Continuous Stereo Self-Calibration by Camera Parameter Tracking

Thao Dang, *Member, IEEE*, Christian Hoffmann, and Christoph Stiller, *Senior Member, IEEE*

***Abstract*—This paper presents a consistent framework for continuous stereo self-calibration. Based on a practical analysis of the sensitivity of stereo reconstruction to camera calibration uncertainties, we identify important parameters for self-calibration. We evaluate different geometric constraints for estimation and tracking of these parameters: bundle adjustment with reduced structure representation relating corresponding points in image sequences, the epipolar constraint between stereo image pairs, and trilinear constraints between image triplets. Continuous, recursive calibration refinement is obtained with a robust, adapted iterated extended Kalman filter. To achieve high accuracy, physically relevant geometric optimization criteria are formulated in a Gauss–Helmert type model. The self-calibration framework is tested on an active stereo system. Experiments with synthetic data as well as on natural indoor and outdoor imagery indicate that the different constraints are complementing each other and thus a method combining two of the above constraints is proposed: While reduced order bundle adjustment gives by far the most accurate results (and might suffice on its own in some environments), the epipolar constraint yields instantaneous calibration that is not affected by independently moving objects in the scene. Hence, it expedites and stabilizes the calibration process.**

***Index Terms*—Active vision, self-calibration, stereo vision.**

I. Introduction

STEREO vision is of growing importance for many applications ranging from automotive driver assistance systems over autonomous robot navigation to 3-D metrology. Probably the most prominent advantage of stereopsis is its ability to provide both instantaneous 3-D measurements and rich texture information that is crucial to many object classification tasks. Camera calibration is indispensable to relate image features acquired with a stereo rig to real world coordinates. The calibration process is typically required to recover camera orientation with an accuracy of say 10^{-3} to 10^{-2} degrees. Traditionally, camera calibration is determined off-line by observing special, well-known reference patterns (see, e.g., [1]). Recent years, however, have seen increasing interest in camera self-calibration methods. Stereo self-calibration refers to the automatic determination of extrinsic and intrinsic camera parameters of a stereo rig from almost arbitrary image sequences. Hence, such methods allow recovery of the camera parameters while the sensor is in use without necessitating any special calibration object.

We believe that self-calibration is an important ability required for the introduction of stereo cameras in the market, especially in the automotive field. Only self-calibration can guarantee maintenance-free long-term operation, since camera parameters may be subject to drift due to adverse environmental conditions such as, e.g., mechanical vibrations or large temperature variations that are commonly encountered in automotive applications. Additionally, reliable self-calibration may render costly initial off-line calibration obsolete, thus reducing time and cost in the production line.

Another important field of application is active vision [2]–[5]. Inspired by the human visual system, an active vision system usually consists of two or more cameras that can adjust gaze direction to currently important areas of the scene. Clearly, such a system can be useful in many applications such as extending the field of view for autonomous vehicles at street crossings or smooth following of objects. We have developed an active camera platform consisting of three cameras as depicted in Fig. 1. The camera system is intended to implement the aforementioned active vision capabilities for autonomous driving. Additionally, since we do not rotate the baseline of the stereo cameras but all cameras individually to reduce packaging size, the platform is an excellent test bed for stereo self-calibration.

The main objective of this paper is to describe a framework for *continuous stereo self-calibration*. Our approach differs from many self-calibration tasks since we assume that an initial guess of the camera calibration is readily available (e.g., camera orientation is given with errors up to a few degrees), and our self-calibration has to refine these initial guesses. A first refinement will be available after processing of a single stereo image pair which is further improved upon availability of more images. Calibration parameters are not constrained to be constant during the continuous calibration process, but may drift over time. Furthermore, we address calibration from almost arbitrary imagery, i.e., we do not impose imagery taken in completely rigid scenes. We thus on one hand address only part of the full self-calibration problem, as we expect a coarse initial guess. However, we believe that this simplification is valid for a variety of applications, e.g., when the parameters of the stereo rig are given within the tolerances of the manufacturing process or in active vision when (perturbed) commanded camera parameters are available. On the other hand the proposed process offers instantaneous and continuous calibration with relaxed constraints on parameter constancy and scene dynamics.

Manuscript received April 29, 2008; revised December 08, 2008. First published June 02, 2009; current version published June 12, 2009. This work was supported by the German Research Foundation (Deutsche Forschungsgesellschaft, DFG) in the Transregional Collaborative Research Centre 28 (TCRC28) "Cognitive Automobiles". The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Amy R. Reibman.

The authors are with the Institut für Mess- und Regelungstechnik, University of Karlsruhe, Germany (e-mail: dang@mrt.uka.de; hoffmann@mrt.uka.de; stiller@mrt.uka.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2009.2017824

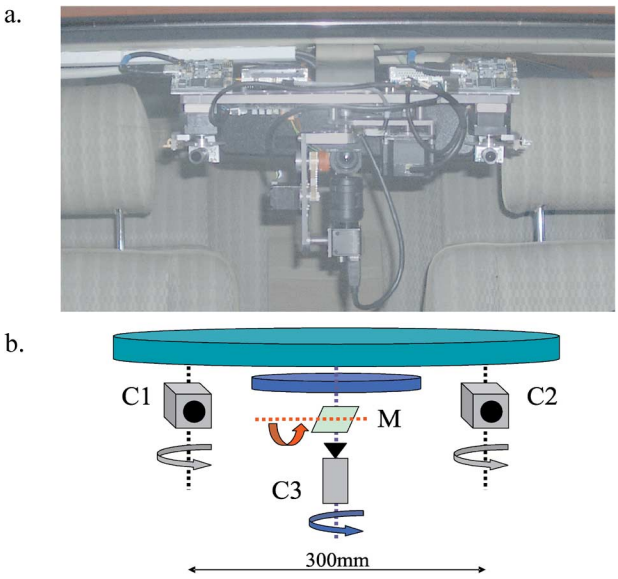

Fig. 1. (a) Active camera platform mounted in our experimental vehicle. The platform consists of two (stereo) cameras with a field of view of 46° and one tele camera. The two stereo cameras can be rotated independently about their yaw axes while the tele camera is capable of both horizontal and vertical rotations to compensate vehicle pitch. (b) Schematic view of the active camera platform (C1, C2: stereo cameras, C3: tele camera, M: mirror).

In our framework, we evaluate different constraints that may be employed when tracking stereo calibration parameters: recursive bundle adjustment with reduced dimension of the parameter state vector, the epipolar constraint between a pair of stereo images, and the trifocal constraint. A Gauss–Helmert type model is employed to ensure that physically relevant errors are minimized. We find that recursive bundle adjustment and the epipolar constraint may complement each other in practical applications: Bundle adjustment provides highest accuracy and might suffice on its own in some environments. Methods employing the epipolar constraint are fast and are not affected by independently moving objects in the scene and may thus expedite and stabilize the calibration process. We propose an algorithm that combines both the epipolar constraint for instantaneous measurements and recursive bundle adjustment that cumulates information from spatio-temporal trajectories of correspondences over time. The algorithms are demonstrated on both synthetic and real imagery.

The paper is organized as follows. Section II briefly categorizes existing literature on camera self-calibration and puts our approach in context with previous work. Section III outlines the stereo camera model used throughout this paper and analyzes the sensitivity of stereo reconstruction to the individual camera parameters. The framework for continuous self-calibration based on different geometric constraints is described in Section IV. Our algorithm is evaluated on synthetic and real-life imagery (Section V). Section VI summarizes our results and concludes the paper.

II. Related Work

Probably the earliest work related to camera self-calibration stems from photogrammetry, addressing the extraction of metric quantities from multiple (often aerial) images. In case of uncertain camera parameters, 3-D reconstruction is usually obtained via *bundle adjustment* (e.g., [6]–[9]): the intrinsic and extrinsic camera parameters as well as the observed 3-D structure are simultaneously refined such that the distance between measured and expected image coordinates becomes minimal. Given accurate correspondence features and a suitable initial guess of all camera parameters, bundle adjustment is known to provide calibration results with high precision. Bundle adjustment always processes a block of images, i.e., a window in a sequence of images. The size of this window is an important design parameter usually limited by the computational resources available. Whenever new images are added to this window, a re-computation of the entire bundle is required which is undesirable in continuous processing.

In the computer vision community, self-calibration of monocular and stereo cameras has been in the focus over the last 15 years and is still an active research topic. While in photogrammetry applications an initial guess of the camera intrinsic and extrinsic camera calibration is usually available from camera data sheets and mechanical specifications, most computer vision research does not assume any *a priori* information about the camera calibration. Self-calibration is commonly formulated as a multistage procedure: First, a projective representation of the camera system is determined in form of fundamental matrices [10] or trifocal tensors (e.g., [11]). Using this representation, it is possible to calculate a projective reconstruction of the scene. Several possibilities have been described to update this projective reconstruction to a metric one: [12]–[14] used the *Kruppa equations* to compute intrinsic camera parameters from the epipolar geometry of two views. However, the Kruppa equations are nonlinear and known to be degenerate in some cases [15]. Another direct method to obtain a metric calibration involves the absolute quadric over several views [16]. Alternatively, a stratified approach has been presented, that computes affine structure from the projective model and subsequently solves the self-calibration problem by upgrading the affine structure to a metric one [17]. This method was extended for the self-calibration of a stereo rig in [18] and [19]; [20] investigated critical motion sequences for autocalibration from multiple stereo image pairs.

The aforementioned multistage methods may yield camera calibration without any prior knowledge of the cameras (except for some minor restrictions on the intrinsic parameters as, e.g., zero skew). However, the attained accuracy is not comparable to classic bundle adjustment. This is mostly due to the utilization of algebraic optimization criteria for autocalibration. Algebraic error criteria neglect available knowledge of observation noise characteristics. In the presence of noisy input data, they provide inferior result as compared to geometric approaches minimizing physically relevant error measures. Thus, bundle adjustment is often used in a nonlinear post processing stage to refine the camera parameters. Other nonlinear methods for refining an initial guess of calibration binocular cameras are proposed in [21] and [22]. They employ a Euclidean parametrization of the fundamental matrices between pairs of images in the stereo sequence and use nonlinear optimization to find the camera parameters that satisfy the *spatial* epipolar constraints between

left and right stereo images as well as *temporal* epipolar constraints between pairs of consecutive images. An important aspect of the work in [21] and [22] is that the nonlinear optimization can be implemented as a recursive approach via an extended Kalman filter. Since such an algorithm can—in theory—run infinitely and continuously integrate new information as soon as it becomes available, we will refer to this class of approaches as *continuous self-calibration*. Apart from Kalman filter based methods for continuous self-calibration, other probabilistic algorithms such as sequential importance sampling have been investigated in [23] and [24] for monocular cameras; [24] demonstrates how to consider critical motion sequences in an algorithm that recovers the focal length of a moving camera.

A natural extension to the usage of the epipolar constraint for continuous self-calibration is to exploit the geometry of three images as captured in the trilinearities (formulated in [25]). The main advantage of the trilinear constraint over a set of epipolar constraints between three images lies in the deficiency of the epipolar transfer (e.g., [26]): given the fundamental matrices between three images and two corresponding points $\mathbf{x}_1$ and $\mathbf{x}_2$ in the first two views, it is not possible to compute the position $\mathbf{x}_3$ if the corresponding object point lies on or near the plane defined by the optical centers of the cameras. The trifocal transfer does not suffer from this degeneracy. For a monocular camera, the trilinear constraint has been used to determine Euclidean camera motion as an initialization step for subsequent self-calibration in [27]. We are not aware of any continuous binocular self-calibration that relies on a Euclidean parametrization of the trilinearities.

Continuous self-calibration is especially important in active vision when the camera calibration is constantly changing and online self-calibration is thus indispensable: [28] employs a variable state dimension filter (VSDF), an efficient implementation of an EKF, for the self-calibration of a monocular active camera from purely rotational motions. A limitation of the VSDF is that it does not allow any dynamics of system or structure parameters. [4] proposed a stereo self-calibration for an active stereo rig that updates the yaw angles of both cameras. The algorithm relies solely on spatial point correspondences between the left and right camera images. A modified version of this approach has lately been implemented in real time using FPGAs [29]. However, these algorithms depend on the assumption that the rotation axes of the two cameras are precisely aligned and can only recover two calibration parameters, namely the yaw angles of both cameras.

Our contribution belongs to the category of continuous self-calibration algorithms. It extends an earlier conference paper [30] and provides a consistent framework for recursive autocalibration that compares and combines different constraints in Euclidean parametrization: Apart from the epipolar constraint that was already utilized in [4] and [21], we also evaluate the trilinear constraint and bundle adjustment with reduced dimension of the parameter state. To the best of our knowledge the latter two constraints have not yet been used in a framework for continuous stereo self-calibration as presented here. Our work is inspired by structure from motion algorithms for monocular cameras [31]–[33], that employ 2-D displacements in an image sequence to retrieve the 3-D structure of a rigid scene.

