	$-.1884 \pm .0053$

Table 4.6: The errors of the proposed calibration with the Seminar room dataset.

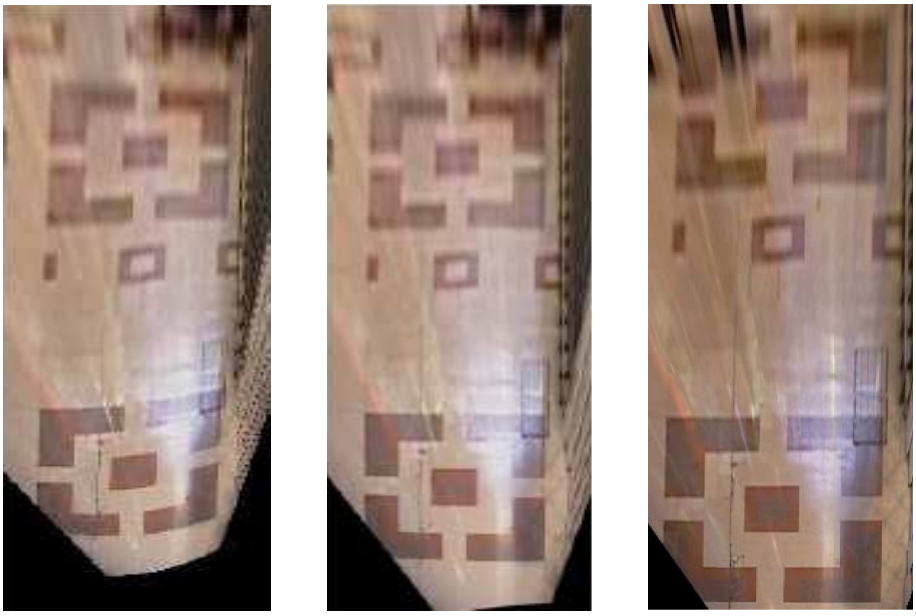
however, in the worst case (Seminar room, Sequence 1, Camera 1), the opposite - a decline by a factor of three - happens. Nevertheless, the worst case absolute difference between the estimate and the ground truth was with 31.84 pixel moderate under the assumed noise level [Liebowitz, 2001] and with this challenging data.

Most images contain 30 % - 40 % noisy line segments and line segments in two orthogonal directions. The third direction is underrepresented by less that 10 % of all line segments in the image (see $\mathbf{v}_X$, $\mathbf{v}_Y$ and $\mathbf{v}_Z$ in Tab. A.2 and Tab. A.3). Kosecka and Zhang also mention this worst 5 % error as upper bound for their approach [Kosecka and Zhang, 2002]. Grammatikopoulos et al. [2007] reported 0.7 % as smallest error for their approach which is comparable to 0.8 % for the case (TechGate, Sequence 2, Camera 2) which we believe is much more difficult than their outdoor house views.

Many authors mention a large variation in the calibration of the principal point [Grammatikopoulos et al., 2007; Zhang et al., 1996]. Our experiments confirm this observation. The largest error for C-1+3 is 26.20 % in the u-axis of Camera 2 in Sequence 2 of the Seminar room dataset, the smallest 0.1 % error occurs in Camera 2, Sequence 2 of the TechGate dataset. Again the challenging data makes a localization with high precision difficult. Nevertheless, the results are still usable in surveillance applications [Zhang et al., 1996].

The radial distortion is in some images underestimated and in some images overestimated but only in a few images close to the ground truth, for example, in Camera 2, Sequence 1 of the Seminar room dataset. The reason for this large variation is the uneven distribution of the line segments in the image. Fig. 4.9 shows the results.

The experiment also shows a strong relationship between the estimate of the radial distortion



(a) Ground truth (b) Auvinet et al. (c) Proposed approach

Figure 4.7: The images show the bird's eye view from Camera 1 of the PETS dataset. A homography obtains the image in (a) which is computed with the ground truth. (b) This homography was recomputed by DLT [Hartley and Zisserman, 2004] with the original point correspondences. (c) Our calibration. The lens undistortion is included. Orthogonal structures should appear orthogonal after rectification. The qualitative errors are obvious in the bottom area of the images. Our calibration is satisfying from the point of view that it is obtained automatically.

and the rotation. This insight is that varying the parameters of the distortion has a direct influence on the vanishing points that encode the rotation. This is an important hint that accurate rotation estimates are impossible without a consideration of the lens distortion. Important papers neglect lens distortion [Kosecka and Zhang, 2002; Rother, 2002a; Schindler and Dellaert, 2004], because they only concentrated on vanishing point estimation. Our results show that an accurate estimate of the vanishing points is hard without considering the lens distortion. Both methods C-2 and C-3 improved the estimates of the Euler angles on average by half a degree, in one case C-0+3 gave an improvement for the pitch larger than 3 deg (PETS dataset, Camera 4). In contrast, C-0+2 gave the best estimates of the distortion parameters which are still moderate to poor. We observed that in scenes with a few line segments and severe lens distortion the estimation then stuck easily in local minima of the cost function. The next two sections contain more evaluation about the rotation.

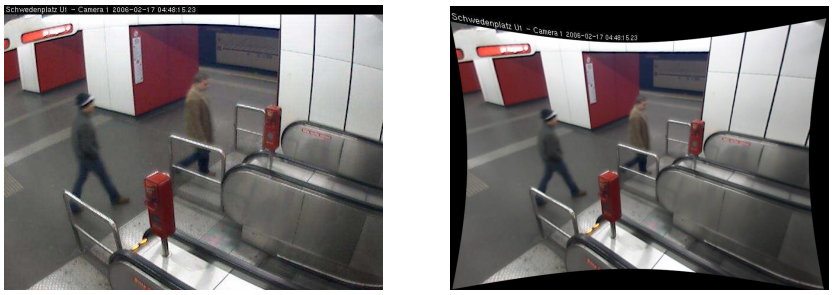
After this careful evaluation of the second experiment we can summarize that C-0 and C-1 gives reasonable estimates in challenging scenes with a few line segments and without much illumination change, although, improvements by further using C-2 and C-3 can be reported.

Figure 4.8: The images show in the same order from top left to bottom right the results of C-0+3 with the PETS dataset (Cameras 1-4), the Seminar room 1, 2 dataset (Camera 1,2) and the TechGate 2, 3 dataset (Camera 1,2). The color coding is explained in Fig. 4.6.

4.3 Experiments with overlapping views

This section demonstrates the success of our approach with distant cameras but overlapping views. Sec. 4.3.1 verifies with synthetic data that the 8-point algorithm [Hartley and Zisserman, 2004] fails in this situation. Sec. 4.3.2 provides evidence on the Seminar room 1 dataset (Sec. 4.1.1) that Thomas Svoboda's Multi-Camera Self-Calibration [Svoboda et al., 2005] is also inapplicable with points that are generated by automatic detectors. Sec. 4.5 exemplifies on the PETS 2006 dataset (Sec. 4.1.3) and the Tech Gate 3 dataset (Sec. 4.1.2) the remarkable robustness of people matching using our calibration.

(a) Seminar room



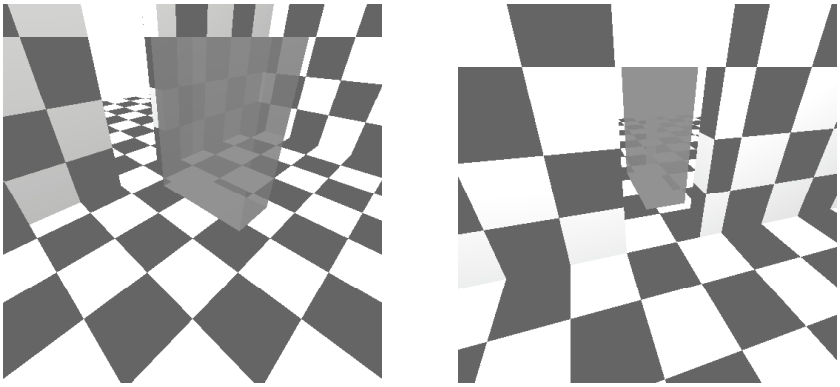
(b) Underground station

Figure 4.9: Qualitative evaluation of the lens estimation. The lens distortion is correctly removed in (a). The blue line coincides well with the edge along the ceiling in the rectified image (right). This is not true in the original image (left). The severe distortion in (b) is significantly reduced, however, not completely removed. This image shows an underground station in Vienna; courtesy of Norbert Brändle, Arsenal research. The upper left background still shows a distortion. The reason are missing and too small line segments in this image area.

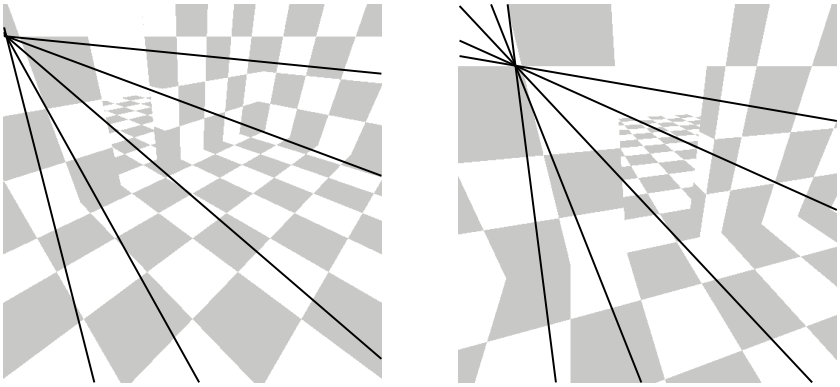
4.3.1 In comparison with the 8-point algorithm

Figures 4.10a and 4.10b show synthetic images of two virtual cameras. All point correspondences were generated in a volume of space which projects for both cameras to approximately a fourth of the whole image area. The point correspondences were moderately distorted using a radial, first-order lens distortion model [Faugeras and Luong, 2001] with the radial center set to the image center and a first-order coefficient set to 0.1. Then, Gaussian noise was added to the points with zero mean and a variance σ^2_{noise} of 0.01 pixel. The ground truth of the epipolar geometry between the two cameras derives directly from PovRay's design parameters. Fig. 4.10c illustrates the ground truth in Camera 1 by five epipolar lines of five points that were manually defined in the image of Camera 2. Fig. 4.10d shows the same vice versa for Camera 2.

Compare now the ground truth with the result of our calibration (Figures 4.11a and 4.11b). We used C-0+2 for each camera separately and then localized both cameras by L-1 with eight particular point correspondences. No large difference is seen. In contrast, Figures 4.11c and 4.11d show the 8-point algorithm's wrong result. To put the error into numbers, we used the



(a) Camera 1 (b) Camera 2

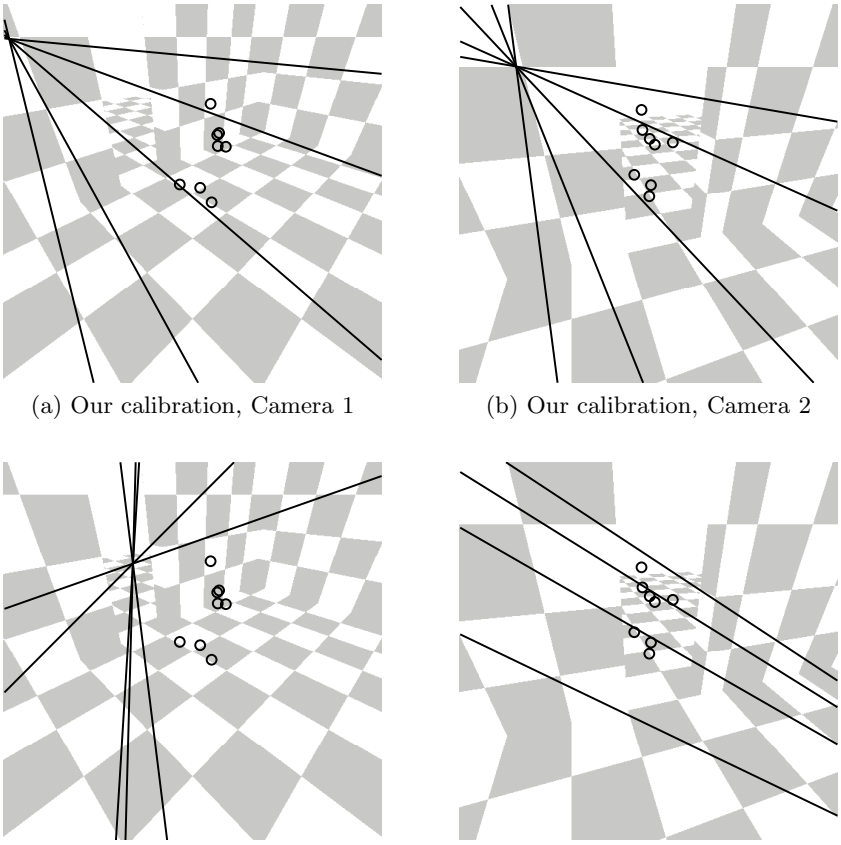


(c) Ground truth, Camera 1 (d) Ground truth, Camera 2

Figure 4.10: Synthetic images of two virtual cameras with small visible overlapping views designed with PovRay which is a raytracing software and is freely available under http://www.povray.org. The highlighted volumes in (a) and (b) are the space where point correspondences are generated. (c) and (d) show the black ground truth's epipolar lines.

error measure of Zhang [Faugeras and Luong, 2001]. It gives us a reasonable error of 6.88 pixel between the ground truth's fundamental matrix and the one estimated by our calibration. The error of the 8-point algorithm's fundamental matrix is with 128.25 pixel unacceptably large. The wrong result of the 8-point algorithm is obvious, although no coplanarity in the configuration of the point correspondences exists. The result suggests that small distances between the points are critical for the 8-point algorithm. In contrast, our two-step calibration uses line segments as additional information. Hence, the result corroborate our argument that the rotation's ambiguity and the camera centers are computable with clustered points.

