Extrinsic Calibration of a Fisheye Multi-Camera Setup Using Overlapping Fields of View

Moritz Knorr¹, Jose Esparza², Wolfgang Niehsen¹, Senior Member, IEEE, and Christoph Stiller³, Senior Member, IEEE

Abstract—It is well known that the robustness of many computer vision algorithms can be improved by employing large field of view cameras, such as omnidirectional cameras. To avoid obstructions in the field of view, such cameras need to be mounted in an exposed position. Alternatively, a multi-camera setup can be used. However, this requires the extrinsic calibration to be known. In the present work, we propose a method to calibrate a fisheye multi-camera rig, mounted on a mobile platform. The method only relies on feature correspondences from pairwise overlapping fields of view of adjacent cameras. In contrast to existing approaches, motion estimation or specific motion patterns are not required. To compensate for the large extent of multi-camera setups and corresponding viewpoint variations, as well as geometrical distortions caused by fisheye lenses, captured images are mapped into virtual camera views such that corresponding image regions coincide. To do this, the scene geometry is approximated by the ground plane in close proximity and by infinitely far away objects elsewhere. As a result, low complexity feature detectors and matchers can be employed. The approach is evaluated using a setup of four rigidly coupled and synchronized wide angle fisheye cameras that were attached to four sides of a mobile platform. The cameras have pairwise overlapping fields of view and baselines between 2.25 and 3 meters.

I. INTRODUCTION

Estimating the motion of a camera with respect to a static scene is a classical problem in computer vision. Cameras with a large field of view have proven to work significantly better than narrow field of view cameras in this regard [5]. Catadioptric cameras, as an example, capture 360 degrees in the horizontal plane, and analysis of the respective motion field is particularly simple [16]. However, a shortcoming of these systems is that they need to be mounted exposed with an unobstructed field of view. Hence, applying them in practice is rather difficult. To circumvent this shortcoming, models have been developed that allow treating many cameras in a unified fashion, thus eliminating the need for a single device ([7],[18]). However, accurate extrinsic and intrinsic calibration is a basic precondition. While cameras can be calibrated intrinsically one at a time, and beforehand, extrinsic calibration requires the cameras to be attached to the rig. The large extent of the rig as well as geometrical distortions render extrinsic calibration troublesome. Calibration from captured image sequences without the requirement for additional hardware or specific conditions is therefore desirable. In this paper we present an approach for estimating the relative pose transformations between cameras using only feature point correspondences from pairwise overlapping fields of view of adjacent cameras. In contrast to existing approaches, motion estimation is not required. In fact, we only require the camera rig to be moving to account for sparse feature matches. Thus, linear motion does not disrupt our method.

Establishing feature point correspondences is challenging, as significant distortions caused by different viewpoints and nonlinear projections are to be expected. To be able to cope with this problem, we use a priori estimates of pairwise relative poses with respect to the ground plane to map captured images into virtual camera views such that corresponding image regions coincide. To this end, we assume the scene to be a combination of the ground plane as well as distant objects. A priori estimates are not required to be consistent across pairs of adjacent cameras and have to be known up to scale only. An outline of a simple method to determine initial estimates is given in the appendix. Feature detection and matching can then be carried out using low complexity algorithms ([19],[3]). To account for sparse feature matches we accumulate data over time and start calibration subsequently.

Once corresponding feature points have been established, the extrinsic calibration is estimated in three steps. For each pair of adjacent cameras with an overlapping field of view, relative pose transformations are derived from a priori estimates and refined independently of one another. Then, a common reference frame is established, and baselines are adjusted by minimizing errors in all cameras simultaneously, while imposing one additional distance constraint to account for global scale ambiguity. Finally, all parameters are refined simultaneously. The detailed calibration method is described in Section IV.