III. Camera Model

A. Mathematical Formulation

Fig. 1 depicts our active camera platform consisting of three independently moving cameras: two stereo camera with one rotational degree of freedom (DOF), respectively, and a tele camera with two DOF. Since the platform will serve as an experimental setup for stereo self-calibration in this article, we will refrain from discussing the tele camera. For convenience, an extrinsic camera model will be employed that is specifically tailored to active stereo cameras. However, this is not a general limitation as any other representation of the extrinsic camera parameters may also be used in our framework.

In this paper, the ideal pinhole model is employed to describe the stereo cameras. This model relates a 3-D point $\mathbf{X} = (X, Y, Z)^T$ and its 2-D image coordinates $\mathbf{x} = (x, y)^T$ by the following projective equation (see, e.g., [34]):

$$\begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} = \lambda \mathbf{P} \begin{pmatrix} \mathbf{X} \\ 1 \end{pmatrix} \tag{1}$$

where λ is an unknown scalar factor. The projection matrix $\mathbf{P}$ of the ideal pinhole camera is defined as follows:

$$\mathbf{P} = \mathbf{K}[\mathbf{R}, -\mathbf{R}\mathbf{C}] \text{ with } \mathbf{K} = \begin{bmatrix} f_x & \alpha & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}. \tag{2}$$

The matrix $\mathbf{K}$ comprises the intrinsic camera parameters, i.e., the focal lengths $[f_x, f_y]^T$, the image center $[c_x, c_y]^T$, and the image skew coefficient α. In the scope of this work, we will assume that the pixels are square, i.e., $f_x = f_y = f$ and $\alpha = 0$, which is true for almost any modern digital camera. The rotation matrix $\mathbf{R}$ and the vector $\mathbf{C}$ specify the extrinsic camera parameters: $\mathbf{R}$ describes the orientation of the camera coordinate system with respect to the world reference frame and $\mathbf{C}$ denotes the position of the camera center in world coordinates.

For simplicity, we abbreviate perspective projection (1) of a 3-D point $\mathbf{X}$ onto its image $\mathbf{x}$ by $\mathbf{x} = \boldsymbol{\pi}(\mathbf{X})$. The inversion of the pinhole projection will be denoted as $\mathbf{X} = \boldsymbol{\Pi}^{-1}(\mathbf{x}, Z)$. Please note that depth component Z of $\mathbf{X}$ is required for a unique reprojection of $\mathbf{x}$.

Fig. 2 depicts the extrinsic parameters of our active camera platform. In many stereo applications, the world coordinate system (WCS) of the stereo rig is chosen to coincide with one of the two camera coordinate systems. In active vision, however, this is not an adequate choice since both cameras may change their orientation with respect to the moving platform. We thus define a WCS whose origin lies in the center of the baseline and whose X-axis is aligned with the baseline. To eliminate the remaining DOF, we impose that the Z-axis of the WCS lies within the plane spanned by the baseline and the optical axes of the right camera. Thus, the pitch angle of the right camera must be equal to zero.

Given the WCS as presented above, it is convenient to represent camera orientations as a concatenation of yaw-pitch-roll rotations. A vector $\boldsymbol{\omega} = [\Psi, \Phi, \Theta]^T$ of yaw, pitch and roll angles is transformed into a rotation matrix as follows:

$$\mathbf{R}(\boldsymbol{\omega}) = \mathbf{R}(\Psi, \Phi, \Theta) = \mathbf{R}_Z(\Theta)\mathbf{R}_X(\Phi)\mathbf{R}_Y(\Psi) \tag{3}$$

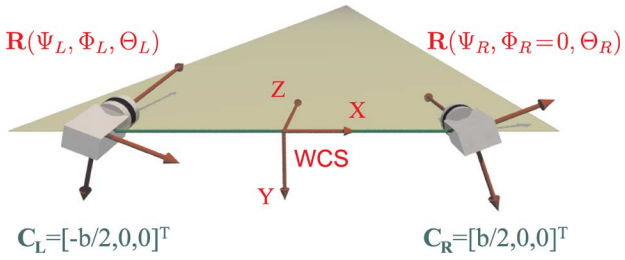

Fig. 2. Extrinsic parameters of active stereo rig. The world coordinate system (WCS) is located in the middle of the baseline of the stereo cameras. Additionally, we impose that the Z-axis of the WCS is parallel to the plane defined by the baseline and the optical axes of the right camera. b denotes the base length of the stereo rig and the rotation matrices are specified in yaw, pitch and roll angles.

where $\mathbf{R}_Z, \mathbf{R}_X, \mathbf{R}_Y$ are rotations about the Z-, X-, and Y-axis, respectively.

Referring to variables of the left or right camera with subscripts L and R, respectively, the projection matrices $\mathbf{P}_L$ and $\mathbf{P}_R$ of the two cameras are defined through Fig. 2

$$\mathbf{P}_R = \mathbf{K}_R \mathbf{R}_R [\mathbf{I}, -\mathbf{C}_R] \tag{4}$$
$$\mathbf{P}_L = \mathbf{K}_L \mathbf{R}_L [\mathbf{I}, -\mathbf{C}_L] \tag{5}$$

with

$$\mathbf{R}_R = \mathbf{R}(\Psi_R, 0, \Theta_R), \quad \mathbf{C}_R = [b/2, 0, 0]^T \tag{6}$$
$$\mathbf{R}_L = \mathbf{R}(\Psi_L, \Phi_L, \Theta_L), \quad \mathbf{C}_L = [-b/2, 0, 0]^T. \tag{7}$$

The stereo camera system is thus fully described by six intrinsic parameters $(f_L, \mathbf{c}_L, f_R, \mathbf{c}_R)$ and six extrinsic parameters: Ψ_L, Φ_L, Θ_L denote the orientation of the left camera with respect to the WCS, Ψ_R, Θ_R specify the yaw and roll angles of the right camera (the pitch angle is omitted since the WCS is aligned with the optical axis of the right camera), and b is the base length of the stereo rig.

B. Sensitivity of 3-D Reconstruction to Erroneous Camera Calibration

The objective of this section is to illustrate the effect of erroneous camera calibration on stereo reconstruction and to analyze the importance of the individual camera parameters for self-calibration. This is not a trivial undertaking since a thorough sensitivity analysis of 3-D reconstruction must not only consider stereo triangulation but also the consequences for stereo correspondence analysis. In practical applications, the search for corresponding points in both stereo images is usually restricted to a 1-D search space: Given a point $\mathbf{x}_R$ in the right image, we can compute the so-called *epipolar line* in the left image which contains all points that may correspond to $\mathbf{x}_R$. The computation of this epipolar line depends on the intrinsic and extrinsic parameters of the stereo rig (cf. Section IV-B). Thus, if camera calibration is inaccurate, the computed 1-D search may not contain the corresponding point and stereo matching may fail completely. Even for a search that is extended beyond the epipolar line, a pair of "corresponding" points will most likely yield nonintersecting associated light rays, necessitating a careful triangulation procedure (e.g., [35]).

A precise, stochastic analysis on how the accuracy of camera calibration affects stereo reconstruction is presented in [36]. In this work, each given point pair is first corrected to satisfy the epipolar constraint and then used for stereo triangulation. The camera parameter uncertainties are propagated through each of these processing stages to assess the influence of calibration errors. However, the obtained results are rather complex and do not allow a straight-forward, illustrative sensitivity analysis. Since our objective is to get more insight into the effect of errors in the individual camera parameters and into the restrictions that apply to enable matching along epipolar lines, we will follow a different approach. Other work on the sensitivity of stereopsis to calibration inaccuracies is presented in [37] and [38], however, with very strict constraints, such as purely vertical misalignments or exactly coplanar optical axes. Recently, an experimental framework was presented in [39] that determines the limits on relative camera alignment errors to allow stereo matching and reconstruction.

For the sensitivity analysis, we assume a calibrated and rectified stereo rig. Rectified or ideal stereo images can be thought of as acquired by two cameras which have coplanar retinas and optical axes that are perpendicular to the baseline. Image coordinate systems in a rectified stereo system may further be chosen such that the epipolar lines are parallel to the x-axes, i.e., corresponding pixels $\tilde{\mathbf{x}}_L$ and $\tilde{\mathbf{x}}_R$ in normalized image coordinates

$$\begin{pmatrix} \tilde{\mathbf{x}}_L \\ 1 \end{pmatrix} = \mathbf{K}_L^{-1} \begin{pmatrix} \mathbf{x}_L \\ 1 \end{pmatrix}; \quad \begin{pmatrix} \tilde{\mathbf{x}}_R \\ 1 \end{pmatrix} = \mathbf{K}_R^{-1} \begin{pmatrix} \mathbf{x}_R \\ 1 \end{pmatrix} \tag{8}$$

fulfill

$$\tilde{\mathbf{x}}_L - \tilde{\mathbf{x}}_R = \begin{pmatrix} \tilde{d} \\ 0 \end{pmatrix}. \tag{9}$$

The (normalized) disparity $\tilde{d}$ of (9) is inversely proportional to the depth Z of an observed object point (e.g., [40])

$$Z = b/\tilde{d}. \tag{10}$$

In case of rectified stereo images, the projection matrices simplify to $\mathbf{P}_R = [\mathbf{I}, -\mathbf{C}_R]$ and $\mathbf{P}_L = [\mathbf{I}, -\mathbf{C}_L]$. Using (1) and (10), stereo reconstruction is easily obtained by

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \frac{b}{\tilde{d}} \begin{pmatrix} \tilde{x}_R \\ \tilde{y}_R \\ 1 \end{pmatrix} + \begin{pmatrix} b/2 \\ 0 \\ 0 \end{pmatrix} = \frac{b}{\tilde{d}} \begin{pmatrix} \tilde{x}_L \\ \tilde{y}_L \\ 1 \end{pmatrix} - \begin{pmatrix} b/2 \\ 0 \\ 0 \end{pmatrix}. \tag{11}$$

Erroneous camera parameters affect the quality of both the rectification and the 3-D reconstruction. More precisely, the image coordinates $\hat{\mathbf{x}}_L, \hat{\mathbf{x}}_R$ obtained with erroneous calibration will differ from the coordinates $\mathbf{x}_L, \mathbf{x}_R$ that had been produced with ideal calibration:

$$\hat{\mathbf{x}}_L = \mathbf{x}_L + \Delta\mathbf{x}_L, \quad \hat{\mathbf{x}}_R = \mathbf{x}_R + \Delta\mathbf{x}_R \tag{12}$$

where $\Delta\mathbf{x}_L$ and $\Delta\mathbf{x}_R$ depend on the true image positions and the erroneous calibration parameters. The vertical and horizontal components of $\Delta\mathbf{x}_{L/R}$ affect stereo vision in different ways.

1) *Vertical misalignment* Δy: Due to the calibration errors, corresponding pixels may no longer lie in the same scan line and most stereo matching algorithms may deteriorate

since the search space may not contain corresponding regions. Thus, in practical applications no correspondence might be found at all, and, hence, the vertical misalignment $\Delta y = \hat{y}_L - \hat{y}_R$ should be small.

2) *Disparity error* Δd: The horizontal position errors induce a perturbation of the stereo disparity $\tilde{\hat{d}} = \tilde{\hat{x}}_L - \tilde{\hat{x}}_R = \tilde{d} + \Delta\tilde{d}$ in normalized image coordinates and $\Delta d_L = \hat{d}_L - d_L$, $\Delta d_R = \hat{d}_R - d_R$ in coordinates of the left and right camera, respectively. Taking the derivative of (10) with respect to $\tilde{d}$, this disparity error results in a deviation ΔZ from the true range Z as follows:

$$\Delta Z = \frac{\partial Z}{\partial \tilde{d}} \Delta\tilde{d} = -\frac{Z^2}{b} \Delta\tilde{d}. \tag{13}$$

Similar expressions hold for the unnormalized disparity errors Δd_L and Δd_R. Thus, for given disparity errors Δd or $\Delta\tilde{d}$ in pixels or normalized coordinates, respectively, the range uncertainty increases quadratically with distance.

For the further sensitivity analysis, we assume small perturbations in the calibration parameters and restrict our consideration to a first order approximation. Hence, for the effect of calibration errors to image coordinates we are left with

$$\Delta\tilde{\mathbf{x}} = \left.\frac{\partial \tilde{\mathbf{x}}}{\partial \mathbf{p}}\right|_{\mathbf{p}=\mathbf{p}_0} \Delta\mathbf{p} \quad \text{and} \quad \Delta\mathbf{x} = \left.\frac{\partial \mathbf{x}}{\partial \mathbf{p}}\right|_{\mathbf{p}=\mathbf{p}_0} \Delta\mathbf{p} \tag{14}$$

in normalized coordinates and in pixels, respectively. The vector $\mathbf{p}$ comprises all camera parameters as defined in Section III-A. The operation point $\mathbf{p}_0$ denotes the nominal camera parameters. For the compact results presented in this section, $\mathbf{p}_0$ is given by an stereo rig with parallel image planes. Given admissible bounds on deviations in image coordinates due to calibration inaccuracy, one can thus impose requirements on the accuracy of the individual calibration parameters.