One might legitimately argument that a configuration of a particular number of point correspondences has no significance. Therefore, we drew randomly 12.070 different configurations of 8, 10, 15, 20, 25 and 30 point correspondences. We were also interested how different noise levels influence the calibration. For a σ^2_{noise} of 0.01 pixel the errors are shown in Fig. 4.12a. Our approach outperforms the 8-point algorithm; considerably when less than 15 point correspondences are available. This number increases with higher noise levels, for example, it



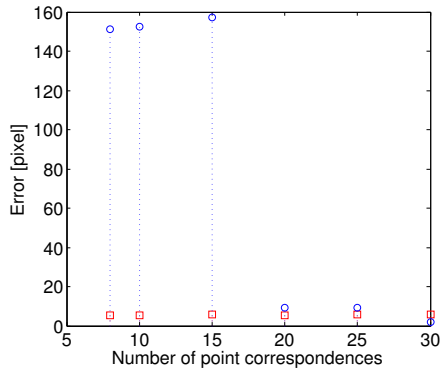
(a) Our calibration, Camera 1 (b) Our calibration, Camera 2

(c) 8-point algorithm, Camera 1 (d) 8-point algorithm, Camera 2

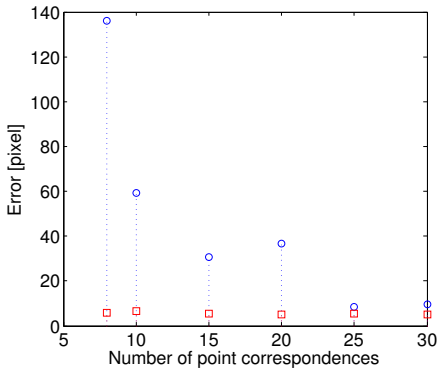
Figure 4.11: An experiment with eight particular point correspondences shown as black circles. (a) and (b) show the black epipolar lines of our calibration. (c) and (d) show the wrong result of the 8-point algorithm. The eight points in each image are neither collinear nor coplanar.

is 20 with $\sigma^2_{noise} = 0.25$ pixel (Fig. 4.12b). These statistical results confirm that the 8-point algorithm is improper to estimate an accurate fundamental matrix with a few clustered point correspondences. In contrast, the errors of our approach suggest independence in the number of point correspondences. Even under different noise levels, the order of magnitude remains approximately 6 pixel.

For a σ^2_{noise} of 1 pixel, this error increased slightly to 10 pixel, however, the 8-point algorithm's mean error is with 104 pixel much larger (Fig. 4.13a)! Although 256 point correspondences reduce drastically this number (Fig. 4.13b), our calibration's lower error bound is still smaller than the 8-point algorithm's one; 2.5 pixel instead of 5.3 pixel. This result suggests that our calibration can be used with a few point correspondences from noisy automatic detectors whereas the 8-point algorithm will fail.

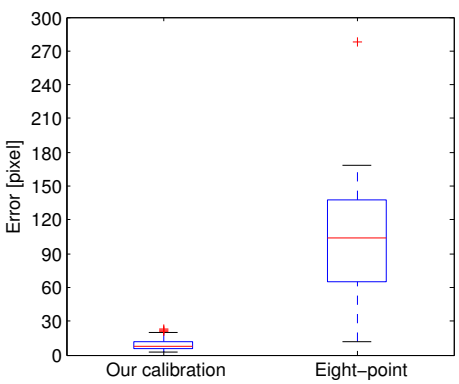


(a) $\sigma^2_{noise} = 0.01$ pixel

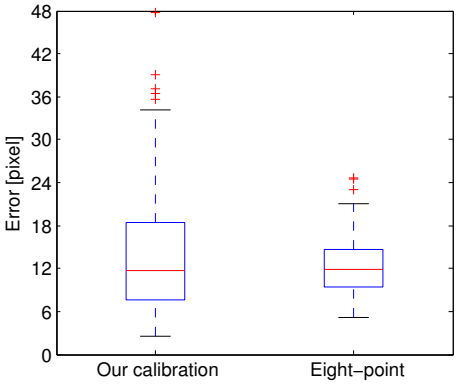


(b) $\sigma^2_{noise} = 0.25$ pixel

Figure 4.12: Error evaluation with 8, 10, 15, 20, 25 and 30 point correspondences. (a) Zhang's error of our calibration vary around 5.6 pixel (red). The largest error of the 8-point algorithm is 157.04 pixel (blue). (b) The errors of our calibration remain in the same order.



(a) 8 point correspondences



(b) 256 point correspondences

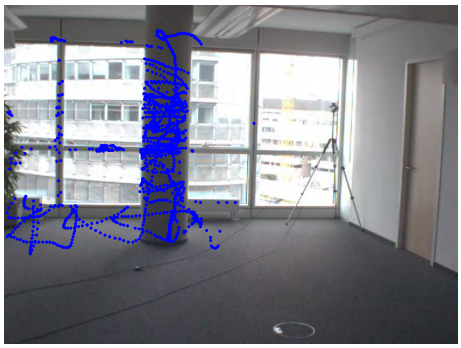
Figure 4.13: Error evaluation with a $\sigma^2_{noise} = 1$ pixel. Notched box plots are shown. (a) shows a large error for the 8-point algorithm while in (b) this error reduces drastically but despite is larger than the error of our calibration which was not much affected by the number of point correspondences.

4.3.2 In comparison with Svoboda's Multi-camera Self-calibration

After a verification with synthetic data it is now time to experiment with the Seminar room 1 dataset. During the acquisition of this dataset, we shaded the room and used a bright green LED light to manually generate subpixel accurate point correspondences in the same way as we did for the ground truth described in Sec. 4.1. Fig. 4.14 shows the LED light imaged in all three cameras at the same time instant and for the whole image sequence.

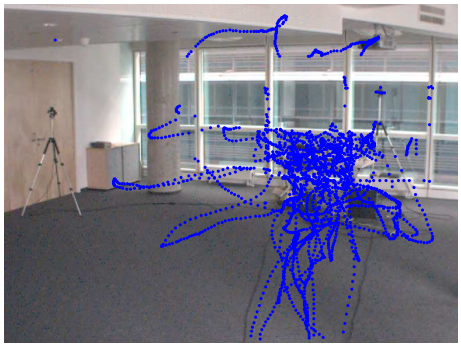
We also let an automatic top of the head detection [Zhao and Nevatia, 2004] run over the whole dataset which generates points far more imprecise. Nevertheless, the top of the head has two advantages among other possible points, for example, the center of gravity of the blob. On the one hand, the image area around this point is less vulnerable to shading effects, because we usually assume the location of a light source above our heads. On the other hand,

(a) Marlin 0, Frame 732



(b) Marlin 0, all frames

(c) Marlin 1, Frame 732



(d) Marlin 1, all frames

(e) Marlin 2, Frame 732

(f) Marlin 2, all frames

Figure 4.14: The generation of point correspondences with a LED light (red circles). (a), (c) and (e) show three images of the three cameras. Blinds shade the room to enable a localization. The images are histogram-equalized to improve their visibility. (b), (d) and (f) show the position of the LED light as blue points within 5,400 image frames. Without surprise, the detection works best in dark areas like in front of the column, because it was impossible at daytime to completely shade the room. This situation will often occur in practice.

the top of the head is a frontier point [Sinha et al., 2004].

The detection performs frame differencing to identify blobs. As by Toyama et al. [1999], we set the minimal pixel difference to 16 pixel. We also tried adaptive background subtraction with an approximated median of the background [McFarlane and Schofield, 1995], though, we observed more troubles with illumination change than with simple frame differencing. To eliminate some blobs caused by image noise, the area of the blob is assessed. We observed that blobs larger than 1000 pixel are predominately caused by walking people. Of course, this threshold depends on the scene depth. Finally, the detection founds the boundary of a blob, smoothes it by averaging within a five point wide window along the boundary and declares the point on the boundary with the smallest v-coordinate to the top of the head. Sometimes more than one point has the smallest v-coordinate. Then, the median of these points is chosen.

Fig. 4.15 shows examples of detections at the same time instant and for the whole image sequence. The many outliers attract attention at a first glance. Fortunately, they are correctly eliminated by RANSAC in the successive point matching step in Svoboda's method as in L-2. One might argue to use more sophisticated classifiers for faces [Viola and Jones, 2004] and for people [Dalal et al., 2006] to reduce substantially this problem. Although these detectors are preferable they frustrate in their poor repeatability of localizing the same point in successive image frames. For example, we observed errors between the center of gravity and the correct point greater than 10 pixel; such errors make these detectors useless.

We run Svoboda's method with the default parameter values in over the LED light and top of the head correspondences. A detailed description of the parameters is on the website (Sec. 4.1). We activated the Bundle adjustment before each outlier removal, we changed the optimization of the nonlinear parameters, that is, only the first coefficient of the radial lens distortion is optimized, and, finally, we enabled square pixels for the image sensor. Tab. 4.7 summarizes the results of the cameras's interior orientation.

Compared to the reference calibration and to the results of C-1, C-1+2 and C-1+3 (Sec. 4.2), the numbers show the poor performance on the Seminar room 1 dataset. The lens distortion is significantly overestimated. Although, in the LED case, the relative error in the focal length (7.36 %) and in the principal point (4.28 % in the u-coordinate and 6.51 % in the v-coordinate) is reasonable for Camera 1, it is severe for Camera 2. As with the 8-point algorithm in Sec. 4.3.1 the method is unable to handle points that are clustered in small image areas. These image areas are even smaller in Camera 2 which is the reason for the larger inaccuracy. Svoboda's method fails completely with the points on top of the heads.

L-2 also run over the same data and on top of the results of both C-1 and C-1+3. The intention was to test the sensitivity of L-2 with respect to the accuracy of the calibration. Marlin 0 was not used during the experiments with L-2, because L-2 works with at least two

Cam	Source	Method	f [pixel]	δf [%]	$\mathbf{p}$ [pixel]	$\mathbf{c}$ [pixel]	k [·]	δk [%]
1		Reference	972.92		$\begin{pmatrix}376.58\\300.11\end{pmatrix}$	$\begin{pmatrix}379.62\\303.72\end{pmatrix}$	−0.1606	
	LED	Svoboda	1044.57	7.36	$\begin{pmatrix}392.69\\280.56\end{pmatrix}$	$\begin{pmatrix}389.00\\290.00\end{pmatrix}$	−0.4907	205.54
	Head	Svoboda	0.02	100.00	$\begin{pmatrix}608.29\\161.45\end{pmatrix}$	$\begin{pmatrix}320.00\\240.00\end{pmatrix}$	0.0000	100.00
2		Reference	974.17		$\begin{pmatrix}393.96\\309.33\end{pmatrix}$	$\begin{pmatrix}398.98\\313.25\end{pmatrix}$	−0.1625	
	LED	Svoboda	1231.63	26.43	$\begin{pmatrix}489.13\\428.81\end{pmatrix}$	$\begin{pmatrix}389.00\\290.00\end{pmatrix}$	−0.5389	231.63
	Head	Svoboda	35.90	96.31	$\begin{pmatrix}339.62\\190.29\end{pmatrix}$	$\begin{pmatrix}320.00\\240.00\end{pmatrix}$	0.0000	100.00

Table 4.7: The interior orientation with Svoboda's Multi-Camera Calibration.

Source	Method	ϕ [deg]	θ [deg]	ψ [deg]	C_X [cm]	δC_X [%]	C_Y [cm]	δC_Y [%]	C_Z [cm]
	Reference	3.29	5.37	16.30	582.20		−773.90		0.00
LED	C-1/L-2	0.90	6.33	16.47	599.74	3.01	−757.91	2.07	1.10
	C-1+3/L-2	4.09	6.67	15.87	572.17	1.72	−778.93	0.65	1.96
Head	C-1/L-2	3.85	4.58	14.24	884.80	51.98	−388.08	49.85	25.57
	C-1+3/L-2	3.92	6.59	16.82	569.63	2.16	−780.80	0.89	0.86

Table 4.8: The exterior orientation with our calibration.

cameras. The results are summarized in Tab. 4.8.

At a first glance, the localization results of L-2 with the LED light are satisfying. Fig. 4.16 shows the same results qualitatively by the epipoles. The accuracy of the orientation was in all three Euler angles below 1 deg. Note, that the inertial sensor has itself an uncertainty of 0.5 deg. The relative error in the camera position was below 3 %. L-2 estimated the two camera positions with approximately 10 cm error which is good compared to the distance of 9.66 m between the two cameras. Remarkable are the results with the points on top of the heads. The estimation of the Euler angles is similar to the result with the LED light, and the accuracy in the camera positions is only slightly worse (approximately 2 cm in all directions).

However, a severe problem is evident. L-2 is sensitive to the calibration. While L-2 on basis of C-1 produces with the LED a large relative error only for the pitch (72.64 %), the same combination fails to localize the cameras with the head points. In contrast, the combination of C-1+3 and L-2 works satisfying. A few degrees deviation in the initial estimation of the

orientation can cause a breakdown of L-2. This deviation depends on the distance between the cameras. The greater this distance the more accurate the initialization must be.

We observed in many experiments that the main reason for a large error in the orientation was caused by a poor estimation of the lens distortion. The estimation of the orientation and the estimation of the lens distortion are interrelated. A substantial improvement in the lens distortion will improve the orientation. Sec. 4.2 showed that C-3 brought such an improvement by simply processing many images instead of only one. Here, we see that this approach was necessary to get a correct localization.

In our next experiment we used the PETS 2006 dataset for evaluation. We calibrated each camera's intrinsic and extrinsic parameters by C-0+3. Known point sets that are provided by the ground truth of the benchmark data were used for testing the accuracy. We triangulated these points and projected them back into the image. The worst re-projection error in Camera 1 was 7.3 pixel, in Camera 3 10.2 pixel and in Camera 4 2.8 pixel. Successively, L-1 localized the three cameras with a short image sequence showing a person walking through the whole scene. Fig. 4.17 presents these results.

Finally, the combination C-1+3/L-2 also succeeds with the Tech Gate 3 dataset. The epipolar geometry and the collected head points are shown in Fig. 4.18.

4.4 Experiments without overlapping views

This section presents experiments that confirms L-3's power in cameras without overlapping views. The parameter values of L-3 are the same in all experiments.

Our first experiment shows L-3's generality to handle two cameras with overlapping and non-overlapping views. We generated 50 synthetic trajectories of a moving object in a 10 m × 10 m × 10 m virtual 3D-space. Each trajectory is at least 5 m in length with a step size of 50 cm between consecutive positions. Q's diagonal is set to (.0001 .0001 .0001 0 .001 .001 .001) and Q's off-diagonal elements are all zero. The worst error that accumulates from one position to the next is approximately three times $\sqrt{.0001}$, that is 3 mm, and the largest change of the velocity is approximately below three times $\sqrt{.001}$, which is 10 cm - a fifth of the step size. This model is similar to the motion of people, except for the motion in the Z-axis which is stable in reality, because people move usually on earth.