To evaluate the approach we use a setup of four synchronized, rigidly coupled wide angle fisheye cameras, mounted on a mobile platform. Distances between adjacent cameras are up to three meters. To the best of our knowledge, no work regarding the calibrating of an equivalent multi-camera rig using only feature point correspondences from overlapping fields of view has been presented so far. The evaluation of our approach is presented in Section V. We conclude with an outlook on future work in Section VI. Related work is discussed in the following section.
II. RELATED WORK

Numerous approaches to extrinsic multi-camera calibration have been presented hitherto. These approaches can be roughly divided into motion-based approaches that do not require overlapping fields of view, approaches that match between frame histories or built maps, and approaches which rely on robust features or prewarped images to establish image-to-image matches. Motion-based approaches estimate rigid relative poses between cameras by solving the hand-eye calibration problem. Methods straightforward in this regard have been presented by Esquivel et al. [6] and Lebraly et al. [14]. Both require the multi-camera rig to rotate about at least two different axes. This requirement is often not met for mobile platforms moving on planar ground. Brookshire and Teller [1] present a more in-depth analysis of singular motion patterns using statistical measures. To circumvent the need for general, unconstrained 3D motion, Pagel and Willersinn [17] as well as Knorr et al. [12] use the ground plane as a global reference object to recover parameters not observable otherwise.

Approaches that do not require cameras to observe the same scene at the same point in time, but perform matching between frame histories have been considered in the work of Lamprech et al. [13] and Heng et al. [11]. Similarly, Carrera et al. [4] fuse maps that have been built using simultaneous localization and mapping. These approaches either employ accurate motion estimation or restrict the motion to be linear.

Matching between current views is the focus in this paper and has been investigated thoroughly for pinhole modeled cameras with a narrow field of view. Hartley [9] presents an algorithm which establishes epipolar rectification with minimal image distortions using a given set of image-to-image matches. Prewarping to restore similarity in corresponding image regions with respect to known slanted planes and camera poses has been considered in the work of Burt et al. [2] for small baselines. The approach presented in this paper adopts this concept and extends it to fisheye cameras and large baselines.

Direct matching between images captured by adjacent cameras is challenging as nonlinear projections and different viewpoints lead to significant distortions in corresponding image regions. In the work of Yu and Morel [20] an extension to the scale-invariant feature transform is presented, making it invariant to affine transformations. Their results are compared to several robust state-of-the-art features. However, using these features we were not able to establish matches in image regions, which correspond to close proximity scene structures.

III. ESTABLISHING IMAGE CORRESPONDENCES

Establishing feature correspondences between images captured by adjacent cameras is fundamental to our calibration approach. Large baselines, disjoint camera orientations, and geometric distortions caused by the fisheye lenses render direct matching challenging. Using robust features, we were able to match sparse features corresponding to distant objects. However, feature matches corresponding to objects in close proximity significantly improve calibration accuracy, as will be shown in Section V.

To establish image correspondences in close proximity as well as in the distance, we map captured images into virtual views in which feature detection and matching is carried out. The virtual views are chosen such that the corresponding image regions look similar. To this end, we assume the scene to be mostly planar in close proximity and cluttered in the distance. Hence, we employ ground plane induced and infinite homographies for mapping. The homographies are constructed from pose estimates given a priori. A simple method for generating a priori estimates is outlined in the appendix in Section VII.

In the following, we consider the situation depicted in Figure 1. For better understanding, the following derivations refer to a single pair of adjacent cameras only. The two camera coordinate systems (CCS$_1$ and CCS$_2$) are related to a reference coordinate system in the ground plane by the $4 \times 4$ homogeneous transformation matrices [8]

\[ T_c = \begin{bmatrix} R_c & C_c \\ 0^T & 1 \end{bmatrix}, \]  

where $R_c$ is an orthonormal rotation matrix, $C_c$ is a displacement vector, and $c \in \{1, 2\}$. The plane normal is aligned with the $z$-axis in the reference coordinate system. Furthermore, we assume the displacement vectors to be normalized such that $C_1 = (c_{1x}, c_{1y}, 1)^T$, and the third component of $C_2$ is equal to $h_{ref}$, i.e. the relative height ratio. Thus metric a priori estimates are not required. Additionally, reference coordinate systems do not have to be consistent across adjacent camera pairs. In other words, the reference coordinate systems are local and only apply to the respective pair of adjacent cameras. The mapping functions relating real and virtual camera views are introduced in the following.