For example, let us first analyze the effect of a yaw error $\Delta\Psi$ in the orientation of the left camera. According to (5), the normalized, erroneous image coordinates in the left stereo frame can be computed as

$$\begin{pmatrix} \tilde{\hat{x}}_L \\ \tilde{\hat{y}}_L \\ 1 \end{pmatrix} = \lambda \Delta\Omega \left[\mathbf{I} \middle| \begin{pmatrix} b/2 \\ 0 \\ 0 \end{pmatrix} \right] \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{15}$$

where

$$\Delta\Omega_L = \begin{bmatrix} \cos\Delta\Psi_L & 0 & \sin\Delta\Psi_L \\ 0 & 1 & 0 \\ -\sin\Delta\Psi_L & 0 & \cos\Delta\Psi_L \end{bmatrix} \tag{16}$$

is the rotation matrix due to the deviation $\Delta\Psi_L$ in the calibration of the left camera about the y-axis. The dependency on 3-D coordinates may be eliminated using (11) and subsequent solving for the scale factor λ finally yields

$$\Delta\tilde{\mathbf{x}}_L = -\begin{pmatrix} 1 + \tilde{x}_L^2 \\ \tilde{x}_L \tilde{y}_L \end{pmatrix} \Delta\Psi_L. \tag{17}$$

Thus, we find that in vicinity of the principal point ($\tilde{x}_L^2 \approx 0$, $\tilde{x}_L \tilde{y}_L \approx 0$) a small error in the yaw rotation yields a negligible vertical misalignment but may result in a significant disparity error $\Delta\tilde{d} = \Delta\Psi_L$ (in normalized coordinates). For cameras with large viewing angles it is worthwhile to consider the increase of the disparity error with horizontal viewing angle, and the increase of vertical misalignment towards the corners of the image.

Using (13), the range uncertainty may be calculated as

$$\Delta Z = \frac{Z^2}{b} \left(1 + \tilde{x}_L^2\right) \Delta\Psi_L \tag{18}$$

i.e., the uncertainty in depth increases slowly and to both sides with horizontal viewing angle, linearly with the yaw error, and quadratically with the distance. It is easy to show that a yaw error in the right image leads to exactly the same results.

Equation (14) may also be used to assess the impact of inaccuracies in the intrinsic camera parameters on image coordinates and on 3-D reconstruction. Consider for example a perturbation $\Delta\mathbf{c}_L = [\Delta c_{Lx}, \Delta c_{Ly}]^T$ in the image center coordinates of the left camera. The resulting image coordinates are given by

$$\begin{pmatrix} \hat{x}_L \\ \hat{y}_L \\ 1 \end{pmatrix} = \lambda \begin{bmatrix} f & 0 & c_{Lx} + \Delta c_{Lx} \\ 0 & f & c_{Ly} + \Delta c_{Ly} \\ 0 & 0 & 1 \end{bmatrix} \left[\mathbf{I} \middle| \begin{pmatrix} b/2 \\ 0 \\ 0 \end{pmatrix} \right] \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}. \tag{19}$$

Together with (11), this yields

$$\Delta\mathbf{x}_L = \Delta\mathbf{c}_L \tag{20}$$

which is identical to the result that had been achieved from (2) and (8). The vertical misalignments, disparity errors and range uncertainties for perturbed image center and other camera parameters are listed in Table I (please note that the given results are valid for calibration errors in both cameras). It is interesting to note that the effect of an image center offset $\Delta\mathbf{c}$ can be regarded as a superposition of a yaw and pitch error for small viewing angles. This suggests, that adjusted yaw and pitch angles of a stereo camera may—at least to a first order approximation—compensate erroneous image center coordinates. Consequently, a self-calibration algorithm that recovers the extrinsic parameters of the stereo rig may neglect the precise determination of the image centers. These analytical results coincide well with the experimental findings reported in [21]. Another important point is that the effect of orientation errors and errors in the focal lengths on the range uncertainty ΔZ increases linearly with the horizontal and vertical distance to the principal point in the image. In other words, objects that are close to the centers of the rectified images tend to be localized with higher accuracy than objects at the image borders.

From the sensitivity analysis presented in this section, we draw the conclusion that our stereo self-calibration algorithm should recover the focal lengths f_L, f_R of both cameras as well as the five extrinsic orientation parameters Ψ_R, Θ_R, Ψ_L, Φ_L, and Θ_L. Both base length b and principal points coordinates $\mathbf{c}_L, \mathbf{c}_R$ are excluded from the autocalibration method. The latter parameters are omitted because our analysis revealed a strong correlation between effects of errors in image center coordinates and errors in yaw and pitch angles. Since cameras are scale blind (i.e., cannot distinguish whether an object is 10 m wide and at a distance of 100 m or 1 m wide and at a distance of 10 m), the base length may not be recovered from image data alone. Depending on the application, additional information as, e.g., known object

TABLE I
SENSITIVITY OF 3-D RECONSTRUCTION TO ERRONEOUS CAMERA CALIBRATION

error source	**effect on normalized image coordinates** ("$\sim$" denotes normalized coordinates)	**effect on pixel coordinates** $\Delta\mathbf{x} = \begin{pmatrix}\Delta d \\ \Delta y\end{pmatrix}$ **disparity error** **scan line error**	**linear sensitivity of 3D reconstruction**
yaw error $\Delta\Psi_L$	$\boldsymbol{\Delta}\tilde{\mathbf{x}}_L \approx \begin{pmatrix}1+\tilde{x}_L^2 \\ \tilde{x}_L\tilde{y}_L\end{pmatrix}\Delta\Psi_L$	$\begin{aligned}\Delta d &\approx f_L(1+\tilde{x}_L^2)\Delta\Psi_L \\ \Delta y &\approx f_L\tilde{x}_L\tilde{y}_L\Delta\Psi_L\end{aligned}$	$\frac{\Delta Z}{\Delta\Psi} \approx -\frac{Z^2}{b}(1+\tilde{x}_L^2)$
pitch error $\Delta\Phi_L$	$\boldsymbol{\Delta}\tilde{\mathbf{x}}_L \approx -\begin{pmatrix}\tilde{x}_L\tilde{y}_L \\ 1+\tilde{y}_L^2\end{pmatrix}\Delta\Phi_L$	$\begin{aligned}\Delta d &\approx -f\tilde{x}_L\tilde{y}_L\Delta\Phi_L \\ \Delta y &\approx -f(1+\tilde{y}_L^2)\Delta\Phi_L\end{aligned}$	$\frac{\Delta Z}{\Delta\Phi_L} \approx \frac{Z^2}{b}\tilde{x}_L\tilde{y}_L$
roll error $\Delta\Theta_L$	$\boldsymbol{\Delta}\tilde{\mathbf{x}}_L \approx \begin{pmatrix}-\tilde{y}_L \\ \tilde{x}_L\end{pmatrix}\Delta\Theta_L$	$\begin{aligned}\Delta d &\approx (y_L - c_y)\,\Delta\Theta_L \\ \Delta y &\approx (x_L - c_x)\,\Delta\Theta_L\end{aligned}$	$\frac{\Delta Z}{\Delta\Theta_L} \approx \frac{Z^2}{b}\tilde{y}_L$
baseline error Δb	$\boldsymbol{\Delta}\tilde{\mathbf{x}}_{L/R} \approx \pm\begin{pmatrix}\frac{d_N}{2b} \\ 0\end{pmatrix}\Delta b$	$\begin{aligned}\Delta d &\approx \frac{\Delta b}{b}\,d \\ \Delta y &\approx 0\end{aligned}$	$\frac{\Delta Z}{\Delta b} \approx -\frac{Z}{b}$
center offset $\boldsymbol{\Delta}\mathbf{c}_L$	$\boldsymbol{\Delta}\mathbf{x}_L \approx \boldsymbol{\Delta}\mathbf{c}_L$	$\begin{aligned}\Delta d &\approx \Delta c_{L,x} \\ \Delta y &\approx \Delta c_{L,y}\end{aligned}$	$\frac{\Delta Z}{\Delta c_{L,x}} \approx -\frac{Z^2}{bf}$
focal length error Δf_L (one camera only)	$\boldsymbol{\Delta}\mathbf{x}_L \approx (\mathbf{x}_L - \mathbf{c})\frac{\Delta f_L}{f}$	$\begin{aligned}\Delta d &\approx (x_L - c_x)\frac{\Delta f_L}{f} \\ \Delta y &\approx (y_L - c_y)\frac{\Delta f_L}{f}\end{aligned}$	$\frac{\Delta Z}{\Delta f_L} \approx -\frac{Z^2}{bf^2}(x_L - c_x)$
focal length error Δf (both cameras)	$\boldsymbol{\Delta}\mathbf{x}_{L/R} \approx (\mathbf{x}_{L/R} - \mathbf{c})\frac{\Delta f}{f}$	$\begin{aligned}\Delta d &\approx \frac{\Delta f}{f}\,d \\ \Delta y &\approx 0\end{aligned}$	$\frac{\Delta Z}{\Delta f} \approx -\frac{Z}{f}$

lengths or given absolute velocities of the observer would be required to estimate b. Such information has not been used in this work. Furthermore, Table I indicates that base length errors may not be too critical for some applications: First, reconstruction errors scale linearly with Δb and small base length tolerances are feasible in the production line. Second, base length errors do not induce vertical misalignments and thus do not affect the reconstruction density of standard stereo matching.

IV. STEREO SELF-CALIBRATION

After modeling the stereo sensor and its parameters, the first step in designing a self-calibration algorithm is the definition of observation models that relate system parameters and (ideal) measurements. Different constraints can be exploited to obtain these relations: our framework utilizes the geometry of stereo image sequences (Section IV-A), stereo image pairs (Section IV-B), and image triplets (Section IV-C). Based on these constraints we derive different cost functions for stereo calibration. These can be used individually but may as well contribute to a combined optimization criterion. To achieve higher accuracy, special care is taken to employ physically relevant geometric instead of algebraic error functions.

Finally, a recursive optimization algorithm is designed that is adequate for continuous stereo self-calibration (Section IV-D). A mathematical derivation of this algorithm is given in the Appendix of this paper. Robustness is an important issue in practical self-calibration since occasional gross errors in the input data are commonly encountered due to, e.g., occlusions or repetitive patterns in the images.

A. Reduced Order Bundle Adjustment

To illustrate bundle adjustment, let us first consider a set of object points $\mathbf{X}_i$, $i \in S_b$. We assume that these points are moving rigidly through space, i.e., obey a simple 3-D motion model $\mathbf{X}_i(k+1) = \mathbf{R}(\boldsymbol{\omega}(k))\mathbf{X}_i(k) + \mathbf{V}(k)$, where the vectors $\boldsymbol{\omega}(k)$ and $\mathbf{V}(k)$ define the camera's 3-D rotation and translation at time k. The projections of these object points into the left and right images are denoted $\mathbf{x}_{R,i}, \mathbf{x}_{L,i}$, respectively, and we are given noisy measurements $\hat{\mathbf{x}}_{R,i}(k) = \mathbf{x}_{R,i}(k) + \mathbf{e}_{R,i}(k)$, $\hat{\mathbf{x}}_{L,i}(k) = \mathbf{x}_{L,i}(k) + \mathbf{e}_{L,i}(k)$. The positions errors $\mathbf{e}_{R,i}(k)$ and $\mathbf{e}_{L,i}(k)$ are assumed to be zero-mean and to have covariances $\mathbf{C}_{R,i}(k)$ and $\mathbf{C}_{L,i}(k)$, respectively.

The objective of bundle adjustment is to find the 3-D structure $\mathbf{X}_i$, the object motion $(\boldsymbol{\omega}, \mathbf{V})$, and the camera parameters $(\boldsymbol{\omega}_R, \boldsymbol{\omega}_L, f_R, f_L)$ such that the distance between the ideal projections $\mathbf{x}_{R,i}, \mathbf{x}_{L,i}$ and the measured coordinates $\hat{\mathbf{x}}_{R,i}, \hat{\mathbf{x}}_{L,i}$ is minimal over all frames k of a sequence. Bundle adjustment has been widely used in photogrammetry and as a refinement step for off-line camera calibration since it can provide highly accurate results. However, it has several shortcomings: First, it requires an initial guess of the parameters with sufficient quality to guarantee convergence to the global optimum. Second, it is usually implemented as a batch approach that requires that all input data is given at once. Third, the parameter space is high

dimensional since each tracked point $\mathbf{X}_i$ introduces three additional DOF, resulting in difficult and time consuming optimization procedures. As stated earlier, we assume that a sufficient initial guess is available and cover only the latter problems. A recursive parameter estimation method will be used (cf. Section IV-D), so that all data will be processed as soon as it arrives. To reduce the state dimension, we decompose each $\mathbf{X}_i$ into its projection onto the right image $\mathbf{x}_{R,i}$ and its depth ρ_i: ρ_i cannot be recovered directly and is thus included in the parameter vector, whereas $\mathbf{x}_{R,i}$ is treated as a (directly accessible) observation in the measurement constraint as will be shown later. Thus, in our formulation each tracked point introduces only one DOF and the dimension of the state vector is reduced significantly.