The two cameras are placed opposite to each other at the same height, that is, C_Z is zero for both cameras. Camera 2 is placed in a fixed position at $(0\,\mathrm{m}\ 10\,\mathrm{m}\ 0\,\mathrm{m})^\top$. Camera 1 is virtually moved along the X-axis to $(\mathrm{x\,m}\ -10\,\mathrm{m}\ 0\,\mathrm{m})^\top$ with $x \in \{.5, 0, -.5, -1, -1.5, -2\}$. These different positions yield from a small overlap of the views (1 m between the borders in direction of the X-axis), over exactly no overlap to a non-overlapping gap (1 m, 2 m, 3 m

Gap [m]	C_X^μ [cm]	C_X^σ [cm]	C_X^Δ [cm]	C_Z^μ [cm]	C_Z^σ [cm]
4	-154.31	7.94	45.69	8.13	14.58
3	-127.46	6.71	22.54	-0.61	12.62
2	-88.66	1.76	11.34	-1.16	3.01
1	-44.95	2.41	5.05	1.08	5.01
0	1.56	0.52	1.56	2.98	3.04
-1	50.03	0.05	0.03	0.20	0.51

(a) 50 trajectories

Gap [m]	C_X^μ [cm]	C_X^σ [cm]	C_X^Δ [cm]	C_Z^μ [cm]	C_Z^σ [cm]
4	-168.38	7.15	31.62	8.51	7.71
3	-131.74	2.96	18.26	-3.04	4.75
2	-87.94	1.73	12.06	-1.46	4.99
1	-44.60	0.92	5.40	2.67	4.10
0	1.23	0.18	1.23	0.23	1.54
-1	50.02	0.04	0.02	0.10	0.47

(b) 200 trajectories

Gap [m]	C_X^μ [cm]	C_X^σ [cm]	C_X^Δ [cm]	C_Z^μ [cm]	C_Z^σ [cm]
4	-83.55	26.88	116.45	9.73	32.12
3	-103.92	14.06	46.08	-0.09	16.65
2	-59.21	15.10	40.79	26.23	33.45
1	-30.05	6.41	19.95	-5.62	18.89
0	5.97	0.90	5.97	4.11	9.52
-1	50.50	0.24	0.50	-0.45	2.11

(c) 30 cm worst case velocity variation

Gap [m]	C_X^μ [cm]	C_X^σ [cm]	C_X^Δ [cm]	C_Z^μ [cm]	C_Z^σ [cm]
4	-177.06	20.03	22.94	-3.62	23.87
3	-142.20	11.54	7.80	-11.60	29.54
2	-100.62	12.57	0.62	9.86	29.32
1	-44.71	3.79	5.29	-0.55	17.01
0	0.51	1.68	0.51	0.63	5.82
-1	50.21	0.59	0.21	1.17	4.40

(d) 15 % outlier

Table 4.9: Results of L-3 for several positions of Camera 1. C_Y is in all cases -10 m. A gap of -1 m means that the views overlap 1 m along the X-axis.

and 4 m). To produce these results, the pan and tilt angles of both cameras are zero and the yaw angle of both cameras is chosen in that way that the plane spanned by Camera 2's right image edge is parallel to the plane spanned by Camera 1's left image edge, that is, 58 deg for Camera 2 and -58 deg for Camera 1. These angles depend also on the camera's interior orientation. We set the focal length of both cameras to 200 pixel which allows for the given camera positions a capturing of the whole space. The principal points are set to $(319\ 239)^\top$ which is with a resolution of 640 pixel $\times$ 480 pixel for both cameras the center of the image.

These interior and exterior orientation project the trajectories into the cameras. For the sake of quantization, each imaged position is disturbed by Gaussian noise with a variance that the worst case error is half a pixel. The reader will imagine that not all trajectories are visible in both cameras, because some trajectories are generated in small parts of the space. However, M-0 is only able to generate trivial matches when a trajectory is visible in both cameras. Between half to three forth of all trajectories are lost with 1 m overlap, and between three fourth to sometimes all trajectories are lost with the 4 m gap. To have a robust evaluation of L-3, we generate not only one but 15 sets of 50 trajectories and compute the median C^μ of the estimated camera positions. The variance C^σ is also computed only over the inlier positions. We know that 15 sets are not enough in a statistical sense but L-3 is currently too slow to evaluate it over a larger number of sets.

Tab. 4.9a summarizes all these numbers for Camera 1. Fig. 4.19 shows them graphically in

the first row. The position of Camera 1 is always multiplied by a scalar that C_Y amounts to -10, because we want to compare the estimated positions to the ground truth in metric units. The positions are estimated accurately by L-3 in the overlapping but also in the non-overlapping case. The maximal absolute error with a 4 m gap is with 45.69 cm less than half a meter with a reasonable variance in the different samples which allows a save use of L-3. The error for a gap of 2 m is 11.34 cm or 1.1 % of the environment's size. Rahimi et al. Rahimi et al. [2004] reported in the same case 1.4 % which is a bit more, however, their method computes also rotation estimates. Fig. 4.20 shows L-3's remarkable power to reconstruct the trajectories even without any overlap.

We repeat the same experiment with 200 trajectories and report an improvement in the accuracy of the positions in the X-axis (Tab. 4.9b). Instead of 45.69 cm the worst error is now 31.62 cm - an improvement of 14.07 cm. An improvement in the Z-axis happens only for the overlapping and the exactly non-overlapping case. L-3 is simply more reliable when using 200 trajectories. Especially in the 4 m gap case, the half a meter upper boundary is now guaranteed (see the second row in Fig. 4.19).

The next experiment shows L-3's breakdown when the maximal possible change of the velocity is increased to 30 cm, that is three times more. L-3 fails with the 2 m gap, because the variance is too large to guarantee an error within 1 m (Tab. 4.9c, second row of Fig. 4.19). More than 200 trajectories would definitely help to reduce the variance and to make L-3 applicable for larger gaps - at least for three meters. For example, only 30 until 48 trajectories out of the 200 are usable in the 2 m gap case.

The last experiment with our synthetic data shows the robustness of L-3 when adding gross outliers. We added 10 pixel to the imaged positions with a probability of 15 %. Tab. 4.9d and the last row of Fig. 4.19 confirm our expectation that the RANSAC framework makes L-3 robust against gross outliers. The variance is large, because many trajectories are outliers. Fig. 4.21 illustrates the trajectories of the second sample set. The reader can see that only two trajectories are inliers, the rest are detected as outliers. Although a gap of 1 m between the cameras exists, the reconstruction and the camera positions are correct.

After L-3 passes bravely the test with the synthetic trajectories, we are interested in how L-3 performs with real image data. We choose for the next experiment Sequence 2 of the Seminar room dataset and let at the very beginning C-1+3 compute the interior orientation. Then, L-3 run over faces which are automatically detected in 2,700 image frames for each camera. The face detector we used is described by Roth et al. [2005]. The first column of Fig. 4.22 shows sample faces in the two cameras at the same time instant. As the Seminar room is not an orthogonal world and L-3 is in such a situation inapplicable, we set the yaw of both cameras manually by measuring the angle between the two window facades. This angle was 62 deg. Tab. 4.10 gives the final estimates of the exterior orientation. The tilt lies

Method	ϕ [deg]	θ [deg]	ψ [deg]	C_X [cm]	C_X^Δ [cm]	C_Y [cm]	C_Y^Δ [cm]	C_Z [cm]
Reference	3.08	5.65	179.38	500.31		837.20		0.00
C-1+3/L-3	2.68	10.55	179.38	515.35	15.04	828.02	9.18	2.55

Table 4.10: L-3's estimates of the exterior orientation.

in the tolerance of the reference, but the pan is underestimated; the error is approximately 5 deg. The reason is the existence of too few line segments and the large lens distortion. Despite the worst error in the position's coordinates is only 15.04 cm, the second and third column in Fig. 4.22 show with 17.84 pixel and 49.30 pixel larger errors between the epipoles and the ground truth; note, that the cameras are still visible in the images. An improved face detector would decrease these errors, although, the error in the tilt indicates a significant contribution of the poor tilt estimate.

4.5 Applications

This section shows two applications in video surveillance for self-calibrating cameras. One application are object trackers in single views [Yilmaz et al., 2006] that are significantly improved in their performance. Due to the perspective distortion, trackers have to resize the object representation, for example a bounding box, however, much work exist in the literature that simply neglect this problem. Trackers and also detectors, that account for resizing the object representation, use on the one hand manually defined homographies to a visible ground plane [Grabner et al., 2007] and on the other hand heuristics based on the object's appearance [Comaniciu et al., 2003]. One can imagine that the former, manual solution is tedious and that the latter solution is susceptible for errors; image features are noisy and can appear or disappear during the tracking, because people are blobs with a large variation in their area during tracking. Sec. 4.5.1 shows a tracking example where tracking based on C-0+3 is superior compared to tracking without the use of geometry.

Sec. 4.5.2 shows results from another important application - the matching of objects across two cameras. We show that in situations where the cameras' visible areas overlap only slightly or even not at all, a reliable matching is possible using geometry or in the non-overlapping case - geometry and the object's dynamics. The calibration results of the previous chapter are used for the examples.

4.5.1 Improved kernel-based tracking

Fig. 4.23 shows the same sequence of someone walking through our lab. We used Comaniciu's kernel-based tracker to follow the head of the person. In the first frame (Frame 1), the bounding box around the head was manually defined. In one case we used Comaniciu et al. [2003]'s approach to resize the bounding box. He proposed to take that box with the smallest error to the template with respect to the Bhattacharya distance between the color histograms. This approach has problems, because the initial template is smaller than all successive template candidates, that is, not enough information is in the template to reliably find the head in later frames. Furthermore, in plane rotations let the color histogram significantly change which also contribute to the bad performance. Now, we used the interior orientation and the rotation to the room that are the results of C-0+3, to initialize a homology between the lower and upper corners of the bounding box [Criminisi et al., 2000]. Thus, the size of the bounding box is deterministically changed and only depends on the location of the box's centroid. The reader can clearly see that the aforementioned problems disappear and that the tracking of the head performs reliable and accurate.

4.5.2 Matching of people

We used triangulation [Hartley and Zisserman, 2004] with the overlapping cameras and the proposed extended triangulation (Sec. 3.3.6) that also incorporates the motion model for matching the people. In all examples, the matching established a match when the reprojection error was smaller than three pixel.

Figures 4.25, 4.26, 4.27 and 4.28 summarize the overlapping case. C-0+3 was used with all datasets. L-0 was only used with the PETS dataset, otherwise, we always used L-1. We believe the reason for L-1's failure on the PETS dataset lied in the poor head detections; the distance between the points was in the order of magnitude of the noise which makes the estimation of the rotation angle sensible. The observation that L-0 worked is another evidence that more knowledge about the orientation will help. The reader can convince himself that all matches are correct. Even more promising, Fig. 4.24 shows correct matches in the non-overlapping case which confirms the feasibility of matching by exploiting geometry and the object's dynamics.

The experiments convince us that the knowledge of geometry plays a key role in robust people tracking. The resolution in parts of the images is low, the perspective distortion is severe and the object's appearance is drastically different. Image features do simply not contain enough information for a robust handover of objects in such scenes. The main problem with geometry is the elaborate, manual calibration which is unrealistic in a practical multi-camera surveillance system.

We experienced during the experiments that the matching was correct in a predominately number of times. Occasionally, a wrong match occurred, because we used sometimes a too large threshold on the re-projection error. It turned out that a threshold of three pixel eliminated these false positives, but on the other hand increased the number of false negatives in particular image frames. Nevertheless, it never happened that a person was completely missed. However, in the case of these datasets an improvement is more a question about the detector than about the calibration which is not a subject of this thesis.

In contrast to Kahn and Shah [2003], we do not believe that observing objects over a period of time provides the most information for robust object handover. Especially, cameras with slightly overlapping views are common in practice. Even if such a robust handover would be conceivable, in such cases, the observation time is simply too short to extract reliable information.

4.6 Discussion

The accuracy of the interior orientation lies in the expected range of 5 %. The obtained results show that the use of many images instead of one image can improve the accuracy significantly. The adaptation is working in scenes when new edge information appears in parts of the image and it is able to prevent the loss of the improved estimates. In the best case, the result of C-3 was ten times better than the result of only using C-0 or C-1. In some scenes the relative error of some estimates was lower than one percent. Interestingly, the focal length of modern lenses gives good initial estimates of the true focal length. In the case of our cameras we have errors below one percent.

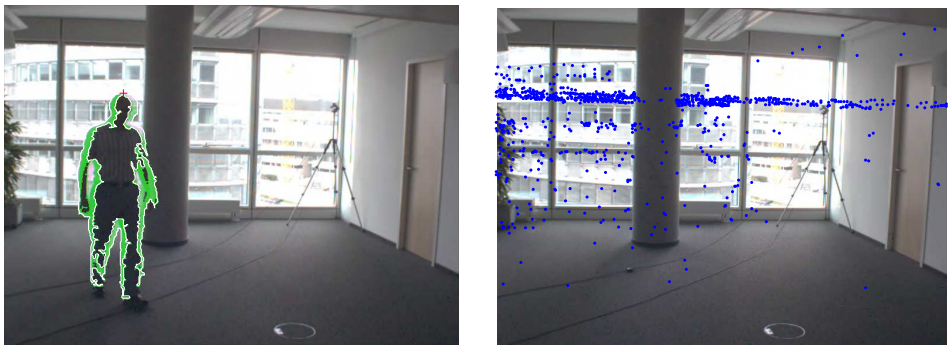
Apart from the success there are also problems. The EM iteration is vulnerable to stuck in local minima, because a severe lens distortion can prevent the method to group correctly the line segments. We observed this problem when only a low number of line segments is present in small areas of the image which is a frequent case in video surveillance. The reader may have the idea to increase the noise level, however, a larger noise level decreases quickly the accuracy of the estimates. We cannot give a definite answer to this problem, but we saw in our experiments that a good initial estimate of the lens distortion can solve the problem. Methods like the one of Devernay Devernay and Faugeras [2001] can help in such situations, unfortunately, we always had to play with the parameters until the method worked successfully.

A further observation is the strong relation between lens distortion and rotation. A poor distortion estimation usually results in a poor rotation estimation which is quickly problematic in the localization. The whole calibration can fail when a poor partial result occurs. Unfortunately, this depends on the one hand on the number and the position of the available

line segments and on the other hand on a potential critical situation of the camera geometry. No automatic method can circumvent this problem, except information from other sensors or by the user is available.

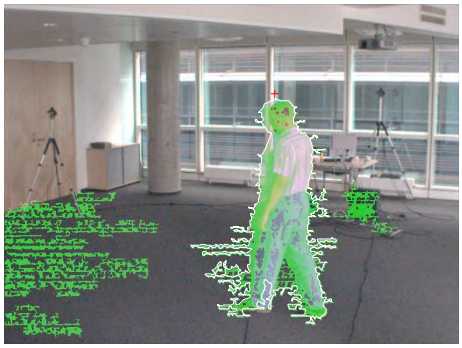
The accuracy of the translation is in all experiments within the expected range of a meter. We show with synthetic and real data of cameras with small overlapping views that our calibration still works while methods like the 8-point algorithm and Svoboda's Multi-Camera Self-Calibration fail. With these methods it is not possible to use automatic detectors, because the delivered positions vary by dozens of pixels between consecutive frames. Our approach is more robust against such variations, because it also uses other information such as the line segments or the a priori knowledge about the cameras. Furthermore, L-3 can be used in both overlapping and non-overlapping views without any constraints on the motion except that it is smooth. With a large non-linear motion of maximal 30 cm change in a half a meter step the results are in the expected range until a three meter gap. With a change of 10 cm which is similar to human motion, L-3 is able to localize the cameras with a four meter gap, although 15 % gross outliers are present. These results encourages further research on L-3.