1) Close proximity: The captured images are mapped into the view of a virtual camera. As neither camera view can be preferred over the other, we decided to install an inter-

---

**Fig. 1:** Geometric relationship between two adjacent cameras and the reference coordinate system in the ground plane. The $z$-axis of the reference coordinate system is aligned with the ground plane normal. All lengths have been rescaled such that the height of the first camera above ground is 1.
mediate virtual camera (see Figure 2, denoted CCS_g). The transformation from the virtual camera coordinate system into the reference coordinate system is \( T_g \), analogically to equation (1). The displacement vector is \( C_g = (C_1 + C_2)/2 \), and \( R_g \) is chosen such that \( R_g e_z \simeq R_1 e_z + R_2 e_z \), where “\( \simeq \)” denotes equality up to scale, and \( e_z = (0, 0, 1)^T \). The x- and y-axis can then be chosen such that \( R_g \) is a proper rotation matrix, leaving one degree of freedom. The corresponding homography mapping functions induced by the ground plane are then given by

\[
\Psi_g = R_g^T R_g - R_g^T (C_g - C_c)(e_z R_g)/h_g
\]

for both cameras, where \( h_g = (h_{rel} + 1)/2 \). It should be noted that the homography relates homogeneous 2D points of calibrated cameras. Thus the inverse world to image projection has to be applied

\[
x_i' = \kappa_g^{-1}(x_i'),
\]

where \( x_i' \) is the \( i \)th image point in the virtual view, \( x_i' \) is the corresponding unit direction vector, and \( \kappa_g(\cdot) \) is the world to image projection function. The resulting mapping functions \( \Psi_g(\cdot) \) from the virtual into the real camera views are then given by

\[
x_c' = \Psi_g(x_i') = \kappa_c(\Psi_g R_c x_i').
\]

2) Distant objects: For distant objects we apply an infinite homography, i.e. the distance from camera to plane is infinite. Hence, the second term in equation (2) vanishes, yielding

\[
e' = R_g^T R_g.
\]

Similar to equation (4), we derive the resulting mapping functions \( \Psi_g \). As the translational parts vanish, the mapping can be interpreted as individually rotated cameras (denoted \( CCS_{\infty,c} \) in Figure 2).

A. FEATURE DETECTION AND MATCHING

In order to establish corresponding feature points in the virtual views, we use the state of the art feature detector by Rosten and Drummond [19] and feature descriptor by Calonder et al. [3]. Descriptor similarity is evaluated efficiently using the Hamming distance. The algorithm provides image correspondences at pixel accuracy at best. However, we found that the matched feature points often do not have minimal Hamming distance in the local neighborhood. Thus we additionally compare feature signatures in the local \( 5 \times 5 \) neighborhood of the matched feature points to find the best match.

Mapping captured images into a virtual camera view is a non-linear transformation. The underlying functions depend on the intrinsic parameters of the camera as well as on the extrinsic configuration. As a result, the spatial resolution in mapped images is locally varying, directional, and depends on the source camera. It is obvious that compared image patches should have the same spatial resolution, as higher frequencies, being present in one image only, will act similar to correlated noise. To this end, we employ locally varying interpolation and smoothing kernels to adjust spatial resolutions and to mitigate aliasing (see e.g. [10]).

IV. EXTRINSIC CALIBRATION

Once corresponding feature points are established, the task at hand is to estimate transformations relating all camera coordinate systems. The estimation is composed of three steps, illustrated schematically in Figure 3. First, corresponding feature points are used to refine the a priori estimates and eliminate gross outliers for each pair of adjacent cameras individually. As the baselines cannot be recovered from pairwise refinement, the camera coordinate systems are transformed into a common coordinate frame in the second step.

Fig. 3: Schematic illustration of the three calibration steps. First, pairwise relative poses are refined (a). Then a common reference coordinate frame is established and baselines are adjusted (b). Additionally, a distance constraint is imposed to fix the scale of the setup \( d_{fixed} \). Finally, estimates are refined simultaneously (c).
Baselines are then recovered by minimizing errors in all views simultaneously while imposing one additional distance constraint. Finally, all parameters, including orientations and displacement vectors, are refined simultaneously.