Assuming we are given the true image position $\mathbf{x}_{R,i}(k)$ and the true depth $\rho_i(k)$ of a tracked point, we can reconstruct $\mathbf{X}_i(k)$ via inverse pinhole projection

$$\mathbf{X}_i(k) = \mathbf{\Pi}_R^{-1}\left(\mathbf{x}_{R,i}(k), \rho_i(k)\right). \tag{21}$$

Using $\mathbf{X}_i(k)$, we are able to predict image positions $\mathbf{x}_{R,i}(k+1)$ and $\mathbf{x}_{L,i}(k)$

$$\begin{aligned}\mathbf{x}_{R,i}(k+1) &= \boldsymbol{\pi}_R\left(\mathbf{R}\left(\boldsymbol{\omega}(k)\right)\mathbf{X}_i(k) + \mathbf{V}(k)\right)\\ \mathbf{x}_{L,i}(k) &= \boldsymbol{\pi}_L\left(\mathbf{X}_i(k)\right).\end{aligned} \tag{22}$$

Thus, for each time instant k, (21) and (22) constitute an implicit constraint between the ideal measurements and the parameter vector $\mathbf{z}_b = [\rho_i, \boldsymbol{\omega}, \mathbf{V}, \boldsymbol{\omega}_R, \boldsymbol{\omega}_L, f_R, f_L]^T$

$$\begin{aligned}&\mathbf{h}_b\left(\mathbf{z}_b, \mathbf{x}_{R,i}(k), \mathbf{x}_{L,i}(k), \mathbf{x}_{R,i}(k+1)\right)\\ &= \begin{bmatrix}\boldsymbol{\pi}_R\left\{\mathbf{R}(\boldsymbol{\omega})\mathbf{\Pi}_R^{-1}\left(\mathbf{x}_{R,i}(k), \rho_i\right) + \mathbf{V}\right\} - \mathbf{x}_{R,i}(k+1)\\ \boldsymbol{\pi}_L\left\{\mathbf{\Pi}_R^{-1}\left(\mathbf{x}_{R,i}(k), \rho_i\right)\right\} - \mathbf{x}_{L,i}(k)\end{bmatrix}\\ &= \mathbf{0}.\end{aligned} \tag{23}$$

Now assume that we are measuring—over the whole image sequence—the temporal displacement of the tracked object point $\mathbf{X}_i$ between two consecutive frames of the right camera as well as the spatial displacement between the object point's coordinates in the right and left image. In other words, we are continuously observing noisy image positions $\hat{\mathbf{x}}_{R,i}(k)$, $\hat{\mathbf{x}}_{R,i}(k+1)$ and $\hat{\mathbf{x}}_{L,i}(k)$. Given these measurements, the objective of our self-calibration is to minimize at each k the pixel error[1]

$$\begin{aligned}\sum_{i\in S_b} &\left\|\hat{\mathbf{x}}_{R,i}(k) - \mathbf{x}_{R,i}(k)\right\|^2_{\mathbf{C}_{R,i}(k)} + \left\|\hat{\mathbf{x}}_{L,i}(k) - \mathbf{x}_{L,i}(k)\right\|^2_{\mathbf{C}_{L,i}(k)}\\ &+ \left\|\hat{\mathbf{x}}_{R,i}(k+1) - \mathbf{x}_{R,i}(k+1)\right\|^2_{\mathbf{C}_{R,i}(k+1)}\end{aligned} \tag{24}$$

subject to a constraint between camera parameters and ideal image positions $\mathbf{x}_{R,i}(k), \mathbf{x}_{R,i}(k+1), \mathbf{x}_{L,i}(k)$

$$\mathbf{h}_b\left(\mathbf{z}_b, \mathbf{x}_{R,i}(k), \mathbf{x}_{L,i}(k), \mathbf{x}_{R,i}(k+1)\right) = \mathbf{0} \tag{25}$$

evaluated for all features $i \in S_b$. Implicit measurement constraints as given by (24) and (25) are related to Gauss–Helmert models (e.g., [41]). We wish to emphasize that (24) minimizes a physically relevant geometric error corresponding to pixel distances in the image. In addition, the dimension of the parameter vector in our optimization problem is $N + 6(M-1) + 7M$ (where N is the number of tracked points and M is the number of stereo image pairs), while standard bundle adjustment extended to parameter drift would require $3N + 6(M-1) + 7M$ elements.

[1] $\|\hat{\mathbf{x}} - \mathbf{x}\|^2_C$ denotes the normalized squared distance $(\hat{\mathbf{x}} - \mathbf{x})^T\mathbf{C}^{-1}(\hat{\mathbf{x}} - \mathbf{x})$, where $\mathbf{C}$ is the covariance matrix representing our uncertainty in the measurement $\hat{\mathbf{x}}$.

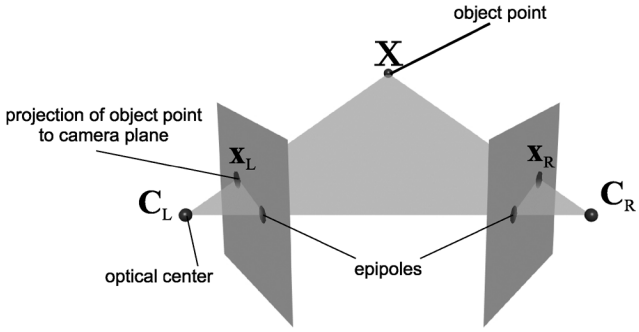

Fig. 3. Epipolar constraint. The optical rays of both images image points $\mathbf{x}_L, \mathbf{x}_R$ have to lie within the same plane.

B. Epipolar Constraint

The epipolar constraint has been known in photogrammetry since the beginning of the 20th century [42] and was introduced to the computer vision community by [43]. It constitutes an elementary relation between two stereo images: Consider two cameras viewing the same scene from different positions and a pair of corresponding points $\mathbf{x}_L$ and $\mathbf{x}_R$ within those images as depicted in Fig. 3. Geometrically, it is clear that the optical centers of both cameras and the image points $\mathbf{x}_L$ and $\mathbf{x}_R$ all lie in the same plane. Mathematically, this relation is expressed via the *fundamental matrix* $\mathbf{F}$ (cf. [44])

$$\begin{pmatrix}\mathbf{x}_L\\ 1\end{pmatrix}^T \mathbf{F} \begin{pmatrix}\mathbf{x}_R\\ 1\end{pmatrix} = 0 \tag{26}$$

where

$$\mathbf{F} = \mathbf{K_L}^{-T}\mathbf{R}_L[\mathbf{C}_R - \mathbf{C}_L]_\times \mathbf{R}_R^T\mathbf{K_R}^{-1} \tag{27}$$

and $[.]_\times$ denotes the skew-symmetric matrix operator, i.e.,

$$[\mathbf{T}]_\times = \begin{bmatrix}0 & -T_Z & T_Y\\ T_Z & 0 & -T_X\\ -T_Y & T_X & 0\end{bmatrix}. \tag{28}$$

Please note that the epipolar constraint (26) does not involve the 3-D position of the observed object point, i.e., the epipolar constraint decouples the extrinsic camera parameters from the 3-D structure of the observed scene.

Given noisy image positions $\hat{\mathbf{x}}_{R,i} = \mathbf{x}_{R,i} + \mathbf{e}_{R,i}$ and $\hat{\mathbf{x}}_{L,i} = \mathbf{x}_{L,i} + \mathbf{e}_{L,i}$, our objective is to find the camera parameters $(\boldsymbol{\omega}_R, \boldsymbol{\omega}_L, f_R, f_L)$ that minimize the sum of squared pixel errors

$$\sum_{i\in S_e} \left\|\hat{\mathbf{x}}_{R,i} - \mathbf{x}_{R,i}\right\|^2_{\mathbf{C}_{R,i}} + \left\|\hat{\mathbf{x}}_{L,i} - \mathbf{x}_{L,i}\right\|^2_{\mathbf{C}_{L,i}} \tag{29}$$

subject to the epipolar constraint (26) abbreviated by

$$h_e(\mathbf{F}, \mathbf{x}_{L,i}, \mathbf{x}_{R,i}) = 0 \quad \forall i \in S_e. \tag{30}$$

Each point correspondence yields one constraint equation. However, as is also clear from Fig. 3, the epipolar constraint constitutes only a necessary condition for corresponding image points: each pair of corresponding points must fulfill the epipolar constraint, but not all points that satisfy (26) may be images of a single object point. Thus, the epipolar constraint neglects matching errors along the epipolar line.

Although the epipolar constraint has some theoretical disadvantages compared to bundle adjustment since it does not provide as much information and cannot achieve the same level of accuracy, it still provides some practical benefits: First, the parameter space for self-calibration is much smaller. Since the epipolar constraint decouples camera parameters from 3-D structure, the cost function (29) only depends on the parameters ω_R and ω_L, while (24) additionally requires one depth for each observed point and the motion of the camera. The lower dimension of the parameter space simplifies the optimization problem and reduces the computational effort significantly. Second, the epipolar constraint does not require temporal feature matching between consecutive image frames. Since all information is gathered instantaneously, the epipolar constraint will still hold in the presence of independently moving objects in the scene.

C. Trilinear Constraints

The geometric relation between coordinates of corresponding points in three images is captured by the trilinear constraints [25]. As in the epipolar constraint described in the previous section, the trilinearities decouple scene structure from camera calibration since they do not require the 3-D position of the observed point explicitly. They provide a sufficient condition for three image coordinates to correspond to the same object point without the deficiency of a set of epipolar constraints (cf. Section II). For more details on the trilinear constraint, the reader is referred to [34] and [41].

Consider a triplet of corresponding image points $\mathbf{x}_R(k), \mathbf{x}_L(k), \mathbf{x}_R(k+1)$ in the current right, left and subsequent right camera frame, respectively. For brevity, we denote these image positions by $\mathbf{x}^A, \mathbf{x}^B, \mathbf{x}^C$ and their associated projection matrices are given by

$$\mathbf{A} = \mathbf{K}_R \mathbf{R}_R [\mathbf{I}, -\mathbf{C}_R] \tag{31}$$

$$\mathbf{B} = \mathbf{K}_L \mathbf{R}_L [\mathbf{I}, -\mathbf{C}_L] \tag{32}$$

$$\mathbf{C} = \mathbf{K}_R \mathbf{R}_R [\mathbf{R}(\boldsymbol{\omega}(k)), \mathbf{V}(k) - \mathbf{C}_R] \tag{33}$$

where $(\mathbf{R}(\boldsymbol{\omega}(k)), \mathbf{V}(k))$ describes the 3-D motion of the stereo rig.

The geometry of three cameras can be captured elegantly by the *trifocal tensor* [45]. The trifocal tensor has 27 elements, but at most 18DOF are required to fully specify the camera configuration. In this contribution, we employ a Euclidean parametrization of the trifocal tensor that has even less DOF since we assume that the intrinsic parameters do not change between two stereo frames and we neglect image skew and aspect ratio. The trifocal tensor $\mathbf{T}$ is computed from the projection matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ as

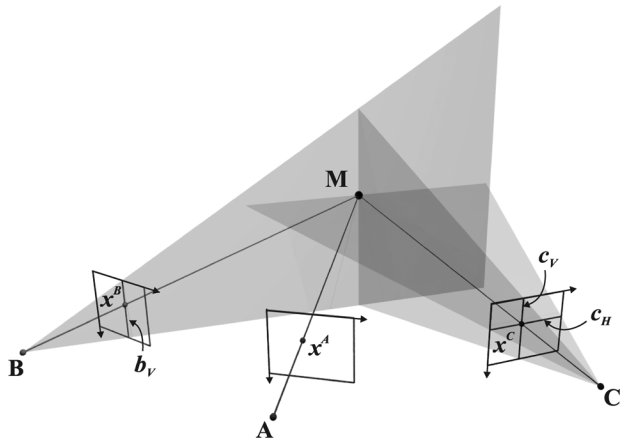

Fig. 4. Geometric interpretation of the trilinear constraints in (37). The 3-D point $\mathbf{M}$ is reconstructed as the intersection of the optical ray associated with $\mathbf{x}^A$ and the plane containing $\mathbf{B}$ and the line $\mathbf{b}_V$, where $\mathbf{b}_V$ is parallel to the y-axis of the camera coordinate system and passes through $\mathbf{x}^B$. The two trilinear constraints used in this paper state that $\mathbf{M}$ should lie in the two planes defined by the optical center $\mathbf{C}$ and the horizontal line $\mathbf{c}_H$ and the vertical line $\mathbf{c}_V$, respectively.

$$\mathbf{T}_l^{qr} = (-1)^{l+1} \det \begin{bmatrix} \sim \mathbf{a}^l \\ \mathbf{b}^q \\ \mathbf{c}^r \end{bmatrix}. \tag{34}$$

$\mathbf{b}^q$ and $\mathbf{c}^r$ refer to the rth and qth row of the matrices $\mathbf{B}$ and $\mathbf{C}$, respectively. $\sim \mathbf{a}^l$ is the matrix $\mathbf{A}$ without the lth row.