The experiments show that the rotation is sensible to noisy point correspondences and that an accurate estimation fails when the error is approximately the distance between the points. Unfortunately, current detectors are not that pixel or sub-pixel accurate in their repeating localization of the same 3D point on a person. Furthermore, people tend to walk along common straight lines that cluster the trajectories in small parts of the image. This all confirmed our thesis that rotation should be computed from other sources and not from moving objects. The camera centers rely on moving objects, because they are the only source of information between cameras.

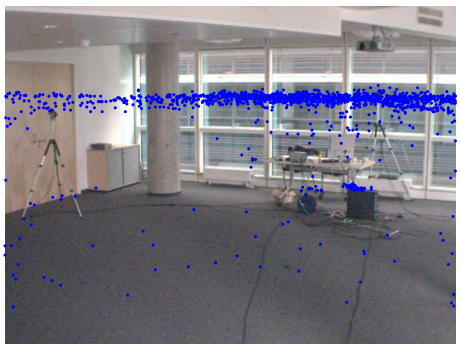


(a) Marlin 0, Frame 1613

(b) Marlin 0, all frames

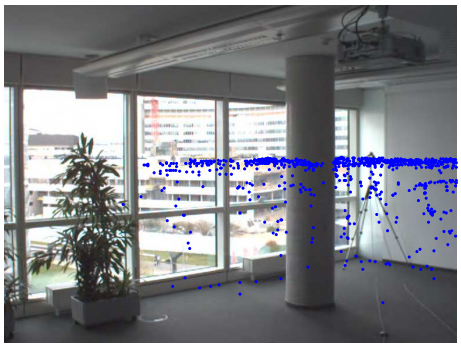


(c) Marlin 1, Frame 1613



(d) Marlin 1, all frames

(e) Marlin 2, Frame 1613



(f) Marlin 2, all frames

Figure 4.15: The generation of point correspondences with our automatic top of the head detection (red crosses). (a), (c) and (e) show three synchronous images of the three cameras. (b), (d) and (f) show the result as blue points within 2,700 image frames. Notice the outliers and the critical collinearity of the points which has its cause in the room floor restricted motion of people.

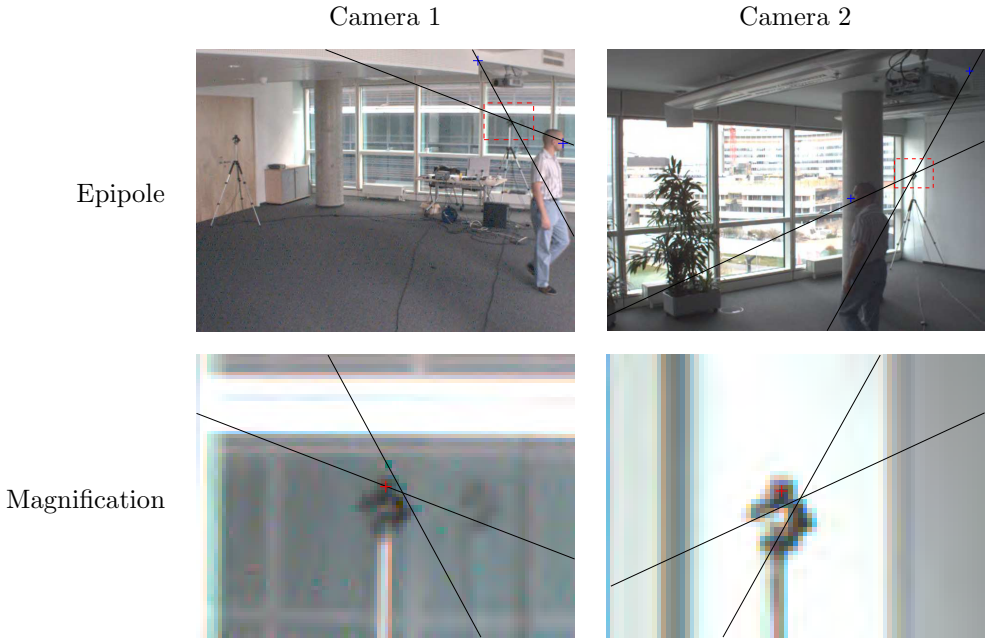


Figure 4.16: The epipoles in Sequence 1 of the Seminar room dataset. The error in the image between the ground truth and the estimates were 3.54 pixel for Camera 1 and 2.96 pixel for Camera 2. The largest error between the manually defined points in the background and their corresponding epipolar lines is 6.27 pixel (the nose in Camera 2). The smallest error was 0.95 pixel (point on ceiling in Camera 1).

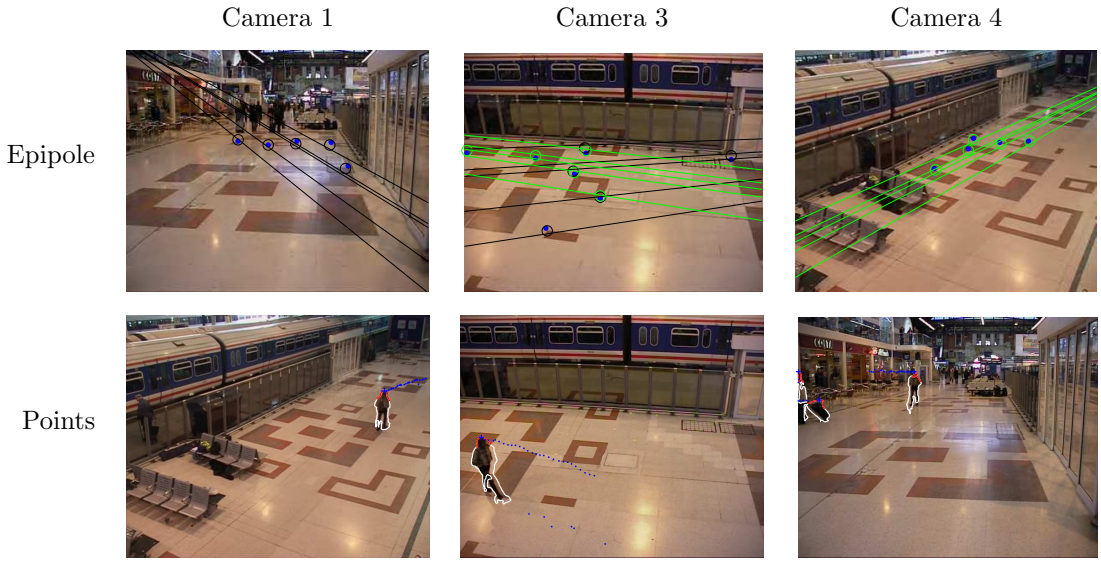


Figure 4.17: (Left column: Results of the camera calibration. The ground truth points are drawn in blue. The circles show the re-projected points. The epipolar lines of the corresponding ground truth points are also drawn. Correspondences between Camera 1 and 3 are shown in black and between Camera 3 and 4 in green. Right column: The output of the head point detector. Blue are the head points. Some outliers can be recognized.

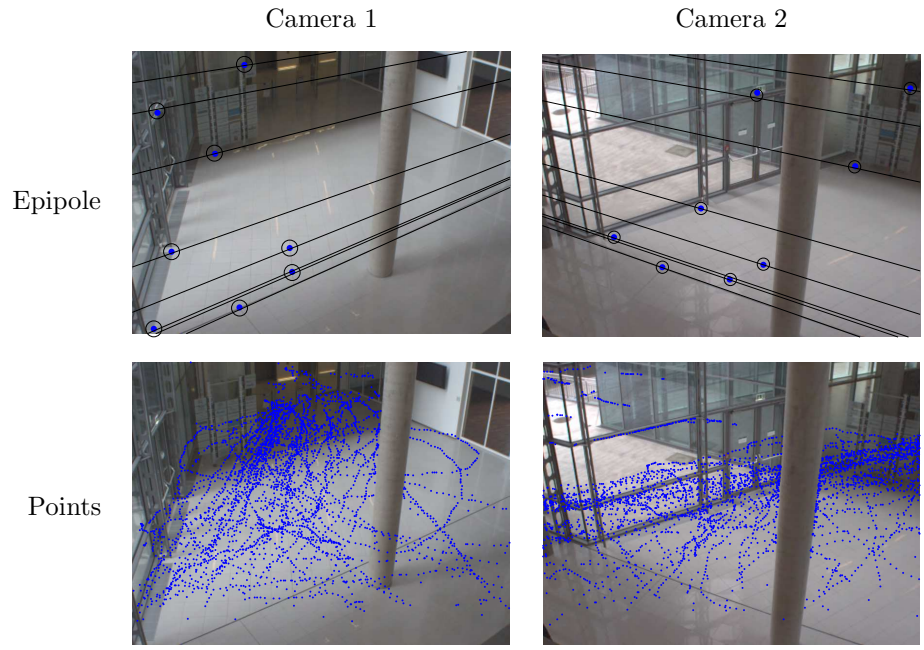
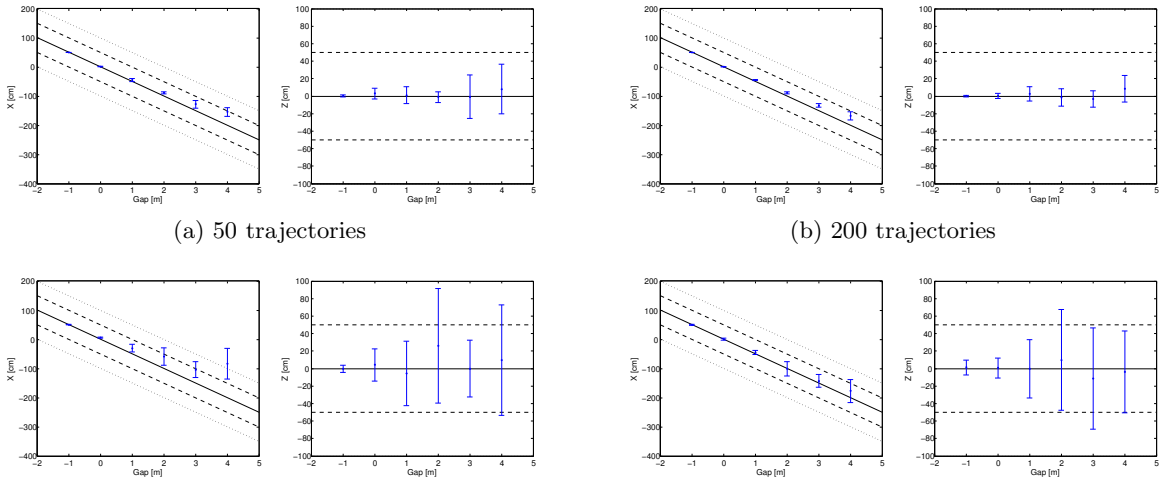


Figure 4.18: The first row shows the correct epipoles and the low error between the manually defined points and their reprojection after triangulation. The worst error was smaller than 3 pixel. The second row shows the point correspondences that are generated by walking people.



(a) 50 trajectories

(b) 200 trajectories

(c) 30 cm worst case velocity variation

(d) 15 % outlier

Figure 4.19: This Figure shows the numbers of Tab. 4.9 graphically. The 95 % confidence intervals around the median values are drawn in blue. The dashed, black line shows the half meter error boundary, the dotted, black line the one meter boundary. The estimates must lie within 1 m, because the localization is otherwise useless for people matching.

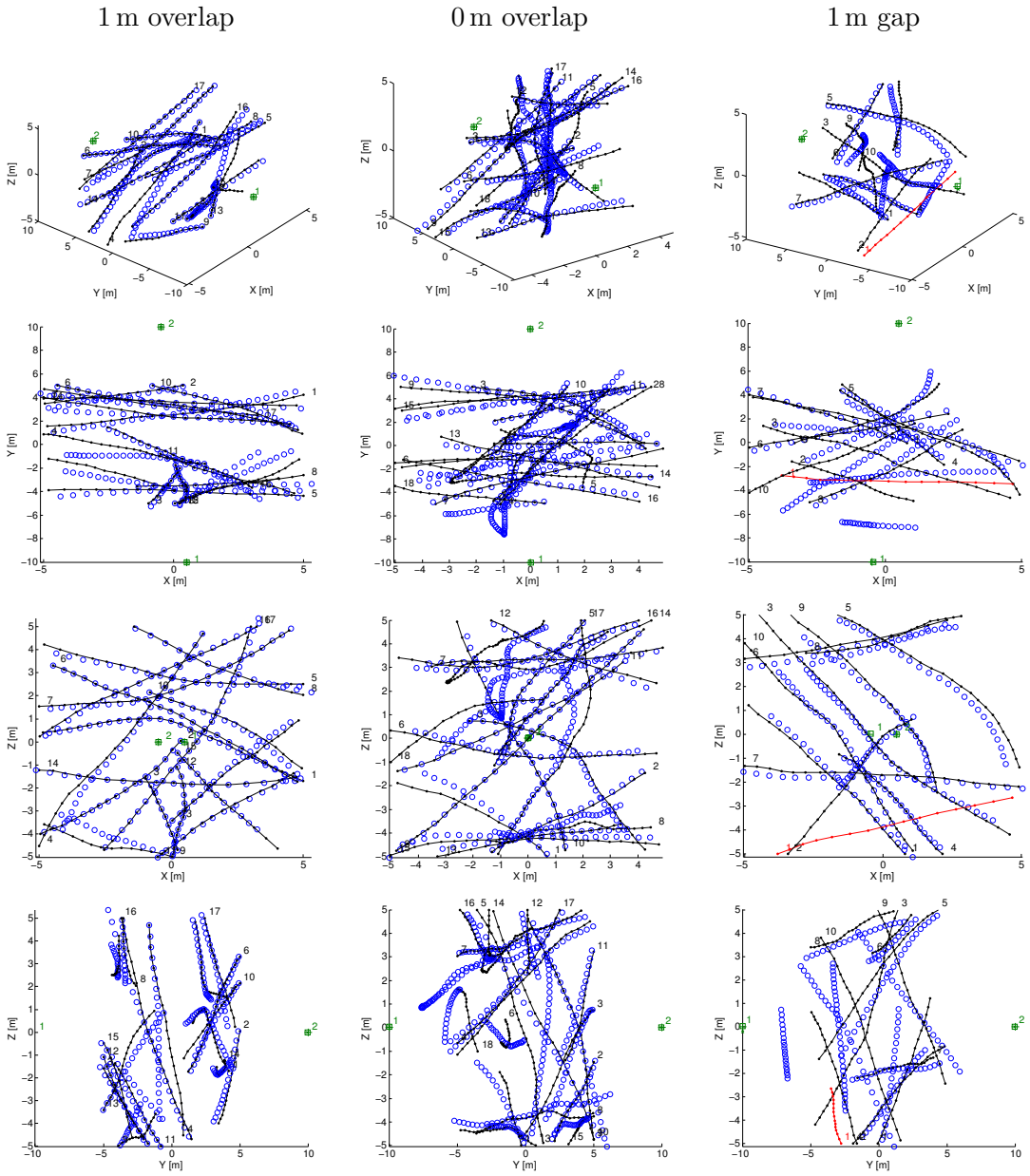


Figure 4.20: Localization by L-3 works with overlapping views (left column) and with non-overlapping views (middle and right column). These 12 images show the successful reconstruction (blue circles) of the given trajectories (black lines). The top row shows the 3D space. To improve the visibility, all other columns show projections of this space onto the plane spanned by two of the three dimensions. The cameras are drawn in green; ground truth with plus, median with a square and the result of the second sample set with a cross. The red line in the last column is an outlier trajectory.