A. ERROR METRIC

For parameter estimation a fixed-scale M-estimator with robust Cauchy cost function is used ([8]). The employed error metric is based on the shortest distance to the epipolar plane, projected into the fisheye camera views. As the epipolar planes are projected to curves, we use a support point based approximation which is not presented in more detail here. Previous to optimization, corresponding feature positions are mapped back into the original views using the mapping functions defined in the previous section.

B. PAIRWISE RELATIVE POSE REFINEMENT

Given the a priori pose estimates of pairs of adjacent cameras with respect to local reference coordinate systems, we compute the direct camera-to-camera relative poses, e.g. for camera 1 to 2 this is $T_2^1 = T_2^1 T_1$. The pose estimates are then refined using the corresponding feature points. This is depicted in Figure 3a. The resulting transformations are only known up to scale, i.e. only five out of six degrees of freedom can be recovered. Without loss of generality, for now, we assume the baselines $b_c$, i.e. the distances between adjacent cameras, to have unit length.

C. COMMON REFERENCE FRAME

As the baselines cannot be recovered from pairwise refinement, concatenating relative pose transformations would result in an inconsistent setup. To be able to adjust the baselines, a common reference coordinate frame is established first. In order to keep the parameterization minimal and to avoid ambiguities during estimation, one of the camera coordinate systems is chosen as the reference coordinate frame. The corresponding camera is called reference camera. We propose choosing the reference camera based on the number of adjacent cameras with overlapping fields of view and the number of corresponding feature points. All cameras are related to the reference coordinate system by relative pose transformations (see Figure 3b)

$$\Delta T_c = \begin{bmatrix} \Delta R_c & b_c \Delta C_c \\ 0^T & 1 \end{bmatrix}.$$ (6)

For adjacent cameras, these pose transformations, up to scale, have already been defined above. For other cameras the transformations can be determined by concatenating relative pose transformations. The baselines are then adjusted using the fixed-scale M-estimator and the parameterization described in the following. Rotation matrices are parameterized using angle-axis representation [8], $R(\omega_{x,c}, \omega_{y,c}, \omega_{z,c})$. Displacement vectors are parameterized using spherical coordinates with angles $\theta_c$, $\phi_c$, and baseline $b_c$, $\Delta C_c = (\sin(\theta_c)\cos(\phi_c), \sin(\theta_c)\sin(\phi_c), \cos(\theta_c))^T$. Explicit separation of the baseline is required to impose additional constraints. The transformation in equation (6) can then be represented by a parameter vector $\xi_c = (\omega_{x,c}, \omega_{y,c}, \omega_{z,c}, \theta_c, \phi_c, b_c)^T$. To improve the condition of the estimation problem, pre-rotation matrices are derived from the initial estimates and applied to the relative orientations and the displacement vectors. The resulting initial parameter vectors are $\xi_c^0 = (0, 0, 0, 0, 0, 0, b_c)^T$. The parameter vector of the whole camera rig is then composed of the $N - 1$ relative pose parameter vectors, where $N$ is the total number of cameras. As the absolute scale cannot be recovered, the distance between the reference camera and another dedicated camera is held fixed (as indicated by $d_{\text{fixed}}$ in Figure 3b), thus reducing the number of parameters to be estimated by one. For experimental evaluation, the fixed value is set to the corresponding known ground truth value, such that results can be compared metrically. Note that only the baselines $b_c$ are adjusted in this step.

D. REFINEMENT

The final step of our calibration process is refinement. All parameters, except for the fixed distance, are refined using the fixed-scale M-estimator with robust cost function and the parameterization introduced above. This is illustrated in Figure 3c. In order to evaluate whether estimation using feature matches corresponding to distant or close-by objects is sufficient, experiments incorporating either subset have been conducted additionally (see Section V).