The trifocal tensor allows the formulation of constraints between point triplets that are multilinear in the elements of the trifocal tensor and in the image coordinates

$$\begin{aligned} &g_{qr}(\mathbf{T}, \mathbf{x}^A, \mathbf{x}^B, \mathbf{x}^C) \\ &\quad = \sum_{l=1}^{3} \mathbf{x}_l^A \left(x_q^B \mathbf{x}_r^C T_l^{33} - x_r^C T_l^{q3} - x_q^B T_l^{3r} + T_l^{qr} \right) \\ &\quad = 0. \end{aligned} \tag{35}$$

The linearity is of course not pertained if a Euclidean parametrization in intrinsic and extrinsic camera parameters is used. However, the experiments in [11], [27] suggest that the integration of nonlinear restrictions between the calibration parameters improves the achievable accuracy of the camera calibration (given an initial guess with sufficient quality for nonlinear optimization). Equation (35) actually yields nine constraints for the possible choices of $q, r \in \{1, 2, 3\}$. Four of these constraints are linearly independent [25], but as shown in [41], the trilinearities impose only three constraints on the geometry of the image triplet if a minimal parametrization is used. An optimal choice of the constraints is non trivial, in fact the selection of the constraints should be adapted to the current motion of the stereo and the position of the observed 3-D point. This has not yet been implemented in our work. Instead, we found that for our stereo rig with fixed base length, a combination of two trilinear constraints $(q, r) = (1,1), (1,2)$ and the epipolar constraint between the left and right stereo frame gives adequate results. A geometric interpretation of these two trilinear constraints based on [46] is given in Fig. 4. It

is obvious from this illustration that the transfer problem using the chosen set of constraints is only degenerate if the the point $\mathbf{M}$ lies on the line containing the optical centers $\mathbf{A}$ and $\mathbf{B}$. This configuration, however, is irrelevant for typical stereo vision applications.

Using the selected constraints and given a set S_t of corresponding point triplets in three images, our self-calibration algorithm has to minimize the cost function

$$\sum_{i\in S_t} \|\hat{\mathbf{x}}_{R,i}(k) - \mathbf{x}_{R,i}(k)\|^2_{\mathbf{C}_{R,i}(k)} + \|\hat{\mathbf{x}}_{L,i}(k) - \mathbf{x}_{L,i}(k)\|^2_{\mathbf{C}_{L,i}(k)} + \|\hat{\mathbf{x}}_{R,i}(k+1) - \mathbf{x}_{R,i}(k+1)\|^2_{\mathbf{C}_{R,i}(k+1)} \quad (36)$$

subject to the constraint

$$\begin{bmatrix} g_{11}\left(\mathbf{T}, \mathbf{x}_{R,i}(k), \mathbf{x}_{L,i}(k), \mathbf{x}_{R,i}(k+1)\right) \\ g_{12}\left(\mathbf{T}, \mathbf{x}_{R,i}(k), \mathbf{x}_{L,i}(k), \mathbf{x}_{R,i}(k+1)\right) \\ h_e\left(\mathbf{F}, \mathbf{x}_{L,i}(k), \mathbf{x}_{R,i}(k)\right) \end{bmatrix} = \mathbf{0} \quad (37)$$

for all $i \in S_t$.

D. Recursive Optimization

Three different geometric constraints for stereo self-calibration have been discussed in the preceding sections and formulated in a common Gauss–Helmert type model. In the Appendix of this paper, we derive a recursive algorithm for such models. The algorithm is an adaptation of an Iterated Extended Kalman Filter (IEKF) and can readily be applied for our continuous self-calibration. It is important to note that the proposed algortihm differs from standard IEKF methods since it recovers not only the camera parameters but also corrected measurements. These corrected observations can be considered nuissance states which are required for subsequent relinearization of the nonlinear implicit constraint functions. The state vector $\mathbf{z}$ determined by our algorithm comprises the depths of all N tracked bundle adjustment features in the set S_b, object motion, and camera parameters

$$\mathbf{z} = [\rho_1, \ldots, \rho_N, \boldsymbol{\omega}, \mathbf{V}, \Psi_R, \Theta_R, \Psi_L, \Phi_L, \Theta_L, f_R, f_L]^T. \quad (38)$$

While the measurement equations of the filter are given by (24), (25), (29), (30), (36), and (37), we still need to define the system model that governs the dynamics of our state vector. For simplicity, we assume that the camera is moving with a constant velocity model pertubated by Gaussian white noise, i.e.,

$$\begin{bmatrix} \boldsymbol{\omega}(k+1) \\ \mathbf{V}(k+1) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\omega}(k) \\ \mathbf{V}(k) \end{bmatrix} + \begin{bmatrix} \mathbf{n}_\omega(k) \\ \mathbf{n}_V(k) \end{bmatrix}. \quad (39)$$

If additional information is available (such as, e.g., commanded steering angles and accelerations or more precise vehicle motion models), it should be incorporated at this point. However, we found that the simple motion model already provides good results in our experiments. The depths ρ_i then evolve as

$$\rho_i(k+1) = [0,0,1]\left(\mathbf{R}\left(\boldsymbol{\omega}(k)\right)\mathbf{X}_i(k) + \mathbf{V}(k)\right), \quad i \in S_b \quad (40)$$

with $\mathbf{X}_i(k) = \boldsymbol{\Pi}_R^{-1}(\mathbf{x}_{R,i}(k), \rho_i(k))$. The dynamics of the extrinsic camera parameters are assumed to be governed by

$$\begin{bmatrix} \Psi_R(k+1) \\ \Theta_R(k+1) \\ \Psi_L(k+1) \\ \Phi_L(k+1) \\ \Theta_L(k+1) \end{bmatrix} = \begin{bmatrix} \Psi_R(k) \\ \Theta_R(k) \\ \Psi_L(k) \\ \Phi_L(k) \\ \Theta_L(k) \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_R(k) \\ u_L(k) \end{bmatrix} + \mathbf{n}_{\omega_R,\omega_L}(k) \quad (41)$$

where u_R, u_L denote the commanded yaw angles of the right and left stereo camera, respectively. $\mathbf{n}_{\omega_R,\omega_L}$ specifies the system noise associated with the extrinsic camera parameters. Similarly, constant focal lengths are assumed with additive Gaussian white noise $\mathbf{n}_f$

$$\begin{bmatrix} f_R(k+1) \\ f_L(k+1) \end{bmatrix} = \begin{bmatrix} f_R(k) \\ f_L(k) \end{bmatrix} + \mathbf{n}_f(k). \quad (42)$$

If no command signals are sent to the active stereo rig, the standard deviations of the noise components $\mathbf{n}_{\omega_R,\omega_L}, \mathbf{n}_f$ are set to small values ($\approx 10^{-5}$) to account for linearization errors. If new gaze directions are commanded, the standard deviations of $\mathbf{n}_{\omega_R,\omega_L}$ are temporarily increased to 1° to make up for mechanical inaccuracies. Equations (39)–(42) represent our knowledge of the stereo rig and its motion. The system model enforces smoothness of the estimated camera parameters over time, especially in configurations when certain camera parameters are difficult to observe (e.g., when all acquired feature points have approximately the same depth).

Feature points for recursive bundle adjustment, epipolar constraint, and trilinear constraints are acquired using Lowe's SIFT feature detector [47]. In addition, the search region used for feature matching is predicted using the current filter state and its uncertainty. We also experimented with a corner finder as described in [48] and correlation based matching with subsequent accuracy evaluation as presented in [49]. However, we found that SIFT gave better self-calibration results for our active stereo rig with horizontal gaze directions ranging from -25° to 25°.

As indicated earlier, robustness is an essential property of our self-calibration algorithm since feature matching is prone to occasional gross errors due to periodic patterns or occlusions. Additionally, there may be independently moving objects in the scene that violate the rigidity assumption required for the trilinear constraints and bundle adjustment. We have thus employed a Least Median of Squares random sampling scheme in the innovation stage of our algorithm as proposed in [50] (cf. Appendix). This method eliminates outliers in the input data and reduces the sensitivity of our algorithm to model violations.

V. Experimental Evaluation

A. Synthetic Data

The objective of our synthetic data experiments was to evaluate and compare the performance of the different self-calibration constraints described in the previous sections. We randomly generated synthetic stereo sequences of a moving point cloud with 40 points. Each sequence was 50 stereo frames long and

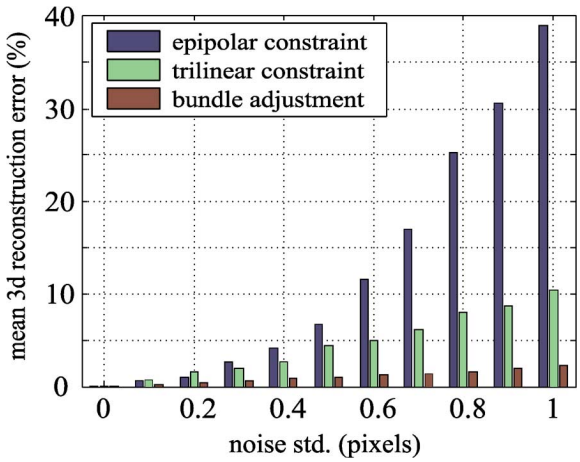

Fig. 5. Comparison of self-calibration results on various noise levels. The plot shows the mean 3-D reconstruction error obtained with the different self-calibration constraints: a: using the epipolar constraint, b: using the trilinear constraint, and c: using reduced order bundle adjustment.

Gaussian white noise was added to the image coordinates of all points in both images. The initial guess for the stereo calibration deviated 2° in each component from the true extrinsic parameters and differed by 10% from the true focal lengths.

To assess the self-calibration results after each simulation run, we compute the mean relative 3-D reconstruction error of all points in the last frame of the sequence. Given the true, noise-free image coordinates $\mathbf{x}_L$ and $\mathbf{x}_R$ in both images and the estimated camera parameters, we can determine the 3-D position $\hat{\mathbf{X}}$ of the corresponding object point using Hartley's triangulation method [35]. The relative 3-D reconstruction error is then computed as

$$\epsilon_{rel} = \frac{||\hat{\mathbf{X}} - \mathbf{X}||}{||\mathbf{X}||} \tag{43}$$

where $\mathbf{X}$ denotes the true 3-D position.

Fig. 5 depicts the results of the proposed algorithm. The standard deviations of the pixel error varied from 0 to 1 pixels and 50 independent simulations were run on each noise level. We compared three different versions of the algorithm: a) using only the epipolar constraint, b) using only the trilinear constraint, and c) using reduced order bundle adjustment only. As indicated above, bundle adjustment is the most complex method since it involves the largest parameter space of all three methods and requires temporal, as well as spatial correspondences. However, it gives the most accurate results (reduced order bundle adjustment outperforms the trilinear constraint by a factor of three; compared to the epipolar constraint, reduction of the mean 3-D reconstruction error is one order of magnitude). Still, the epipolar constraint has advantages in practical application: Since it does not require any restrictions like rigidity on the observed scene, it is unaffected by independently moving objects. Furthermore, it yields calibration parameters instantaneously after processing of a single stereo image pair. In the real imagery examples presented in the next section, we thus employ the epipolar constraint in combination with bundle adjustment to expedite and stabilize the self-calibration process.

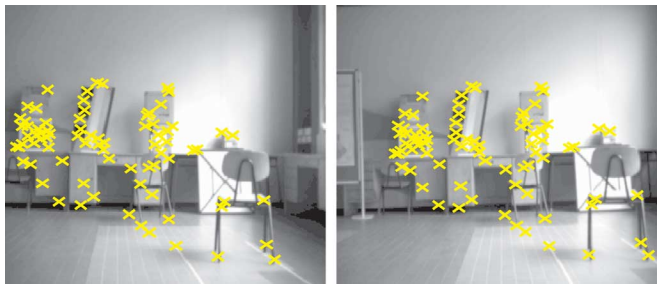

Fig. 6. Sample stereo frame of the indoor sequence. Stereo correspondences obtained with the SIFT key point detector are marked with "×". The sequence consisted of 305 stereo frames and both cameras were rotated about roughly 20° to the right after 199 frames.