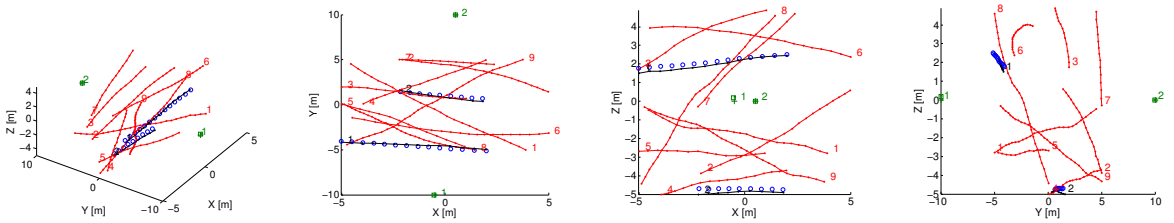


Figure 4.21: These images show the successful outlier detection (red lines) of L-3. The gap between the views is 1 m. The two inlier trajectories (black lines) are correctly reconstructed (blue circles). The positions of the cameras are green (see Fig. 4.20 for more description).

Figure 4.22: Camera localization with Sequence 2 of the Seminar room dataset. Left: Sample face detections that are used. Middle: The estimated epipoles in both cameras. Right: A magnification of the middle images within the red rectangles.

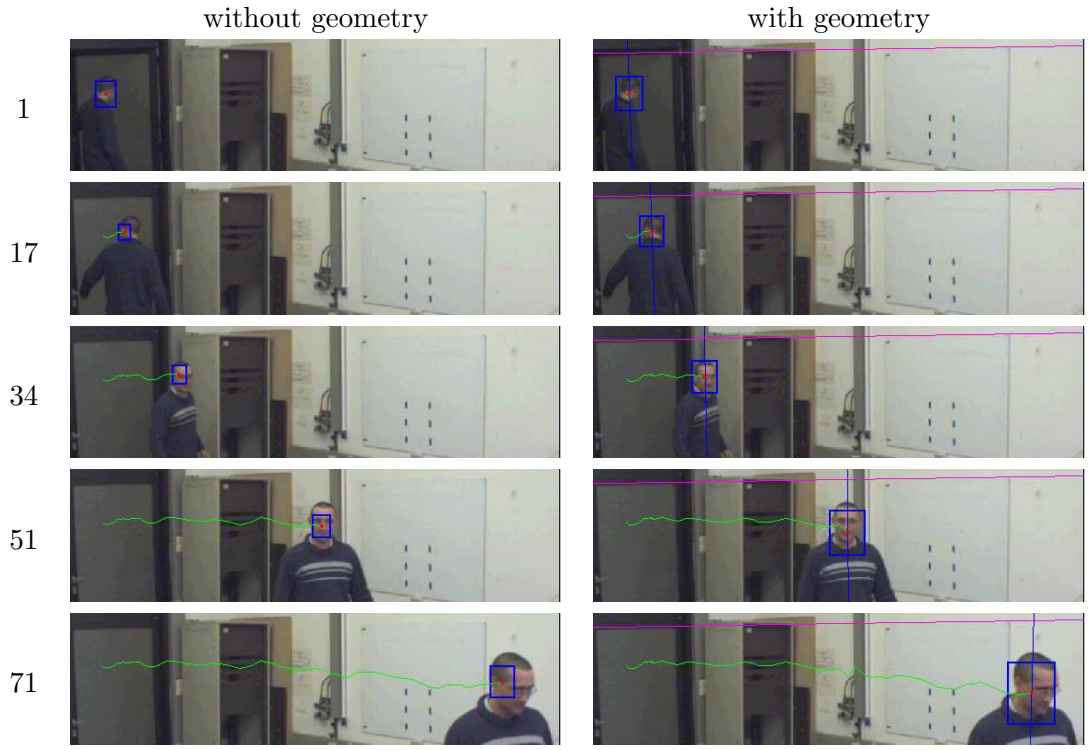
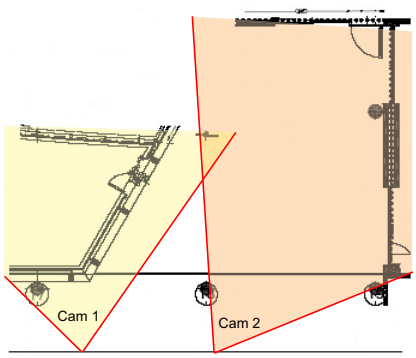


Figure 4.23: The performance improvement of kernel-based tracking (left column) by deterministic resizing of the bounding box is shown (right column). Each row shows the same image frame and the result of both variants. Note especially in the last frame the large error of resizing by image intensities. The head is not reliably followed, because in plane rotations and large changes of the head's size happen. The color histogram of the template (Frame 1) that is manually selected contains only parts of the information that is available in successive frames. The purple line shows the horizon and the blue line the vertical axis of the person.



(a) Floorplan

(b) A person walks from Cam 1 (yellow) to Cam 2 (orange)

Figure 4.24: People matching without overlap. Entrance hall: (a) Camera setup, (b) Successful tracking (blue rectangles) of a person, despite the large gap of 2 m between the views. Black rectangles are detections without a match. Note the change in the person's appearance due to shade effects.

Figure 4.25: People matching with the Seminar room 1 dataset. The images show the matches at three different time instants. Black rectangles are detected people, blue rectangles show a match. Shade and illumination influence drastically the appearance of the same person. In some images only parts of the body are captured.

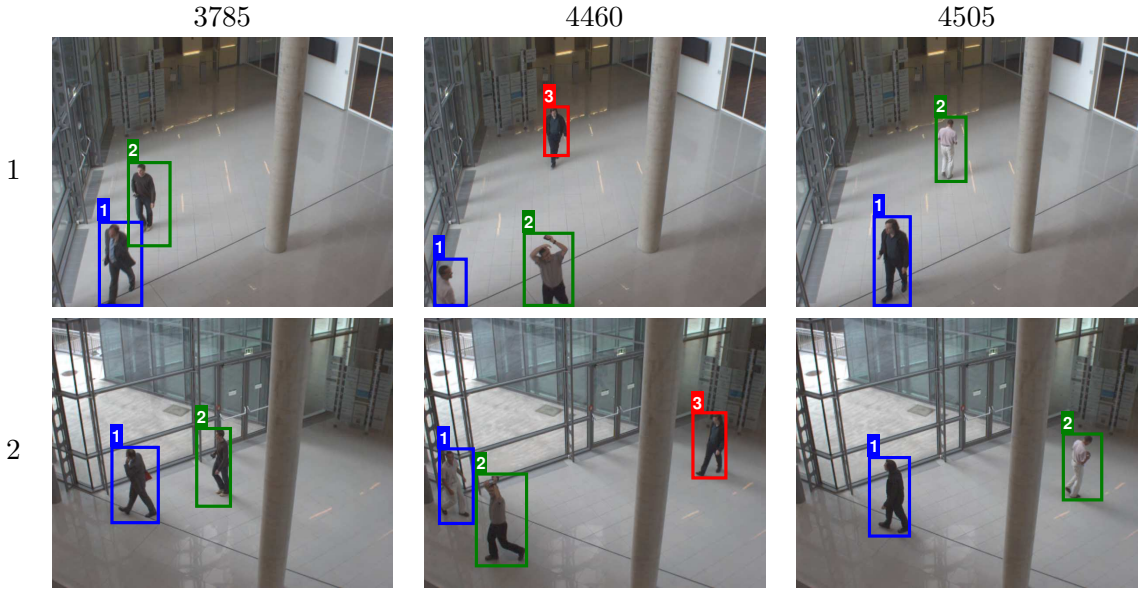


Figure 4.26: Correct people matching results on the Tech-Gate 3 dataset.

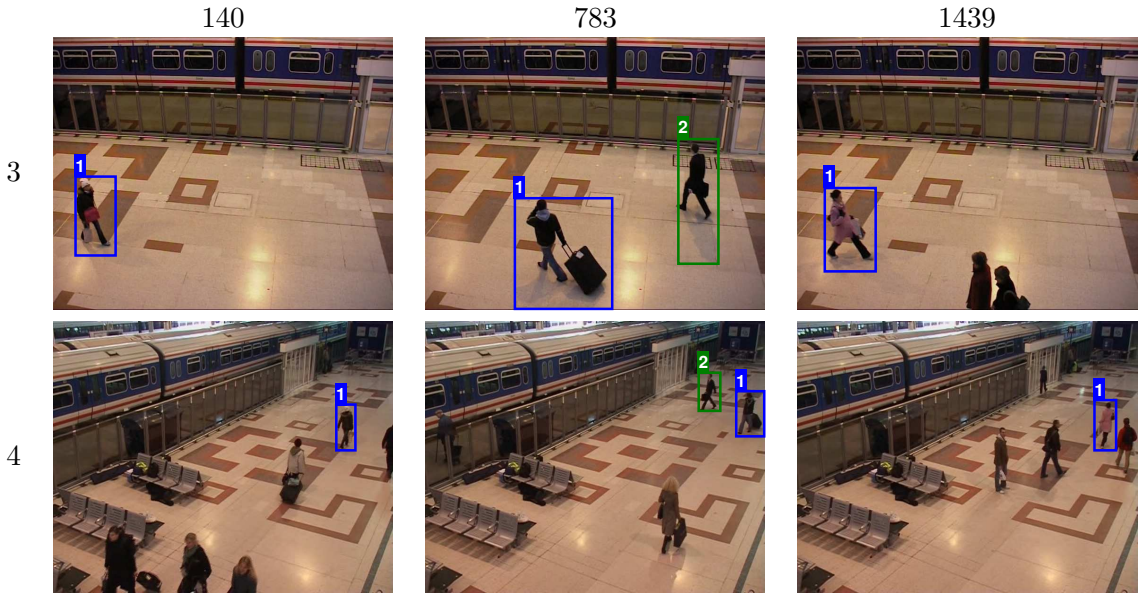


Figure 4.27: People matching between Camera 3 and Camera 4 of the PETS 2006 dataset. Consider the person in Frame 140. The red bag is visible in Camera 3 but is invisible in Camera 4; this will mislead color-based feature matching.

Figure 4.28: People matching between Camera 3 and Camera 1 of the PETS 2006 dataset. The correct match of Frame 1352 is hard to realize only with image features, because of similar intensities. Further difficulties are the low resolution and the wide baseline.

Chapter 5

Conclusion

This thesis treated self-calibrating cameras in video surveillance. After a careful introduction (Chap. 1), the thesis motivated a two step approach, namely, the self-calibration of the interior orientation and parts of the exterior orientation for each single camera separately and then on top of these results the self-calibration of the remaining parts of the exterior orientation between the cameras, that is, the rotational ambiguity and the camera centers. Jaynes [2004] already followed this approach, however, he never made explicit that the use of line segments is necessary to successfully calibrate cameras with slightly overlapping views. Apart from this, his proposed method on the one hand does not work in non-overlapping views and on the other hand assumes a known interior orientation.

5.1 A single camera

A prominent paper that use line segments for the self-calibration of the interior-orientation is the one by Caprile and Torre [1990], who exploit the rich geometric information of vanishing points. Unfortunately, vanishing points are often inaccurate, because small changes in the endpoints of the concurrent line segments can cause large changes in their intersection point, that is, the vanishing point. Therefore, Chap. 2 gave a survey of the bunch of work that tries to overcome this problem. We collected all the observations of the last couple of years and proposed a new method that incorporates Liebowitz's error in the endpoints [Liebowitz and Zisserman, 1999], a novel RANSAC based initialization with similarities mentioned in Rother's PhD thesis [Rother, 2003], the direct optimization in the parameters instead optimizing the vanishing points [Schindler and Dellaert, 2004], the EM-framework [Antone and Teller, 2000], necessary geometric criterions [Van den Heuvel, 1998] and the consideration of the lens distortion [Bräuer-Burchardt, 2004; Van den Heuvel, 1999]. Then, we extended this method to a novel incremental method that has clear advantages when the line segments appear and disappear over time, because they are on moving objects or because the illumination

influences the edge information in an image. We do not use ordinary exponential averaging, because the method would loose a good estimation in case information disappears. Instead, the method has a novel update scheme based on the uncertainty of the line segments and in the spirit of Simulated Annealing.

The experiments with realistic scenarios in Chap. 4 showed that the relative error in the estimates compared to the ground truth is below five percent. This error still allows the use of the parameters for surveillance applications, because the resulting errors in the world were always between half a meter and a meter which is the average width of a person. When the line segments are evenly distributed within the image and all three orthogonal directions are well represented, the relative error drops below one percent which is in the range of conventional methods using calibration patterns [Tsai, 1987].

5.2 Two cameras

Chap. 3 presented methods to self-calibrate the remaining rotational ambiguity and the camera centers. Contrary to a bunch of work that compute homographies between cameras and a common ground plane by assuming and tracking points on planar walking people, we use Nister's 5-point algorithm to estimate the Essential matrix and thus the rotational ambiguity and the camera centers. Although not necessary, we assume a man-made world, because three of the five point correspondences are then known by the vanishing points which makes the method more robust. Remark: We assume that the scene contains a common ground plane, but the people must not walk on this plane.

Perhaps the most significant contribution is the extension of Rother's DRP-method by considering the points as moving. Now the method is useable in overlapping as in non-overlapping views. The motion of the objects compensates the lack of point correspondence. The reader can think about generalizing correspondence from the level of points to the level of trajectories. This allows a similar optimization as it is done by Rahimi et al. [2004] where the positions in the trajectories and the camera centers are simultaneously estimated, but without assuming planar motion.

The chapter showed further that an exhaustive search over the four possible rotations finds the correct rotation. The set is finite thanks to the Manhattan world assumption. The method is able to find the correct rotation by evaluating the re-projection error. The smallest re-projection error gives the correct rotation. In contrast to Rahimi's method, the extended DRP-method is linear and finds a global solution in closed-form by SVD. We showed that incorporating the method into a RANSAC framework makes it robust. This embedding is only possible by extending the triangulation with the motion model. However, the error that is then minimized is algebraic and has no geometric meaning.

The estimation accuracy of the camera centers is in all realistic experiments within the expected range of a meter. Thereby it does not matter whether the views overlap or not. With a large non-linear motion of maximal 30 cm change in a half a meter step the results are in the expected range until a three meter gap. With a change of 10 cm which is similar to human motion, the extended DRP-method is able to localize the cameras with a four meter gap, although 15 % gross outliers are present.