B. Natural Imagery

Our algorithm was also evaluated on a variety of real imagery sequences. Here we describe two representative examples in indoor and outdoor environments. All stereo images have been acquired with an active stereo rig as described in Section III. In both examples, the intrinsic camera parameters were constant while the extrinsic parameters haven been varied throughout the sequences. Radial and tangential distortion effects have been removed prior to our calibration tracking algorithm.

The first sequence was recorded by a mobile platform traveling through a laboratory environment (Fig. 6). To obtain a reference calibration, we determined the camera parameters at the beginning of the sequence with Bouguet's excellent off-line camera calibration toolbox.[2] However, we used a manual initial guess for the calibration parameters that differed from the reference calibration ($\sim 3°$ in camera orientation and $\sim$10% in the focal lengths) to start our self-calibration algorithm. The sequence was 305 frames long and included a change in the gaze direction after 199 frames. The self-calibration algorithm combined reduced order bundle adjustment and the epipolar constraint, where at most 50 points were tracked over time and no more than 20 points constituted the epipolar constraint. Fig. 7 depicts the results of our algorithm.

To assess the performance of our algorithm, we used the mean of the relative 3-D reconstruction error (43) for the stereo features in Fig. 6. The off-line calibration results were used to generate a ground truth for the 3-D positions. This ground truth was compared to the triangulation obtained with the self-calibration parameters after 100 frames. Table II summarizes the self-calibration results. Using the manual guess for the stereo calibration, only 63 of the 91 stereo features yielded a position in front of the cameras. However, these 63 positions still had a mean 3-D reconstruction error of approximately 50%. After 100 frames, all features could be reconstructed and the results are comparable to off-line calibration.

The stereo self-calibration was also tested in an outdoor scenario with our experimental vehicle. Fig. 9 shows sample frames of a sequence with extracted correspondence features. We have chosen a version of our algorithm combining at most 30 tracking features in bundle adjustment and 30 stereo features in the epipolar constraint. The vehicle was first driving straight for about 40 frames and then made a left turn. This is also

[2]http://www.vision.caltech.edu/bouguetj/calib_doc.

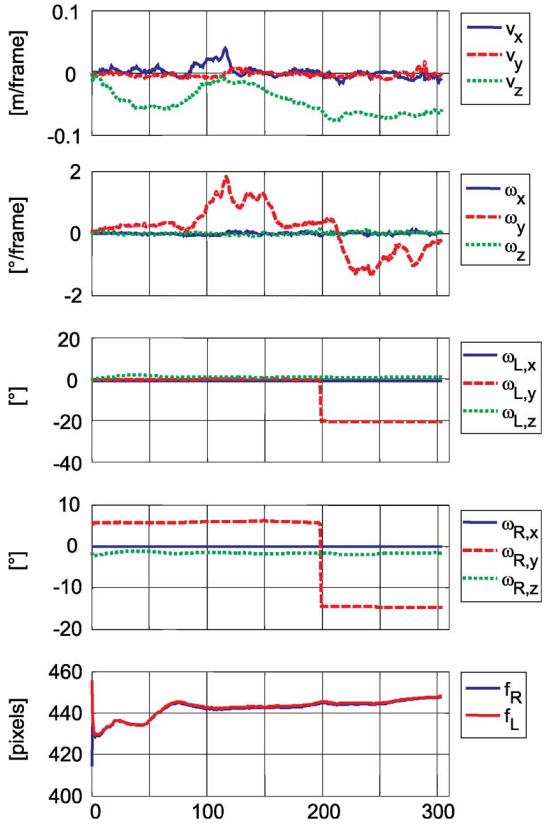

Fig. 7. Extracted ego motion $(\mathbf{v}, \boldsymbol{\omega})$, extrinsic camera orientations $(\boldsymbol{\omega}_R, \boldsymbol{\omega}_L)$ and focal lengths (f_R, f_L) for indoor sequence (Fig. 6). The camera rotations are tracked accurately.

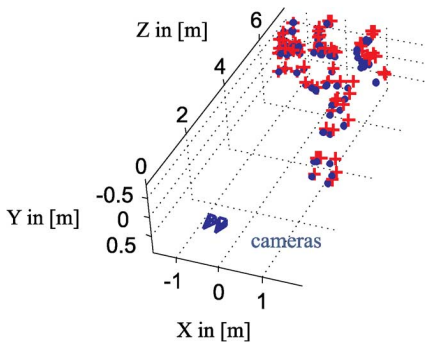

Fig. 8. Three-dimensional reconstruction of the stereo features from Fig. 6. "." indicates reconstructed 3-D positions using camera parameters obtained with Bouguet's offline calibration. "+" marks the reconstruction results using the self-calibration parameters obtained after 100 frames.

TABLE II
COMPARISON OF 3-D RECONSTRUCTION RESULTS

	no. of valid reconstructions	rel. reconstruction error
offline calibration	91 of 91	—
initial guess	63 of 91	49.46%
online calibration	91 of 91	2.36%

clearly shown in the estimated motion parameters $\mathbf{v}$ and $\boldsymbol{\omega}$ (Fig. 10). The cameras were rotated twice in the sequence: first about 15° to the left before starting the turn at frame 38, second about −15° after completing the turn (frame 168). Both

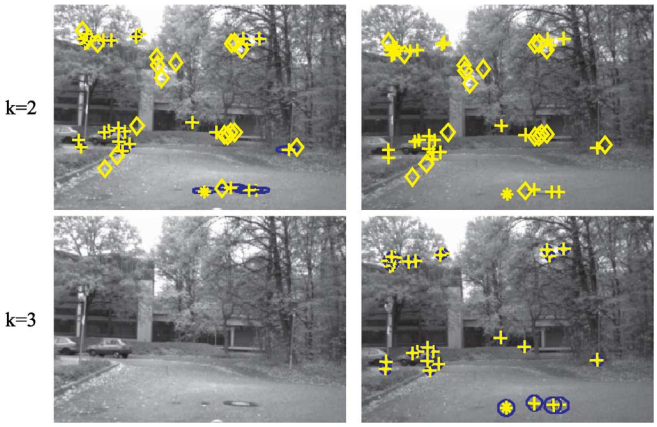

Fig. 9. Second and third stereo frame of sample sequence. The automatically selected features are also shown: +: successfully tracked features, ◇: stereo features for epipolar constraint, ∗: invalid tracking features. The predicted positions of the tracked features are marked by their covariance ellipses.

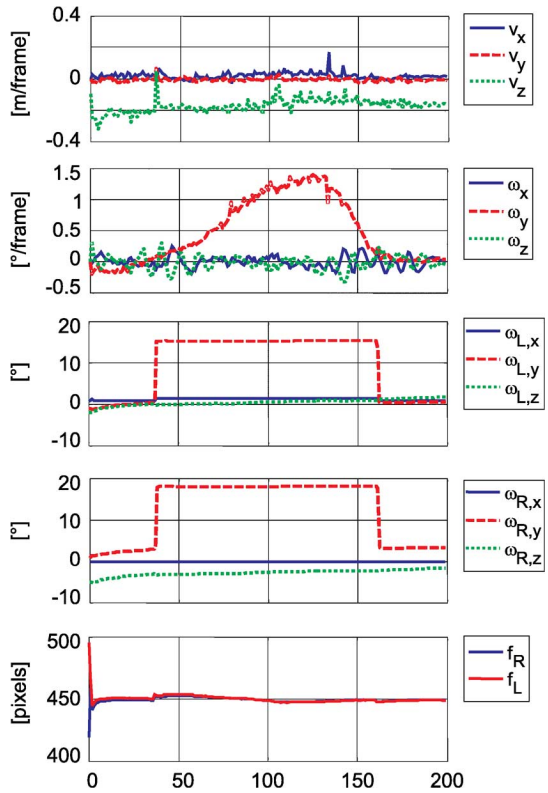

Fig. 10. Extracted 6d ego motion (rotation and translation) and camera parameters for the sequence depicted in Fig. 9. The cameras were rotated twice: 15° to the left before starting the turn (frame 38) and then 15° to the right after leaving the curve in frame 162.

changes in the gaze direction are captured by the self-calibration.

For a quality analysis of the calibration tracking in the outdoor example, we used the self-calibration results in a standard stereo vision process: Stereo reconstruction was obtained by first rectifying the images with the estimated camera parameters (using the method proposed in [51]), so that all corresponding pixels in both rectified frames should have the same y-coordinate. Then, correlation based matching as described in [52] was performed. Please note that we fully relied on stereo rectification

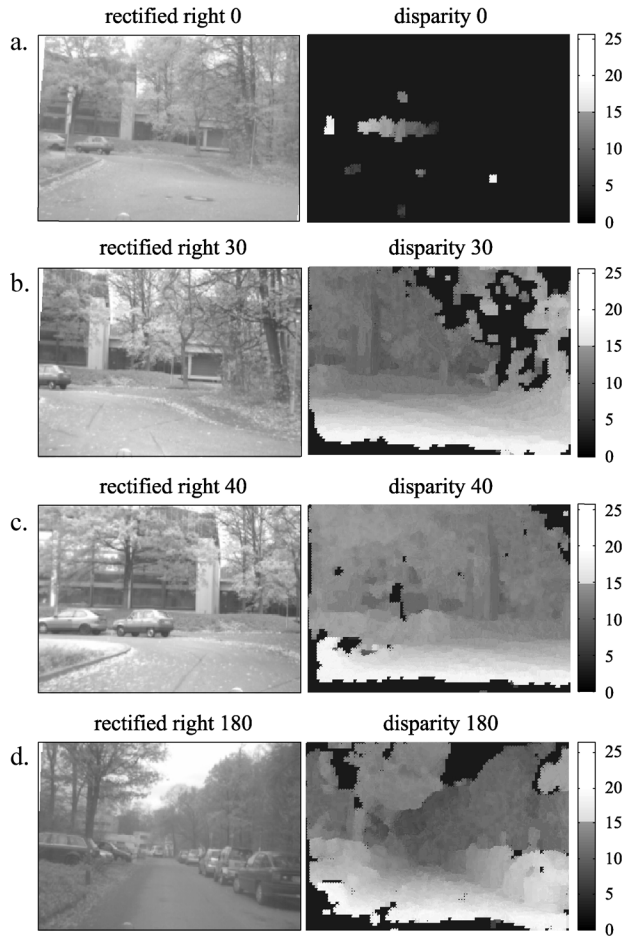

Fig. 11. Stereo reconstruction results obtained with the stereo calibration from Fig. 10. (a) Initial guess of the camera parameters at frame 0. Note that no offline stereo calibration was employed, the initial calibration parameters are just manually set and do not allow meaningful 3-D reconstruction. (b) Reconstruction at frame 30. (c) Reconstruction after first camera rotation (frame 40). (d) Reconstruction after second camera rotation (frame 180).

and used only a 1-D search region for stereo matching, so that erroneous camera parameters greatly influence the matching performance.

The top row of Fig. 11 displays the stereo reconstruction results using the initial stereo parameters. As the initial parameter setting was just a manual guess of the camera calibration, stereo reconstruction was not possible here. At frame 30—before rotating the cameras to the left—, the self-calibration has converged to camera parameters and allows reliable stereopsis. The estimated stereo calibration even remains valid after the camera rotations in frames 38 and 168, respectively, and gives satisfying results over the whole sequence.

VI. Conclusion

This article proposes a novel algorithm for continuous self-calibration of stereo cameras based on geometric error criteria. It relies on a consistent derivation of a robust, recursive optimization scheme for Gauss–Helmert models. The algorithm allows combining different geometric constraints (i.e., epipolar constraints for stereo images, trilinear constraints for image triplets and collinearity constraints in reduced order bundle adjustment) in a common framework.

Examples on synthetic and real imagery demonstrate the effectiveness and advantages of our algorithm: Our iterative method has proven the ability to refine an initial guess of the camera parameters as well as to continuously track drift in the stereo calibration parameters. The combination of bundle adjustment and epipolar constraint features sets ensures both high accuracy as well as robustness against independently moving objects that are often encountered in real imagery. It was shown that our algorithm can provide reliable results that allow stereo reconstruction with relative 3-D errors of less than 5% as compared to offline calibration.

Another contribution of this work is a thorough sensitivity analysis of 3-D stereo reconstruction to errors in the camera calibration parameters. This analysis allows to quantify the importance of the individual camera parameters for stereo self-calibration and to compute tolerance limits for camera calibration that guarantee desired 3-D reconstruction accuracy. The analysis reveals correlations between the effects of calibration error on stereo reconstruction and helps to decide which parameters should be considered for online self-calibration.

Future work beyond the scope of this article includes the autocalibration of lens distortion parameters. The reliable estimation of these parameters is a necessary step to replace costly offline calibration by online self-calibration. Additionally, open issues arise from the sensitivity analysis presented in Section V. First, the correlation between the effects of several camera parameters on 3-D reconstruction errors gives rise to the question for "optimal" camera parameterisations for stereo self-calibration. Another interesting subject is to evaluate the influence of feature points with various positions or various motions on 3-D reconstruction accuracy. This topic is closely related to critical motion sequences. It may help to automatically identify input data points with highest information gain in the current situation and to decide online whether or not to update certain parameters in a given scenario.

Appendix
Robust Iterated Extended Kalman Filter With Implicit Measurement Constraint

Our optimization problem is fully specified by a nonlinear system equation and an implicit measurement constraint equation: The state vector $\mathbf{z}(k)$ evolves according to a stochastic difference equation

$$\mathbf{z}(k+1) = \mathbf{f}\left(\mathbf{z}(k), \mathbf{u}(k) + \mathbf{w}(k)\right) + \mathbf{v}(k) \tag{44}$$

where the initial state is given by $\mathbf{z}(0) = \mathbf{z}_0$, $\mathbf{f}$ is the nonlinear transition function and $\mathbf{u}(k)$ is the control vector. In our case, both noise components $\mathbf{w}(k)$ and $\mathbf{v}(k)$ are modeled as realizations of independent Gaussian random variables with covariance matrices $\mathbf{Q}_B(k)$ and $\mathbf{Q}_S(k)$, respectively. We are also given noisy observations

$$\hat{\mathbf{x}}(k) = \mathbf{x}(k) + \mathbf{e}(k) \tag{45}$$

where again $\mathbf{e}(k)$ represents Gaussian white noise with covariance $\mathbf{R}(k)$, and the ideal observations $\mathbf{x}(k)$ need to satisfy a nonlinear implicit constraint

$$\mathbf{h}\left(\mathbf{z}(k), \mathbf{x}(k)\right) = \mathbf{0}. \tag{46}$$

For self-calibration, we seek to minimize the sum of normalized squared observation errors $\mathbf{e}$. This optimization problem can be solved recursively by an adaptation of a robust iterated extended Kalman filter (IEKF) for implicit measurement constraints. The IEKF alternates between three different stages—prediction, robust innovation, and relinearization of the measurement constraint—to determine an optimal state estimate $\mathbf{z}$ and its associated covariance matrix $\mathbf{P}$.

Prediction: Let $\mathbf{z}^+(k)$ denote the *a posteriori* state estimate at time k (i.e., our best estimate of the system parameters given all previous information) and $\mathbf{P}^+(k)$ the corresponding covariance. Given $\mathbf{z}^+(k)$ and $\mathbf{P}^+(k)$, we can predict the *a priori* state vector $\mathbf{z}^-(k+1)$ and its covariance matrix $\mathbf{P}^-(k+1)$ in the next time step $k+1$ using the standard formulas of an extended Kalman filter (e.g., [53] and [54])

$$\mathbf{z}^-(k+1) = \mathbf{f}\left(\mathbf{z}^+(k), \mathbf{u}(k)\right) \tag{47}$$

$$\mathbf{P}^-(k+1) = \mathbf{F}\mathbf{P}^+(k)\mathbf{F}^T + \mathbf{B}\mathbf{Q}_B(k)\mathbf{B}^T + \mathbf{Q}_S(k) \tag{48}$$

where $\mathbf{F} = \partial\mathbf{f}/\partial\mathbf{z}|_{\mathbf{z}^+(k),\mathbf{u}(k)}$ and $\mathbf{B} = \partial\mathbf{f}/\partial\mathbf{u}|_{\mathbf{z}^+(k),\mathbf{u}(k)}$.

Robust Innovation: After completing the prediction step, we incorporate the current measurement $\hat{\mathbf{x}}(k)$ to compute both the *a posteriori* state estimate $\mathbf{z}^+(k)$ and $\mathbf{x}^+(k)$, the best estimate of the ideal observation vector $\mathbf{x}(k)$. This is achieved by minimizing (for brevity, the time index k is omitted in the following paragraphs)

$$\begin{pmatrix} \mathbf{x}^+ - \hat{\mathbf{x}} \\ \mathbf{z}^+ - \mathbf{z}^- \end{pmatrix}^T \begin{bmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}^- \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{x}^+ - \hat{\mathbf{x}} \\ \mathbf{z}^+ - \mathbf{z}^- \end{pmatrix} \tag{49}$$

subject to the nonlinear constraint in (46).

To solve this optimization problem, we first linearize (46) about an operation point $(\check{\mathbf{z}}, \check{\mathbf{x}}) = (\mathbf{z}^-, \hat{\mathbf{x}})$ that corresponds to our best available guess of the true parameters. We obtain

$$\mathbf{h}(\mathbf{z}^+, \mathbf{x}^+) \approx \mathbf{A}\mathbf{z}^+ + \mathbf{B}\mathbf{x}^+ + \mathbf{y} \tag{50}$$

where $\mathbf{A} = -(\partial\mathbf{h}/\partial\mathbf{z})|_{\mathbf{z}^-,\hat{\mathbf{x}}}$, $\mathbf{B} = \partial\mathbf{h}/\partial\mathbf{x}|_{\mathbf{z}^-,\hat{\mathbf{x}}}$ and $y = \mathbf{h}(\mathbf{z}^-, \hat{\mathbf{x}}) - \mathbf{A}\mathbf{z}^- - \mathbf{B}\hat{\mathbf{x}}$. Using Lagrangian multipliers, we then have to find the extremum of the cost function

$$J = \begin{pmatrix} \mathbf{x}^+ - \hat{\mathbf{x}} \\ \mathbf{z}^+ - \mathbf{z}^- \end{pmatrix}^T \begin{bmatrix} \mathbf{R}^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{P}^-)^{-1} \end{bmatrix} \begin{pmatrix} \mathbf{x}^+ - \hat{\mathbf{x}} \\ \mathbf{z}^+ - \mathbf{z}^- \end{pmatrix} - 2\boldsymbol{\eta}(\mathbf{A}\mathbf{z}^+ + \mathbf{B}\mathbf{x}^+ + \mathbf{y}). \tag{51}$$

Taking the derivatives of J with respect to the corrected measurement $\mathbf{x}^+$, the corrected state vector $\mathbf{z}^+$, and the Lagrangian multiplier $\boldsymbol{\eta}$, we get after some algebra

$$\frac{\partial J}{\partial \mathbf{x}^+} = \mathbf{0} \iff \mathbf{x}^+ = \hat{\mathbf{x}} + \mathbf{R}\mathbf{B}^T\boldsymbol{\eta} \tag{52}$$

$$\frac{\partial J}{\partial \mathbf{z}^+} = \mathbf{0} \iff \mathbf{z}^+ = \mathbf{z}^- + \mathbf{P}^-\mathbf{A}^T\boldsymbol{\eta} \tag{53}$$

$$\frac{\partial J}{\partial \boldsymbol{\eta}} = \mathbf{0} \iff \mathbf{A}\mathbf{z}^+ + \mathbf{B}\mathbf{x}^+ + \mathbf{y} = 0. \tag{54}$$

Substituting (52) and (53) in (54) yields

$$(\mathbf{A}\mathbf{P}^-\mathbf{A}^T + \mathbf{B}\mathbf{R}\mathbf{B}^T)\boldsymbol{\eta} = -\mathbf{A}\mathbf{z}^- - \mathbf{B}\hat{\mathbf{x}} - \mathbf{y}. \tag{55}$$

For simplicity, we define a transformed observation vector

$$\hat{\mathbf{x}}^* = -\mathbf{B}\hat{\mathbf{x}} - \mathbf{y}, \quad \mathbf{R}^* = \mathbf{B}\mathbf{R}\mathbf{B}^T \tag{56}$$

and obtain from (55)

$$\boldsymbol{\eta} = (\mathbf{A}\mathbf{P}^-\mathbf{A}^T + \mathbf{R}^*)^{-1}(\hat{\mathbf{x}}^* - \mathbf{A}\mathbf{z}^+). \tag{57}$$

Please note that in our application, the matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{R}$ and $\mathbf{P}^-$ have full rank and thus the inverse in (57) exists. Combining (52), (53), (57), and we have derived the formulas for computing our corrected observations and state parameters

$$\mathbf{x}^+ = \hat{\mathbf{x}} + \mathbf{R}\mathbf{B}^T\mathbf{S}(\hat{\mathbf{x}}^* - \mathbf{A}\mathbf{z}^+) \tag{58}$$

$$\mathbf{z}^+ = \mathbf{z}^- + \mathbf{K}(\hat{\mathbf{x}}^* - \mathbf{A}\mathbf{z}^+) = \mathbf{z}^- - \mathbf{K}\mathbf{h}(\mathbf{z}^-, \hat{\mathbf{x}}) \tag{59}$$

$$\mathbf{K} = \mathbf{P}^-\mathbf{A}^T\mathbf{S} \tag{60}$$

$$\mathbf{S} = (\mathbf{A}\mathbf{P}^-\mathbf{A}^T + \mathbf{R}^*)^{-1}. \tag{61}$$

From (59), it is straightforward to determine the covariance matrix $\mathbf{P}^+$ of the updated state estimate

$$\mathbf{P}^+ = (\mathbf{I} - \mathbf{K}\mathbf{A})\mathbf{P}^-(\mathbf{I} - \mathbf{K}\mathbf{A})^T + \mathbf{K}\mathbf{R}^*\mathbf{K}^T. \tag{62}$$

Please note that (59)–(62) are equivalent to the standard Kalman filter innovation equations for a transformed observation model with measurement vector $\hat{\mathbf{x}}^*$ and correspond to the minimization of a normalized Sampson error [55]. Similar results (yet with different derivation) have been presented in [56], [57]. The main difference to our algorithm, however, is the correction of the observation vector in (58). Our best estimate $\mathbf{x}^+$ of the ideal image positions can be considered a *nuisance state vector*: $\mathbf{x}^+$ is irrelevant for subsequent stereo rectification and 3-D reconstruction. However, both $\mathbf{z}^+$ and $\mathbf{x}^+$ define a new operation point for subsequent relinearization (see next section). Sparse matrix operations can be used to implement (58)–(62) efficiently.

Several possibilities exist to make this optimization algorithm robust against gross outliers in the input data: First, one could check the (cumulated) χ^2-distribution of the residuals associated with each measurement $\hat{\mathbf{x}}_i$ as proposed in [58]. Thus, identified outliers could be excluded from the relinearization and further tracking process. Second, random sampling as described in [50] could be employed in the filter innovation stage: subsets of all given input features are randomly selected and used in (59)–(62). The best subset is the one with the least median residual error of all correspondence features. Subsequently, a state update is computed based on all features that are consistent with this best subset. The second method was used for the examples in Section V.

Relinearization: The iterated extended Kalman filter (IEKF) is a common approach to improve the *a posteriori* state estimates in case of nonlinear observation models. The IEKF uti-

lizes $\mathbf{z}^+$ of (59) as a new operation point for an additional linearization of the model equation and iteratively refines the parameter state vector. However, since our cost function is minimized with respect to both optimal state parameters and optimal observations, the nonlinear constraint (46) should be relinearized about $(\mathbf{z}^+, \mathbf{x}^+)$ instead of $(\mathbf{z}^+, \hat{\mathbf{x}})$.

The IEKF for implicit measurement constraints can be formulated directly following, e.g., [54]. The resulting filter equations are stated here for completeness: The nonlinear constraint equation $\mathbf{h}$ is first linearized about our best available guess for the state vector and the ideal measurements, i.e., $(\check{\mathbf{z}}_0, \check{\mathbf{x}}_0) = (\mathbf{z}^-, \hat{\mathbf{x}})$. Then, for $l = 0 \ldots L-1$, where L denotes the number of iterations in each time step, compute

$$\mathbf{A}_l = \partial \mathbf{h} / \partial \mathbf{z}|_{\check{\mathbf{x}}_l, \check{\mathbf{z}}_l}, \quad \mathbf{B}_l = \partial \mathbf{h} / \partial \mathbf{x}|_{\check{\mathbf{x}}_l, \check{\mathbf{z}}_l} \tag{62}$$

$$\mathbf{K}_l = \mathbf{P}^- \mathbf{A}_l^{\mathrm{T}} \left[\mathbf{A}_l \mathbf{P}^- \mathbf{A}_l^{\mathrm{T}} + \mathbf{B}_l \mathbf{R} \mathbf{B}_l^{\mathrm{T}} \right]^{-1} \tag{64}$$

$$\mathbf{r}_l = \mathbf{h}(\check{\mathbf{x}}_l, \check{\mathbf{z}}_l) + \mathbf{B}_l \cdot (\hat{\mathbf{x}} - \check{\mathbf{x}}_l) + \mathbf{A}_l \cdot (\mathbf{z}^- - \check{\mathbf{z}}_l) \tag{65}$$

$$\check{\mathbf{z}}_{l+1} = \mathbf{z}^- - \mathbf{K}_l \mathbf{r}_l \tag{66}$$

$$\check{\mathbf{x}}_{l+1} = \hat{\mathbf{x}} - \mathbf{R} \mathbf{B}_l^{\mathrm{T}} \left[\mathbf{B}_l \mathbf{R} \mathbf{B}_l^{\mathrm{T}} \right]^{-1} \mathbf{r}_l. \tag{67}$$

The final *a posteriori* state vector and its covariance matrix are given by

$$\mathbf{z}^+ = \check{\mathbf{z}}_L, \quad \mathbf{P}^+ = [\mathbf{I} - \mathbf{K}_{L-1} \mathbf{A}_{L-1}] \mathbf{P}^-. \tag{68}$$

Equations (63)–(67) may be iterated until the difference between two refined state estimates falls below a predefined threshold or a fixed number of iterations is reached. Based on our experiments in [59], we found that more than 60% of the improvement by iterated innovation steps is already obtained after one additional linearization. After the fourth linearization, the accuracy of the estimation results does not change significantly. We have thus chosen $L = 3$ for the experiments presented in this paper.