5.3 Future work

There is still plenty of work that could be done:

Gradients instead of line segments: To utilize the full structural information in the images, one could use the gradients directly instead of using line segment detectors (Sec. 2.1). One reason why we used line segment detectors was computational performance. More research is here necessary to speed up the processing on gradients.

Self-calibration with low structure: Missing line segments or noisy line segments have a great influence. Hence, it would be beneficial to combine further approaches with the existing method and incorporate other sources of information. For example, walking humans contain information of the interior orientation as they might walk along a straight line or at constant speed or they do not change in height. Another interesting approach is by Saxena et al. [2008] who used Markov Random Fields to model the depth between different areas in the image. They used sample images to train a functional relationship between the images and the given depth map. Then they use this function to compute the depth map of unknown images. The idea of learning the interior orientation would be interesting to investigate. Some work is already done by Gallagher [2002].

Points on people: The experiments showed that the top of the head is much more stable as the centroid and in contrast to the centroid, it is a frontier point [Sinha et al., 2004]. However, the motion segmentation is still not robust, thus much more work has to be done to reliably and robustly find with high accuracy the top of the head. Some approaches using head-shoulder models for tracking seem appropriate [Jin and Mokhtarian, 2007].

Incremental SVD: The more trajectories are available and the larger the video frame-rate is, the more intractable the computation of the SVD will be. Thus, an incremental SVD is necessary [Brand, 2002], but how can we compute the rotation and how can we make the optimization robust to gross outliers? Some ideas, how to compute the

rotation can be found in the recent work of Li and Hartley [2007]. They search with a branch-and-bound method in the space of all rotations.

Algebraic error: We have seen that the error being minimized is algebraic. Instead of defining the error on the vector product one could formulate the error on the dot product which has a geometric meaning, because it relates to the angle between the rays.

Moving cameras: The DRP-method could also be extended to the case when the cameras move. This would mean that the camera centers are also positions of camera trajectories with proper motion constraints.

Synchronization: This thesis assumes synchronized cameras, but in practice this is hard to require. Synchronization is establishable when the geometry is known [Caspi and Irani, 2002; Piao and Sato, 2007]. So it is a similar problem to correspondence which is found by the Hungarian algorithm, once the geometry is known. Perhaps all the synchronization, the correspondence and the geometry can be solved by one single approach.

More than two cameras: The DRP-method was developed to reconstruct structure with many cameras, hence, it is naturally extendable to more than two cameras.

Distributed algorithm: All methods in this thesis are centralized. They all run on a single computer. When we consider camera networks, it would be desirable to have a scalable and distributed solution.

Appendix A

Image datasets

The absolute values and their uncertainties of our results are summarized in Tab. A.1, A.2 and A.3. ϕ, θ and ψ are in the same order the roll, pitch and yaw angle of the rotation matrix. $\mathbf{v}_X$, $\mathbf{v}_Y$ and $\mathbf{v}_Z$ are the vanishing points that corresponds to the X-, $Y-$ and Z-axis of the world coordinate system.

Cam.	Met.	f [pixel]	$\mathbf{p}$ [pixel]	$\mathbf{c}_k$ [pixel]	k [·]	ϕ [deg]	θ [deg]	ψ [deg]	$\Pr(\mathbf{v}_X)$ [%]	$\Pr(\mathbf{v}_Y)$ [%]	$\Pr(\mathbf{v}_Z)$ [%]
1	C-0	1082.21 ± 11.60	$\begin{pmatrix}359.00 \pm 13.31\\287.00 \pm 06.42\end{pmatrix}$	$\begin{pmatrix}359.00 \pm 0.00\\287.00 \pm 0.00\end{pmatrix}$	$.0123 \pm .0019$	-0.03 ± 10.85	10.57 ± 102.35	7.04 ± 8.00	.28	.28	.02
	C-0+2	no difference to C-0									
	C-0+3	854.05 ± 22.95	$\begin{pmatrix}371.71 \pm 26.47\\277.57 \pm 12.53\end{pmatrix}$	$\begin{pmatrix}369.31 \pm 10.86\\278.71 \pm 08.07\end{pmatrix}$	$.0264 \pm .0021$	-0.12 ± 09.01	10.29 ± 101.83	5.84 ± 5.96	.13	.29	.28
2	C-0	888.58 ± 5.64	$\begin{pmatrix}359.00 \pm 5.60\\287.00 \pm 4.29\end{pmatrix}$	$\begin{pmatrix}359.00 \pm 0.00\\287.00 \pm 0.00\end{pmatrix}$	$.0008 \pm .0013$	2.17 ± 38.65	2.83 ± 154.64	-13.98 ± 4.65	.29	.27	.07
	C-0+2	no difference to C-0									
	C-0+3	no difference to C-0									
3	C-0	792.75 ± 4.01	$\begin{pmatrix}359.00 \pm 4.38\\287.00 \pm 2.51\end{pmatrix}$	$\begin{pmatrix}341.65 \pm 3.01\\332.33 \pm 7.60\end{pmatrix}$	$.0335 \pm .0012$	1.18 ± 4.93	28.58 ± 106.13	10.98 ± 13.92	.50	.22	.04
	C-0+2	828.54 ± 2.99	$\begin{pmatrix}359.00 \pm 3.16\\286.98 \pm 1.94\end{pmatrix}$	$\begin{pmatrix}390.48 \pm 3.42\\310.10 \pm 4.81\end{pmatrix}$	$.0377 \pm .0011$	1.15 ± 4.56	30.52 ± 116.63	10.11 ± 11.87	.09	.44	.21
	C-0+3	799.88 ± 3.72	$\begin{pmatrix}351.51 \pm 4.04\\236.74 \pm 2.31\end{pmatrix}$	$\begin{pmatrix}341.65 \pm 3.01\\332.33 \pm 7.60\end{pmatrix}$	$.0353 \pm .0011$	1.49 ± 4.76	27.95 ± 103.22	10.55 ± 12.93	.09	.48	.20
4	C-0	895.30 ± 3.16	$\begin{pmatrix}359.00 \pm 3.12\\287.00 \pm 2.23\end{pmatrix}$	$\begin{pmatrix}359.00 \pm 0.00\\287.00 \pm 0.00\end{pmatrix}$	$-.0189 \pm .0008$	-1.03 ± 0.53	21.35 ± 32.35	29.49 ± 2.26	.32	.18	.13
	C-0+2	895.11 ± 2.31	$\begin{pmatrix}359.02 \pm 2.24\\286.99 \pm 1.62\end{pmatrix}$	$\begin{pmatrix}359.02 \pm 2.25\\286.98 \pm 1.64\end{pmatrix}$	$-.0164 \pm .0007$	-1.08 ± 0.54	21.41 ± 32.43	29.52 ± 2.37	.37	.20	.14
	C-0+3	890.98 ± 2.07	$\begin{pmatrix}386.48 \pm 2.03\\224.10 \pm 1.45\end{pmatrix}$	$\begin{pmatrix}384.33 \pm 2.17\\227.55 \pm 1.61\end{pmatrix}$	$-.0228 \pm .0009$	-0.90 ± 1.13	17.94 ± 28.10	29.06 ± 3.42	.36	.19	.17

Table A.1: Results with the PETS dataset.

Seq.	Cam.	Met.	f [pixel]	$\mathbf{p}$ [pixel]	$\mathbf{c}_k$ [pixel]	k [·]	ϕ [deg]	θ [deg]	ψ [deg]	$\Pr(\mathbf{v}_X)$ [%]	$\Pr(\mathbf{v}_Y)$ [%]	$\Pr(\mathbf{v}_Z)$ [%]
1	1	C-1+0	963.86 ± 0.00	$\begin{pmatrix} 389.00 \pm 0.00 \\ 290.00 \pm 0.00 \end{pmatrix}$	$\begin{pmatrix} 400.06 \pm 07.28 \\ 323.36 \pm 13.66 \end{pmatrix}$	$.0475 \pm .0038$	0.12 ± 0.5	8.89 ± 1.31	21.72 ± 2.42	.42	.23	.00
		C-1+2	963.86 ± 0.00	$\begin{pmatrix} 389.00 \pm 0.00 \\ 290.00 \pm 0.00 \end{pmatrix}$	$\begin{pmatrix} 367.72 \pm 6.39 \\ 313.63 \pm 9.07 \end{pmatrix}$	$.0502 \pm .0028$	0.08 ± 0.55	8.72 ± 1.37	22.43 ± 2.82	.03	.40	.22
		C-1+3	941.08 ± 12.31	$\begin{pmatrix} 407.85 \pm 14.56 \\ 275.44 \pm 6.74 \end{pmatrix}$	$\begin{pmatrix} 402.47 \pm 06.32 \\ 321.03 \pm 10.74 \end{pmatrix}$	$.0474 \pm .0030$	0.00 ± 0.70	8.03 ± 3.05	21.31 ± 4.46	.03	.42	.23
	2	C-1+0	963.86 ± 0.00	$\begin{pmatrix} 389.00 \pm 0.00 \\ 290.00 \pm 0.00 \end{pmatrix}$	$\begin{pmatrix} 356.25 \pm 7.79 \\ 307.70 \pm 5.65 \end{pmatrix}$	$.0418 \pm .0022$	0.21 ± 0.53	2.31 ± 0.63	-37.68 ± 0.49	.21	.21	.13
		C-1+2	963.86 ± 0.00	$\begin{pmatrix} 389.00 \pm 0.00 \\ 290.00 \pm 0.00 \end{pmatrix}$	$\begin{pmatrix} 382.74 \pm 5.84 \\ 295.37 \pm 3.47 \end{pmatrix}$	$.0511 \pm .0019$	0.25 ± 0.55	2.28 ± 0.66	-38.36 ± 0.96	.15	.21	.22
		C-1+3	1005.98 ± 5.23	$\begin{pmatrix} 350.51 \pm 5.23 \\ 296.93 \pm 3.97 \end{pmatrix}$	$\begin{pmatrix} 377.54 \pm 4.89 \\ 298.97 \pm 3.95 \end{pmatrix}$	$.0458 \pm .0020$	0.36 ± 0.70	2.34 ± 2.25	-38.24 ± 6.70	.16	.22	.23
2	1	C-1+0	963.86 ± 0.00	$\begin{pmatrix} 389.00 \pm 0.00 \\ 290.00 \pm 0.00 \end{pmatrix}$	$\begin{pmatrix} 343.29 \pm 07.98 \\ 374.09 \pm 16.34 \end{pmatrix}$	$.0358 \pm .0027$	0.07 ± 1.08	8.49 ± 1.64	22.06 ± 3.34	.40	.21	.01
		C-1+2	963.86 ± 0.00	$\begin{pmatrix} 389.00 \pm 0.00 \\ 290.00 \pm 0.00 \end{pmatrix}$	$\begin{pmatrix} 369.54 \pm 05.66 \\ 338.35 \pm 10.63 \end{pmatrix}$	$.0446 \pm .0025$	0.10 ± 1.13	8.70 ± 1.75	22.45 ± 3.75	.02	.44	.31
		C-1+3	949.06 ± 10.91	$\begin{pmatrix} 395.03 \pm 12.82 \\ 287.88 \pm 6.07 \end{pmatrix}$	$\begin{pmatrix} 377.35 \pm 06.00 \\ 351.11 \pm 11.69 \end{pmatrix}$	$.0419 \pm .0025$	0.05 ± 1.26	8.75 ± 4.35	22.13 ± 5.53	.02	.41	.23
	2	C-1+0	963.86 ± 0.00	$\begin{pmatrix} 389.00 \pm 0.00 \\ 290.00 \pm 0.00 \end{pmatrix}$	$\begin{pmatrix} 398.05 \pm 7.85 \\ 304.01 \pm 5.08 \end{pmatrix}$	$.0408 \pm .0015$	0.46 ± 0.49	2.11 ± 0.59	-37.37 ± 0.56	.23	.26	.11
		C-1+2	963.86 ± 0.00	$\begin{pmatrix} 389.00 \pm 0.00 \\ 290.00 \pm 0.00 \end{pmatrix}$	$\begin{pmatrix} 404.83 \pm 5.65 \\ 297.71 \pm 4.22 \end{pmatrix}$	$.0500 \pm .0017$	0.42 ± 0.51	1.94 ± 0.61	-37.80 ± 0.95	.13	.24	.34
		C-1+3	1000.44 ± 7.00	$\begin{pmatrix} 312.16 \pm 6.77 \\ 297.15 \pm 5.57 \end{pmatrix}$	$\begin{pmatrix} 360.28 \pm 6.59 \\ 300.91 \pm 4.18 \end{pmatrix}$	$.0429 \pm .0012$	0.58 ± 0.56	2.20 ± 2.06	-36.44 ± 5.90	.14	.24	.29

Table A.2: Results with the Seminar room dataset.