REFERENCES

[1] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3-D machine vision metrology using off-the-shelf tv cameras and lenses," *IEEE J. Robot. Autom.*, vol. RA-3, no. 4, pp. 323–344, Aug. 1987.

[2] E. Dickmanns, "The development of machine vision for road vehicles in the last decade," in *Proc. IEEE Intelligent Vehicle Symp.*, Versailles, France, 2002, pp. 268–281.

[3] S. K. Gehrig, "Large-field-of-view stereo for automotive applications," in *OmniVis*, Beijing, China, 2005.

[4] M. Bjorkman and J. Eklundh, "Real-time epipolar geometry estimation of binocular stereo heads," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 425–432, Mar. 2002.

[5] D. W. Murray, F. Du, P. F. McLauchlan, I. D. Reid, P. M. Sharkey, and J. M. Brady, "Design of stereo heads," in *Active Vision*, A. Blake and A. Yuille, Eds. Cambridge, MA: MIT, 1992, pp. 155–174.

[6] H. Schmid, "Eine allgemeine analytische Lösung für die Aufgabe der Photogrammetrie," *Bildmessung und Luftbildwesen (BuL); Zeitschr. für Photogrammetrie u. Fernerkundung*, vol. 2: 1959/1-12, pp. 103–113, 1958.

[7] S. Granshaw, "Bundle adjustment methods in engineering photogrammetry," *Photogramm. Record: Int. J. Photogramm.*, vol. 10, no. 56, pp. 181–207, 1980.

[8] K. Kraus, *Photogrammetrie*. Bonn, Germany: Dümmler Verlag, 1986.

[9] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle adjustment—A modern synthesis," in *Vision Algorithms: Theory and Practice*, ser. Lecture Notes Comput. Sci., B. Triggs, A. Zisserman, and R. Szeliski, Eds. New York: Springer-Verlag, 2000, vol. 1883, pp. 298–372.

[10] Z. Zhang, "Determining the epipolar geometry and its uncertainty: A review," *Int. J. Comput. Vis.*, vol. 27, no. 2, pp. 161–195, 1998.

[11] P. Torr and A. Zisserman, "Robust parameterization and computation of the trifocal tensor," *Image Vis. Comput.*, vol. 15, pp. 591–605, 1997.

[12] S. Maybank and O. Faugeras, "A theory of self calibration of a moving camera," *Int. J. Comput. Vis.*, vol. 8, pp. 123–152, 1992.

[13] O. D. Faugeras, Q.-T. Luong, and S. J. Maybank, "Camera self-calibration: Theory and experiments," in *Proc. Eur. Conf. Computer Vision*, 1992, pp. 321–334.

[14] Q. Luong and O. Faugeras, "Camera calibration, scene motion, and structure recovery from point correspondences and fundamental matrices," *Int. J. Comput. Vis.*, vol. 22, no. 3, pp. 261–289, 1997.

[15] P. Sturm, "A case against kruppa's equations for camera self-calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1199–1204, Oct. 2000.

[16] B. Triggs, "Autocalibration and the absolute quadric," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 609–614.

[17] M. Pollefeys, "Self-Calibration and Metric 3-D Reconstruction From Uncalibrated Image Sequences," Ph.D. dissertation, Katholieke Univ. Leuven, Belgium, 1999.

[18] A. Zisserman, P. Beardsley, and I. Reid, "Metric calibration of a stereo rig," in *Proc. IEEE Workshop on Representations of Visual Scenes*, Boston, MA, 1995, pp. 93–100.

[19] R. Horaud, G. Csurka, and D. Demirdijian, "Stereo calibration from rigid motions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1446–1452, Dec. 2000.

[20] P. Sturm, "Critical motion sequences for the self-calibration of cameras and stereo systems with variable focal length," *Image Vis. Comput.*, vol. 20, pp. 415–426, 2002.

[21] Z. Zhang, Q.-T. Luong, and O. Faugeras, "Motion of an uncalibrated stereo rig: Self-calibration and metric reconstruction," *IEEE Trans. Robot. Autom.*, vol. 12, pp. 103–113, 1996.

[22] Q. Luong and O. Faugeras, "Self-calibration of a moving camera from point correspondences and fundamental matrices," *Int. J. Comput. Vis.*, vol. 22, no. 3, pp. 261–289, Mar. 1997.

[23] G. Qian and R. Chellappa, "Structure from motion using sequential monte carlo methods," in *Proc. Int. Conf. Computer Vision*, 2001, pp. 614–621.

[24] G. Qian and R. Chellappa, "Bayesian self-calibration of a moving camera," in *Proc. Eur. Conf. Computer Vision*, 2002, pp. 277–293.

[25] A. Shashua, "Algebraic functions for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, pp. 779–789, Aug. 1995.

[26] A. Zisserman and S. Maybank, "A case against epipolar geometry," in *Applications of Invariance in Computer Vision LNCS 825*. New York: Springer-Verlag, 1994.

[27] S. Abraham, "Kamera-Kalibrierung und Metrische Auswertung Monokularer Bildfolgen," Ph.D. dissertation, Univ. Bonn, Germany, 2000.

[28] P. F. McLauchlan and D. W. Murray, "Active camera calibration for a head-eye platform using the variable state-dimension filter," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 1, pp. 15–22, Jan. 1996.

[29] N. Pettersson and L. Petersson, "Online stereo calibration using FPGAs," presented at the IEEE Intelligent Vehicles Symp., Las Vegas, NV, Jun. 2005.

[30] T. Dang and C. Hoffmann, "Tracking camera parameters of an active stereo rig," in *28th Annu. Symp. German Association for Pattern Recognition*, Berlin, Germany, Sep. 12–14, 2006.

[31] A. Azarbayejani and A. Pentland, "Recursive estimation of motion, structure, and focal length," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, pp. 562–575, 1995.

[32] S. Soatto and P. Perona, "Reducing structure from motion: A general framework for dynamic vision part 1: Modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 9, pp. 933–942, Sep. 1998.

[33] S. Soatto and P. Perona, "Reducing structure from motion: A general framework for dynamic vision, part 2: Implementation and experimental assessment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 9, pp. 943–960, Sep. 1998.

[34] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. New York: Cambridge Univ. Press, 2002.

[35] R. I. Hartley and P. Sturm, "Triangulation," *Comput. Vis. Image Understand.*, vol. 68, no. 2, pp. 146–157, 1997.

[36] K. Kanatani, *Statistical Optimization for Geometric Computation: Theory and Practice*, ser. Machine Intelligence and Pattern Recognition. New York: Elsevier, 1996, vol. 18.
[37] Y. Xiong and L. Matthies, "Error analysis of a real-time stereo system," in *IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 1097–1093.
[38] S. Das and N. Ahuja, "Performance analysis of stereo, vergence, and focus as depth cues for active vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 12, pp. 1213–1219, Dec. 1995.
[39] M. Bansal, A. Jain, T. Camus, and A. Das, "Towards a practical stereo vision sensor," in *IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops*, 2005, p. 63ff.
[40] D. Scharstein, *View Synthesis Using Stereo Vision*, ser. Lecture notes in computer science. New York: Springer, 1999, vol. 1583.
[41] W. Förstner, "On weighting and choosing constraints for optimally reconstructing the geometry of image triplets," in *Proc. ECCV*, 2000, vol. 2, pp. 669–684.
[42] H. von Sanden, "Die Bestimmung der Kernpunkte der Photogrammetrie," Ph.D. dissertation, Univ. Göttingen, Germany, 1908.
[43] H. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, pp. 133–135, 1981.
[44] Q.-T. Luong and O. D. Faugeras, "The fundamental matrix: Theory, algorithms, and stability analysis," *Int. J. Comput. Vis.*, vol. 17, no. 1, pp. 43–75, 1996.
[45] R. Hartley, "A linear method for reconstruction from lines and points," in *Proc. Int. Conf. Computer Vision*, 1995, pp. 885–887.
[46] R. I. Hartley, "Lines and points in three views and the trifocal tensor," *Int. J. Comput. Vis.*, vol. 22, no. 2, pp. 125–140, 1997.
[47] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[48] C. Tomasi and J. Shi, "Good features to track," in *Proc. IEEE Computer Vision and Pattern Recognition Conf.*, 1994, pp. 593–600.
[49] C. Stiller, S. Kammel, J. Horn, and T. Dang, "The computation of motion," in *Digital Image Processing: Compression and Analysis*. Boca Raton, FL: CRC, Sep. 2004, ch. 3, pp. 73–108.
[50] T. Dang and C. Hoffmann, "Stereo calibration in vehicles," in *Proc. IEEE Intelligent Vehicles Symp.*, Parma, Italy, Jun. 2004, pp. 268–273.
[51] A. Fusiello, E. Trucco, and A. Verri, "A compact algorithm for rectification of stereo pairs," *Mach. Vis. Appl.*, vol. 12, no. 1, pp. 16–22, 2000.
[52] H. Hirschmüller, P. R. Innocent, and J. M. Garibaldi, "Real-time correlation-based stereo vision with reduced border errors," *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 229–246, 2002.
[53] A. Gelb, *Applied Optimal Estimation*. Cambridge, MA: MIT Press, 1994.
[54] D. Simon, *Optimal State Estimation: Kalman, H-Infinity, and Nonlinear Approaches*. New York: Wiley, 2006.
[55] P. Sampson, "Fitting conic sections to very scattered data: An iterative refinement of the bookstein algorithm," *Comput. Graph. Image Process.*, vol. 18, pp. 97–108, 1982.
[56] Z. Zhang and O. Faugeras, *3-D Dynamic Scene Analysis*. New York: Springer, 1992, vol. 27.
[57] S. Soatto, R. Frezza, and P. Perona, "Motion estimation via dynamic vision," *IEEE Trans. Autom. Control*, vol. 4, no. 3, pp. 393–413, Mar. 1996.
[58] T. Dang, C. Hoffmann, and C. Stiller, "Fusing optical flow and stereo disparity for object tracking," in *Proc. IEEE Int. Conf. Intelligent Transportation Systems*, Singapore, 2002, pp. 112–117.
[59] T. Dang, "Kontinuierliche Selbstkalibrierung von Stereokameras," Ph.D. dissertation, Univ. Karlsruhe (TH), Germany, 2007.

Thao Dang (M'04) studied electrical engineering at the University of Karlsruhe, Germany, the Massachusetts Institute of Technology, Cambridge, and the University of Massachusetts, Dartmouth. He received the Dr.-Ing. degree (Ph.D.) with distinction from the University of Karlsruhe, Germany, in 2007.

His research interests include stereo vision, 3-D reconstruction, camera self-calibration, and driver assistance systems. Since September 2007, he has been a research engineer at Daimler AG, Germany.

Christian Hoffmann received the Diploma in mechanical engineering and the Ph.D. degree from the University of Karlsruhe, Germany, in 2001 and 2007, respectively.

Until 2006, he researched vision sensor technology for driver assistance systems at the Institute of Measurement and Control Engineering, University of Karlsruhe. Since November 2006, he has been working at Heidelberger Druckmaschinen AG, Germany.

Christoph Stiller (S'93–M'95–SM'99) studied electrical engineering at the Universities in Aachen, Germany, and Trondheim, Norway. He received the Dr.-Ing. degree (Ph.D.) with distinction from Aachen University in 1994.

He worked in the Research Department, INRS-Telecommunications, Montreal, QC, Canada, and in development for Robert Bosch GmbH, Hildesheim, Germany. In 2001, he became a chaired Professor and Head of the Institute for Measurement and Control Engineering, Karlsruhe University, Germany. His present interests cover cognition of mobile systems, computer vision, and real-time applications thereof. He is the author or coauthor of more than 100 publications and patents in these fields.

Dr. Stiller is Vice President Member Activities of the IEEE Intelligent Transportation Systems Society. He served as Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING (1999–2003) and, since 2004, he has served as an Associate Editor for the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS. In January 2009, he was nominated Editor-in-Chief of the *IEEE Intelligent Transportation Systems Magazine*. He is the Speaker of the Transregional Collaborative Research Center "Cognitive Automobiles" of the German Research Foundation.