Seq.	Cam.	Met.	f [pixel]	$\mathbf{p}$ [pixel]	$\mathbf{c}_k$ [pixel]	k [·]	ϕ [deg]	θ [deg]	ψ [deg]	$\Pr(\mathbf{v}_X)$ [%]	$\Pr(\mathbf{v}_Y)$ [%]	$\Pr(\mathbf{v}_Z)$ [%]
2	1	C-1+0	963.86 ± 0.00	$\begin{pmatrix}389.00 \pm 0.00\\290.00 \pm 0.00\end{pmatrix}$	$\begin{pmatrix}396.32 \pm 7.10\\299.66 \pm 4.32\end{pmatrix}$	$.0499 \pm .0019$	-1.16 ± 0.56	41.72 ± 1.02	-6.84 ± 0.67	.27	.19	.06
		C-1+2	963.86 ± 0.00	$\begin{pmatrix}389.00 \pm 0.00\\290.00 \pm 0.00\end{pmatrix}$	$\begin{pmatrix}382.23 \pm 5.59\\309.48 \pm 3.90\end{pmatrix}$	$.0482 \pm .0016$	-1.19 ± 0.59	41.86 ± 1.26	-6.82 ± 0.72	.10	.24	.41
		C-1+3	971.89 ± 8.01	$\begin{pmatrix}376.96 \pm 8.48\\290.42 \pm 5.32\end{pmatrix}$	$\begin{pmatrix}383.21 \pm 6.52\\300.60 \pm 4.71\end{pmatrix}$	$.0438 \pm .0017$	-0.73 ± 1.70	41.93 ± 3.73	-6.37 ± 1.94	.09	.22	.29
	2	C-1+0	963.86 ± 0.00	$\begin{pmatrix}389.00 \pm .00\\290.00 \pm .00\end{pmatrix}$	$\begin{pmatrix}375.81 \pm 3.85\\342.78 \pm 6.50\end{pmatrix}$	$.0395 \pm .0012$	0.48 ± 0.87	25.03 ± 1.01	-23.90 ± 1.56	.48	.05	.27
		C-1+2	963.86 ± 0.00	$\begin{pmatrix}389.00 \pm .00\\290.00 \pm .00\end{pmatrix}$	$\begin{pmatrix}366.53 \pm 3.12\\339.26 \pm 4.53\end{pmatrix}$	$.0427 \pm .0008$	0.49 ± 0.87	25.02 ± 1.04	-23.96 ± 1.59	.08	.22	.44
		C-1+3	974.98 ± 7.39	$\begin{pmatrix}380.87 \pm 8.27\\274.04 \pm 4.17\end{pmatrix}$	$\begin{pmatrix}364.99 \pm 3.89\\305.23 \pm 5.83\end{pmatrix}$	$.0410 \pm .0012$	0.67 ± 1.27	24.2 ± 2.76	-23.6 ± 2.93	.07	.22	.46
3	1	C-1+0	963.86 ± 0.00	$\begin{pmatrix}389.00 \pm 0.00\\290.00 \pm 0.00\end{pmatrix}$	$\begin{pmatrix}389.00 \pm 0.00\\290.00 \pm 0.00\end{pmatrix}$	$.0246 \pm .0006$	-1.14 ± 1.35	29.66 ± 1.02	-35.84 ± 0.65	.46	.17	.00
		C-1+2	no difference to C-1+0									
		C-1+3	996.56 ± 5.56	$\begin{pmatrix}414.97 \pm 5.02\\296.26 \pm 3.79\end{pmatrix}$	$\begin{pmatrix}409.72 \pm 3.76\\296.98 \pm 2.99\end{pmatrix}$	$.0250 \pm .0006$	-1.78 ± 2.67	30.21 ± 1.79	-37.05 ± 3.85	.03	.17	.43
	2	C-1+0	963.86 ± 0.00	$\begin{pmatrix}389.00 \pm 0.00\\290.00 \pm 0.00\end{pmatrix}$	$\begin{pmatrix}423.35 \pm 5.79\\287.48 \pm 5.71\end{pmatrix}$	$.0314 \pm .0012$	-0.35 ± 0.44	30.4 ± 0.54	-33.31 ± 1.07	.38	.18	.00
		C-1+2	963.86 ± 0.00	$\begin{pmatrix}389.00 \pm 0.00\\290.00 \pm 0.00\end{pmatrix}$	$\begin{pmatrix}412.52 \pm 5.11\\288.49 \pm 4.60\end{pmatrix}$	$.0332 \pm .0010$	-0.37 ± 0.45	30.53 ± 0.60	-33.22 ± 1.12	.01	.20	.39
		C-1+3	980.13 ± 25.16	$\begin{pmatrix}398.49 \pm 25.95\\287.40 \pm 15.00\end{pmatrix}$	$\begin{pmatrix}405.18 \pm 5.89\\287.71 \pm 5.53\end{pmatrix}$	$.0319 \pm .0013$	-0.62 ± 1.98	30.73 ± 2.20	-33.35 ± 2.93	.01	.21	.38

Table A.3: Results with the Tech-Gate dataset.

Bibliography

Ahmed, M. and Farag, A. (2005). Nonmetric calibration of camera lens distortion: Differential methods and robust estimation. *IEEE Transactions on Image Processing (PAMI)*, 14(8):1215–1230.

Anjum, N., Taj, M., and Cavallaro, A. (2007). Relative position estimation of non-overlapping cameras. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 281–284, Honolulu.

Antone, M. E. and Teller, S. (2000). Automatic recovery of relative camera rotations for urban scenes. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 282–289, Hilton Head Island, SC, USA.

Auvinet, E., Grossmann, E., Rougier, C., Dahmane, M., and Meunier, J. (2006). Left-luggage detection using homographies and simple heuristics. In *Proceedings of the 9th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, pages 51–58, New York. IEEE.

Badler, N. I. (1974). Three dimensional motion from two-dimensional picture sequences. *Proceedings of the 2nd International Joint Conference on Pattern Recognition*, pages 157–161.

Ballard, D. and Brown, C. (1982). *Computer Vision*. Prentice Hall.

Barnard, S. T. (1983). Interpreting perspective images. *Artificial Intelligence*, 21(4):435–462. Elsevier Science B.V.

Bay, H., Ferrari, V., and Gool, L. V. (2005). Wide-baseline stereo matching with line segments. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 329–336, San Diego, US. IEEE.

Becker, S. and Bove, M. V. (1995). Semiautomatic 3-d model extraction from uncalibrated 2-d camera views. In *SPIE Symposium on Electronic Imaging*, San Jose.

Bingham, C. (1974). An antipodally symmetric distribution on the sphere. *The anals of statistics*, 2(6):1201–1225.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Black, J. and Ellis, T. (2001). Multi view image surveillance and tracking. In *Proceedings of the 2th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, Kauai, Hawaii.

Black, J., Ellis, T., and Rosin, P. (2002). Multi view image surveillance and tracking. In *Proceedings of the Workshop on Motion and Video Computing*, pages 169–174, Orlando, Florida, US. IEEE.

Brand, M. (2002). Incremental singular value decomposition of uncertain data with missing values. Technical Report 24, MERL - A Mitsubishi Electric Research Laboratory.

Bräuer-Burchardt, C. (2004). A simple new method for precise lens distortion correction of low cost camera systems. In *Proceedings of the 24th Pattern Recognition Symposium (DAGM)*, pages 570–577.

Bräuer-Burchardt, C. and Voss, K. (2002). Automatic correction of weak radial lens distortion in single views of urban scenes using vanishing points. In *Proceedings of the International Conference on Image Processing (ICIP)*, volume 3, pages 865–868.

Brillault-O'Mahony, B. (1991). New method for vanishing point detection. *Journal of Computer Vision, Graphics and Image Processing*, 54(2):289–300.

Brown, D. C. (1971). Close-range camera calibration. In *Proceedings of the Symposium on Close-Range Photogrammetry*, Urbana, Illinois.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 8(6):679–698.

Cantoni, V., Lombardi, L., Porta, M., and Sicard, N. (2001). Vanishing point detection: Representation analysis and new approaches. In *Proceedings of the 11th International Conference on Image Analysis and Processing*, Palermo.

Caprile, B. and Torre, V. (1990). Using vanishing points for camera calibration. *International Journal of Computer Vision (IJCV)*, 4:127–140.

Caspi, Y. and Irani, M. (2002). Aligning non-overlapping sequences. *International Journal of Computer Vision (IJCV)*, 48(1):39–51.

Caspi, Y., Simakov, D., and Irani, M. (2006). Feature-based sequence-to-sequence matching. *International Journal of Computer Vision (IJCV)*, 68(1):53–64.

Cipolla, R. and Boyer, E. (1998). 3d model aquisition from uncalibrated images. In *Proceedings of the APR Workshop on Machine Vision Applications*, pages 559–568, Chiba, Japan.

Cipolla, R., Drummond, T., and Robertson, D. (1999). Camera calibration from vanishing points in images of architectural scenes. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 2, pages 382–391, Nottingham, UK.

Clarke, J. (1998). Modelling uncertainty: A primer. Technical Report 2161, University of Oxford, Dept. Engineering Science.

Clarke, T. A. and Fryer, J. G. (1998). The development of camera calibration methods and models. *Photogrammetric Record*, 16(91):51–66.

Coifman, B., Beymer, D., McLauchlan, P., and Malik, J. (1998). A real-time computer vision system for vehicle tracking and traffic surveillance. *Transportation research. Part C, Emerging technologies (Transp. res., Part C Emerg. technol.)*, 6(4):271–288.

Collins, R. (1993). *Model acquisition using stochastic projective geometry*. PhD thesis, University of Massatchusetts.

Collins, R. T., Lipton, A. J., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., Hasegawa, O., Burt, P., and Wixson, L. (2000). A system for video surveillance and monitoring. Technical Report CMU-Ri-TR-00-12, Carnegie Mellon University.

Collins, R. T. and Weiss, R. S. (1990). Vanishing point calculation as statistical inference on the unit sphere. In *Proceedings of the 3rd International conference on Computer Vision (ICCV)*, pages 400–403.

Comaniciu, D., Ramesh, V., and Meer, P. (2003). Kernel-based object tracking. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 25(5):564–577.

Cornelis, K., Pollefeys, M., and Van Gool, L. (2002). Lens distortion recovery for accurate sequential structure and motion recovery. In Heyden, A., editor, *Proceedings of the 7th European Conference on Computer Vision (ECCV)*, volume LNCS 2351, pages 186–200. Springer.

Coughlan, J. M. and Yuille, A. L. (1999). Manhattan world: Compass direction from a single image by bayesian inference. In *Proceedings of the International Conference on Computer Vision (ICCV)*, volume 2, pages 941–947, Corfu, Greece. IEEE.

Coughlan, J. M. and Yuille, A. L. (2000). The manhattan world assumption: Regularities in scene statistics which enable bayesian inference. In *Neural Information Processing Systems (NIPS '00)*, Denver, CO.

Coughlan, J. M. and Yuille, A. L. (2003). Manhattan world: Orientation and outlier detection by bayesian inference. *Neural computation*, 15(5):1063–1088.

Criminisi, A., Reid, I., and Zisserman, A. (2000). Single view metrology. *International Journal of Computer Vision*, 40(2):123–148.

Dalal, N., Triggs, B., and Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *Proceedings of the 9th European Conference on Computer Vision (ECCV)*.

Dellaert, F. (2002). The expectation maximization algorithm. Technical Report GIT-GVU-02-20, College of Computing, Georgia Institute of Technology.

Dellaert, F., Seitz, S. M., Thorpe, C. E., and Thrun, S. (2000). Structure from motion without correspondence. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 557–564.

Demazure, M. (1981). Sur deux problemes de reconstruction. Technical Report 882, INRIA, Rocquencourt, France.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

Deutscher, J., Isard, M., and MacCormick, J. (2002). Automatic camera calibration from a single manhattan image. In *Proceedings of the 7th European Conference on Computer Vision*, volume 2353, pages 175–188. IEEE.

Devernay, F. and Faugeras, O. (2001). Straight lines have to be straight. *Machine Vision and Applications*, 13(1):14–24.

Duda, R. and Hart, P. (1972). Use of the hough transform to detect lines and curves in pictures. *Comm ACM*, 15:11–15.

Duda, R., Hart, P., and Stork, D. (2001). *Pattern Classification*. John Wiley & Sons.

Engels, C., Stewenius, H., and Nister, D. (2006). Bundle adjustment rules. In Förstner, W. and Steffen, R., editors, *In Photogrammetric Computer Vision (PCV)*, pages 266–271, Bonn. ISPRS.

Faugeras, O. and Luong, Q.-T. (2001). *The Geometry of Multiple Images*. MIT Press.

Faugeras, O. and Maybank, S. (1990). Motion from point matches: Multiplicity of solutions. *International Journal of Computer Vision*, 4(3):225–246.

Faugeras, O. D., Luong, Q.-T., and Maybank, S. J. (1992). Camera self-calibration: Theory and experiments. In *Proceedings of the European Conference on Computer Vision*, pages 321–334.

Fischler, M. and Bolles, R. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. of the ACM*, 24:381–395.

Fisher, R. B. (2002). Self-organization of randomly placed sensors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 4, page 146 ff.

Forsyth, D. A. and Ponce, J. (2002). *Computer Vision: A modern approach*. Prentice Hall.

Freeman, J. (2007). The worldwide intelligent video and smart camera market. Technical report, J.P. Freeman.

Freyer, J. G. and Brown, D. C. (1986). Lens distortion for close-range photogrammetry. *Photogrammetric Engineering and Remote Sensing*, 52(1):51–58.

Funiak, S., Guestrin, C., Paskin, M., and Sukthankar, R. (2006). Distributed localization of networked cameras. In *Proceedings of the 5th international conference on Information processing in sensor networks (IPSN)*, pages 34–42, Nashville.

Gallagher, A. C. (2002). A groundtruth based vanishing point detection algorithm. *Pattern Recognition*, 35:1527–1543.

Golub, G. and van Loan, C. (1996). *Matrix Computations*. The Johns Hopkins University Press, Baltimore.

Grabner, H., Roth, P., and Bischof, H. (2007). Is pedestrian detection really a hard task? In *Proceedings of the 10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*.

Gracie, G. (1968). Analytical photogrammetry applied to single terrestrial photograph mensuration. In *In XIth International Conference of Photogrammetry*, Lausanne, Switzerland. The International Society for Photogrammetry and Remote Sensing.

Grammatikopoulos, L., Karras, G., and Petsa, E. (2003). Camera calibration approaches using single images of man-made objects. In *Proceedings of the XIX CIPA International Symposium*, pages 328–332, Antalya, Turkey.

Grammatikopoulos, L., Karras, G., and Petsa, E. (2007). An automatic approach for camera calibration from vanishing points. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62:64–76.

Gurdjos, P. and Payrissat, R. (2000). About conditions for recovering the metric structures of perpendicular planes from the single ground plane to image homography. In *Proceedings of the International Conference on Pattern Recognition*, volume 1, pages 1358–1361, Barcelona, Spain. IEEE Computer Scociety.

Guru, D. S., Shekar, B. H., and Nagabhushan, P. (2004). A simple and robust line detection algorithm based on small eigenvalue analysis. *Pattern Recognition Letters*, 25(1):1–13.

Hartley, R. (1995). In defence of the 8-point algorithm. In *Proceedings of the 5th International conference on Computer Vision (ICCV)*, pages 1064–1070, Cambridge, MA.

Hartley, R. and Kang, S. B. (2005). Parameter-free radial distortion correction with centre of distortion estimation. Technical Report MSR-TR-2005-42, Microsoft research.

Hartley, R. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press.

Henderson, D. W. and Taimina, D. (2005). *Experiencing Geometry - Euclidean and Non-Euclidean with History*. Pearson Prentice Hall.

Henrik Stewenius, C. E. and Nister, D. (2006). Recent developments on direct relative orientation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60(4):284–294.

Heuel, S. (2004). *Uncertain Projective Geometry*. Springer Verlag GmbH.

Heyden, A. and Aström, K. (1995). A canonical framework for sequences of images. In *IEEE Workshop on Representation of Visual Scenes*, Boston.

Horn, B. (1990). Relative orientation. *International Journal of Computer Vision (IJCV)*, 4:59–78.

Hough, P. V. (1962). Method and means for recognizing complex patterns. US Patent 3,069,654.

Javed, O., Rasheed, Z., Alatas, O., and Shah, M. (2003). M-knight: A real time surveillance system for multiple and non-overlapping cameras. In *IEEE International conference on Multimedia and Expo*, pages 649–652, Baltimore.

Javed, O., Shafique, K., Rasheed, Z., and Shah, M. (2008). Modeling inter-camera space-time and apprearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109:146–162.

Jaynes, C. (2004). Multi-view calibration from planar motion trajectories. *Image and Vision Computing*, 22(7):535–550.

Jin, Y. and Mokhtarian, F. (2007). Variational particle filter for multi-object tracking. In *Proceedings of the 11th International conference on Computer Vision (ICCV)*.

Kahn, S. and Shah, M. (2003). Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1355–1360.

Kanatani, K. (1996). *Statistical optimization for geometric computation*. Elsevier.

Kanazawa, Y. and Kanatani, K. (2001). Do we really have to consider covariance matrices for image features? In *Proceedings of the 8th International conference on Computer Vision (ICCV)*, pages 301–306.

Karras, G. and Petsa, E. (1999). Metric information from single uncalibrated images. In *Proceedings of the XVII CIPA International Symposium*, Olinda, Brasil.

Kaucic, R., Hartley, R., and Dano, N. (2001). Plane-based projective reconstruction. In *Proceedings of the International conference on Computer Vision (ICCV)*, volume 1, pages 420–427.

Kemp, M. (1992). *The science of art: optical themes in western art from Brunelleschi to Seurat*. Yale University Press.

Kosecka, J. and Zhang, W. (2002). Video compass. In *Proceedings of the 7th European Conference on Computer Vision (ECCV)*, volume 2353, page 476f. Springer-Verlag.

Krahnstoever, N. and Mendoca, P. (2005). Bayesian autocalibration for surveillance. In *Proceedings of the 5th International conference on Computer Vision (ICCV)*.

Krahnstoever, N. and Mendonca, P. (2006). Autocalibration from tracks of walking people. In *Proceedings of the British machine vision conference (BMVC)*.

Kruppa, E. (1913). Zur ermittlung eines objektes aus zwei perspektiven mit innerer orientierung. *Sitz.-Bericht Akademie der Wissenschaften*, 122:1939–1948.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics*, 2:83–97.

Lee, L., Romano, R., and Stein, G. (2000). Monitoring activities from multiple video streams: Establishing a common coordinate frame. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 22(8):758–767.

Li, H. and Hartley, R. (2005). A non-iterative method for correcting lens distortion from nine point correspondences. In *Proceedings of the 6th workshop on omnidirectional vision, camera networks and non-classical cameras*.

Li, H. and Hartley, R. (2007). The 3d-3d registration problem revisited. In *Proceedings of the 11th International conference on Computer Vision (ICCV)*, pages 1–8, Rio de Janeiro.

Liebowitz, D. (2001). *Camera Calibration and Reconstruction of Geometry from Images*. PhD thesis, University of Oxford, Dept. Engineering Science. D.Phil. thesis.

Liebowitz, D. and Zisserman, A. (1998). Metric rectification for perspective images of planes. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 482–488, Santa Barbara, CA, USA. IEEE.

Liebowitz, D. and Zisserman, A. (1999). Combining scene and auto-calibration constraints. In *Proceedings of the International Conference on Pattern Recognition (ICCV)*, volume 1, pages 293–300. IEEE.

Longuet-Higgins, H. C. (1981). A computer algorithm for reconstructing a scene from two projections. *Nature*, 293(10):133–135.

Lourakis, M. and Argyros, A. (2005). Is levenberg-marquardt the most efficient optimization algorithm for implementing bundle adjustment? In *Proceedings of the 10th International conference on Computer Vision (ICCV)*, pages 1526–1531.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110.

Lutton, E., Maitre, H., and Lopez-Krahe, J. (1994). Contribution to the determination of vanishing points using hough transform. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 4(16):430–438.

Lv, F., Zhao, T., and Nevatia, R. (1999). Self-calibration of a camera from video of a walking human. In *Proceedings of the International Conference on Computer Vision*. IEEE.

Ma, Y., Soatto, S., Kosecka, J., and Sastry, S. S. (2004). *An Invitation to 3-D Vision*. Springer.

Magee, M. J. and Aggarwal, J. K. (1984). Determining vanishing points from perspective images. *Journal of Computer Vision, Graphics and Image Processing*, 26(2):256–267.

Martinec, D. and Pajdla, T. (2002). Structure from many perspective images with occlusions. In A. Heyden, G. Sparr, M. N. and Johansen, P., editors, *Proceedings of the 4th European Conference on Computer Vision (ECCV)*, volume 2, pages 355–369, Berlin. Springer.

Matas, J., Chum, O., Urban, M., and Pajdla, T. (2002). Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British machine vision conference*, volume 1, pages 384–393, London. Stephens & George Print Group.

McFarlane, N. and Schofield, C. (1995). Segmentation and tracking of piglets. *Machine Vision and Applications*, 8(3):187–193.

McGlone, editor (2004). *Manual of Photogrammetry*. ASPRS.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley.

McLean, G. F. and Kotturi, D. (1995). Vanishing point detection by line clustering. *IEEE Transactions on pattern analysis and machine intelligence*, 17(11):1090–1095.

Minagawa, A., Tagawa, N., Moriya, T., and Gotoh, T. (1999). Line clustering with vanishing point and vanishing line. In *Proceedings of the International Conference on Image Analysis and Processing*, pages 388–393.

Nakatani, H. and Kitahashi, T. (1980). Extraction of vanishing point and its application to scene analysis based on image sequence. In *Proceedings of the 5th International Conference of Pattern Recognition (ICPR)*, pages 370–372.

Nister, D. (2004). An efficient solution to the five-point relative pose problem. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 26(6):756–777.

Nister, D., Kahl, F., and Stewenius, H. (2007). Structure from motion with missing data is np-hard. In *Proceedings of the 11th International conference on Computer Vision (ICCV)*, Rio de Janeiro.

Oscar Pizzaro, R. E. and Singh, H. (2003). Relative pose estimation for instrumented calibrated imaging platforms. In C. Sun, H. Talbot, S. O. and Adriaansen, T., editors, *Proceedings of the 7th Digital Image Computing: Techniques and Applications (DICTA)*.

Pavlidis, I., Morellas, V., Tsiamyrtzis, P., and Harp, S. (2001). Urban surveillance systems: From the laboratory to the commercial world. *Proceedings of the IEEE*, 89(10):1478–1497.

Pflugfelder, R. and Bischof, H. (2003). Vanishing points and lorries. In Burger, W. and Scharinger, J., editors, *Digital Imaging in Media and Education, Proceedings of the 28th AAPR Workshop*, volume 179, pages 205–212. AAPR/ÖAGM - Österreichische Arbeitsgemeinschaft Mustererkennung, Österreichische Computer Gesellschaft.

Pflugfelder, R. and Bischof, H. (2005). On-line auto-calibration in man-made worlds. In *Proceedings of the Digital Image Computing: Techniques and Applications (DICTA)*, Cairns. APRS, IEEE.

Pflugfelder, R. and Bischof, H. (2006a). Computation of the epipolar geometry in slightly overlapping views. In *Proceedings of the 11th Computer Vision Winter Workshop (CVWW)*. Czech Pattern Recognition Society Group, Center for Machine Perception at CTU Prague and Czech Society for Cybernetics and Informatics.

Pflugfelder, R. and Bischof, H. (2006b). Fundamental matrix and slightly overlapping views. In *Proceedings of the 18th International Conference of Pattern Recognition (ICPR)*, pages 527–530. IEEE.

Pflugfelder, R. and Bischof, H. (2007a). Kameramatrix. Austrian Patent AT 502.356.

Pflugfelder, R. and Bischof, H. (2007b). People tracking across two distant self-calibrated cameras. In *Proceedings of the IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*. IEEE, IEEE.

Pflugfelder, R., Bischof, H., Fernandez, G., Nölle, M., and Schwabach, H. (2005). Influence of camera properties on image analysis in visual tunnel surveillance. In *Proceedings of the 8th international conference on Intelligent Transportation Systems (ITSC)*. ITSS, IEEE.

Philip, J. (1996). A non-iterative algorithm for determining all essential matrices corresponding to five point pairs. *Photogrammetric Record*, 15(88):589–599.

Philip, J. (1998). Critical point configurations of the 5-, 6-, 7-, and 8-point algorithms for relative orientation. *TRITA-MAT-1998-MA-13*.

Piao, Y. and Sato, J. (2007). Computing epipolar geometry from unsychronized cameras. In *Proceedings of the 4th International Conference on Image Analysis and Processing (ICIAP)*.

Quan, L. and Mohr, R. (1989). Determining perspective structures using hierarchical hough transform. *Pattern Recognition Letters*, 9:279–286.

Rahimi, A., Dunagan, B., and Darrell, T. (2004). Simultaneous calibration and tracking of non-overlapping sensors. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 187–194.

Roth, P., Grabner, H., Skocaj, D., Bischof, H., and Leonardis, A. (2005). On-line conservative learning for person detection. In *Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*.

Rother, C. (2002a). A new approach for vanishing point detection in architectural environments. In *Proceedings of the British Machine Vision Conference*, volume 20, pages 647–656.

Rother, C. (2002b). A new approach to vanishing point detection in architectural environments. *Image and vision computing*, 20:647–655.

Rother, C. (2003). *Multi-View Reconstruction and Camera Recovery using a Real and Virtual Reference Plane*. PhD thesis, Royal Institute of Technology.

Rudoy, M. and Rohrs, C. E. (2006). Simultaneous sensor calibration and path estimation. In *In Proc. IEEE Asilomar Conference on Signals, Systems, and Computers*.

Sanden, H. V. (1908). *Die Bestimmung der Kernpunkte in der Photogrammetrie*. PhD thesis, Universität Göttingen.

Saxena, A., Chung, S., and Ng, A. (2008). 3-d depth reconstruction from a single still image. *International Journal of Computer Vision (IJCV)*, 76:53–69.

Schaffalitzky, F. and Zisserman, A. (2000). Planar grouping for automatic detection pf vanishing lines and points. *Image and vision computing*, 18:647–658.

Schindler, G. and Dellaert, F. (2004). Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex manmade environments. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 203–209.

Sekita, I. (1994). On fitting several lines using the em algorithm. In *Proceedings on CVVC*, pages 107–109.

Semple, J. G. and Kneebone, G. T. (1998). *Algebraic Projective Geometry*. Oxford Press.

Seo, K. S., Lee, J. H., and Choi, H. M. (2006). An efficient detection of vanishing points using inverted coordinates image space. *Pattern Recognition Letters*, 27:102–108.

Shah, M., Javed, O., and Shafique, K. (2007). Automated visual surveillance in realistic scenarios. *IEEE Multimedia*, 14(1):30–39.

Shashua, A. and Navab, N. (1994). Relative affine structure: Theory and applications to 3d reconstruction from perspective views. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 483–489, Seattle.

Sheikh, Y., Li, X., and Shah, M. (2007). Trajectory association across non-overlapping moving cameras in planar scenes. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7.

Shufelt, J. A. (1999). Performance evaluation and analysis of vanishing point detection techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(3):282–288.

Siebel, N. T. and Maybank, S. J. (2004). The advisor visual surveillance system. In *In Proceedings of the ECCV Workshop on Applications of Computer Vision*, Prague.

Sinha, S. N., Pollefeys, M., and McMillan, L. (2004). Camera network calibration from dynamic silhouettes. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington DC.

Stauffer, C. and Tieu, K. (2003). Automated multi-camera planar tracking correspondence modeling. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 259–266.

Stein, G. (1999). Tracking from multiple view points: Self-calibration of space and time. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1521–.1527.

Strang, G. (2003). *Introduction to Linear Algebra*. Wellesley-Cambridge Press.

Sturm, P. and Maybank, S. J. (1999). A method for interactive 3d reconstruction of piecewise planar objects from single images. In *Proceedings of the British machine vision conference (BMVC)*, pages 265–274.

Sturm, P. and Triggs, B. (1996). A factorization based algorithm for multi-image projective structure and motion. In *Proceedings of the 4th European Conference on Computer Vision (ECCV)*, pages 709–720.

Sturm, R. (1869). Das problem der projektivität und seine anwendung auf die flächen zweiten grades. *Mathematische Annalen*, 1:533–573.

Svoboda, T., Martinec, D., and Pajdla, T. (2005). A convinient multi-camera self-calibration for virtual environments. *PRESENCE: Teleoperators and Virtual Environments*, 14(4):407–422.

Taylor, C., Rahimi, A., Bachrach, J., Shrobe, H., and Grue, A. (2006). Simultaneous localization, calibration, and tracking in an ad hoc sensor network. In *Proceedings of the 5th international conference on Information processing in sensor networks (IPSN)*, pages 27–33, Nashville.

Thirde, D., Borg, M., Valentin, V., Fusier, F., Aguilera, J., Ferryman, J., Bremond, F., Thonnat, M., and Kampel, M. (2005). Visual surveillance for aircraft activity monitoring. In *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*.

Thirde, D., Li, L., and Ferryman, J. (2006). An overview of the pets 2006 dataset. In *Proceedings of the 9th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, pages 47–50, New York.

Thrun, S., Burghard, W., and Fox, D. (2005). *Probabilistic Robotics*. MIT Press.

Tomasi, C. and Kanade, L. (1992). Shape and motion from image streams under orthography: A factorization approach. *International Journal of Computer Vision (IJCV)*, 9(2):137–154.

Toyama, K., Krumm, J., Brumitt, B., and Meyers, B. (1999). Wallflower: Principles and practice of background maintenance. In *Proceedings of the 8th International conference on Computer Vision (ICCV)*, pages 255–261.

Triggs, B. (2000a). Plane + parallax, tensors and factorization. In *Proceedings of the 4th European Conference on Computer Vision (ECCV)*, pages 522–538, Dublin.

Triggs, B. (2000b). Routines for relative pose of two calibrated cameras from five points. Technical report, INRIA.

Triggs, B., McLauchlan, P., Hartley, R., and Fitzgibbon, A. (2000). *Vision Algorithms: Theory and Practice*, volume 1883 of *LNCS*, chapter Bundle Adjustment: A modern synthesis, pages 298–375. Springer.

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 Journal of Robotics and Automation*, RA-3(4):323–344.

Tuytelaars (1998). The cascaded hough transform as an aid. In *Proceedings of the International conference on Computer Vision (ICCV)*.

Van den Heuvel, F. A. (1998). Vanishing point detection for architectural photogrammetry. *International archives of photogrammetry and remote sensing*, 32(5):652–659.

Van den Heuvel, F. A. (1999). Estimation of interior orientation parameters from constraints on line measurements in a single image. In *International archives of photogrammetry and remote sensing*, volume 32, pages 81–88.

Viola, P. and Jones, M. (2004). Robust real-time object detection. *International Journal of Computer Vision (IJCV)*, 57(2):137–154.

Yilmaz, A., Javed, O., and Shah, M. (2006). Object tracking: A survey. *ACM Computing Surveys (CSUR)*, 38(4):13.

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334.

Zhang, Z., Luong, Q.-T., and Faugeras, O. (1996). Motion of an uncalibrated stereo rig: self-calibration and metric reconstruction. *IEEE Transactions on Robotics and Automation*, 12(1):103–113.

Zhao, T. and Nevatia, R. (2004). Tracking multiple humans in complex situations. *Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1208–1221.