E. SINGULAR CONFIGURATIONS

There exist several camera configurations in which a single additional constraint as shown in Section IV-C is not sufficient, e.g. if all camera centers are aligned or if camera centers are located at the vertices of a parallelogram. As only epipolar constraints are enforced, relative distances between camera centers have to be recovered by enforcing global consistency. However, in the first example, one could shift cameras along the displacement vector without introducing additional errors. To detect singular configurations, we recommend evaluating the covariance matrix during refinement.

V. EXPERIMENTAL EVALUATION

In order to evaluate our approach we use a setup of four synchronized, rigidly coupled 1.2 megapixel wide angle fisheye cameras ($N = 4$), which were attached to four sides of a mobile platform. Overlapping fields of view only exist between adjacent cameras. The baselines range from 2.25 to 3 meters, whereas the respective distances to the ground plane are 0.75 to 1.2 meters. The fixed distance between the dedicated cameras is set to the ground truth value of approximately 5 meters.

Ten sequences with lengths between 10 and 75 seconds have been captured at 30 frames per second. Due to the high frame rate, the observed scene changes little in successive frames. Thus, sequences were sub-sampled to 7.5 frames per second.

1In this example, baseline and fixed distance coincide. In the experimental setup this is not the case.
For the intrinsic camera calibration we use a method similar to [15]. The estimation results are compared to ground truth which has been obtained using special calibration targets as well as additional cameras and bundle adjustment. The error metric used for quantitative evaluation is introduced in the following. Given the ground truth relative poses $\Delta T_c^{GT}$ and the estimated relative poses $\Delta T_c$ with respect to the reference camera, the residual transformations $T^{res}_c = \Delta T_c^{-1} \Delta T_c^{GT}$ are computed. From this, we compute the mean position error and angular error as

$$e_p = \frac{1}{N-1} \sum_{c=1}^{N-1} ||C^{res}_c||_2,$$

(7)

$$e_A = \frac{1}{N-1} \sum_{c=1}^{N-1} \cos^{-1} \left( \frac{\text{tr}(R^{res}_c) - 1}{2} \right).$$

(8)

In order to evaluate the influence of feature points corresponding to distant and close-by objects on the estimation, experiments were repeated using either subset and the combined set, respectively. The results are given in Table I. It can be seen that the combination of feature sets results in superior position estimates. This, however, does not hold for the angular estimates for all sequences, and is subject to further analysis. Calibration based on distant scene points yields poor position and angular estimates. We want to draw attention to the fact that feature points corresponding to distant scene points were located in flat and wide regions in the images. Feature points corresponding to objects in close proximity were located in regions with larger vertical extend. As both regions merely overlap, they complement one another well and should both be considered for calibration.

To be able to assess the attained accuracy, an additional, application-related experiment was conducted. Similar to the work in [12],2 corresponding feature points in successive frames were used to estimate egomotion. The resulting track of concatenated motion estimates using ground truth is given in light blue in Figure 4. In the same way, the orange track was generated using the estimation results of Sequence 01 with combined feature sets (bold in Table I). The experiment was designed to detect errors in loop-closure. Thus, tracks are supposed to overlap at start and end, which was found to be almost the case for the track with ground truth parameters. The offset of the track with estimated parameters is approximately 0.5 meters at start and end.

### VI. CONCLUSION AND OUTLOOK

In this paper we have presented an approach to extrinsic calibration of a multi-camera rig. The approach solely relies on feature point correspondences from pairwise overlapping fields of view of adjacent cameras. In order to establish correspondences, captured images were mapped into virtual camera views in which feature detection and matching was carried out. Attention has been paid to the problem of resampling and local spatial resolution. A three-step approach was used to calibrate the cameras with respect to a dedicated reference camera. Experiments showed that feature points corresponding to distant objects have to be combined with feature points corresponding to close-by objects to obtain higher accuracy. In the future we will concentrate on improving resampling and smoothing for feature detection and matching, as this is a fundamental step in our approach. Furthermore, we plan on incorporating narrow field of view cameras.

### VII. APPENDIX

#### INITIAL RELATIVE POSE ESTIMATES

In this section, we will briefly outline a heuristic but robust approach for obtaining initial pose estimates with respect to local reference coordinate systems for camera pairs. The solution is obtained using frame to frame homographies induced by the ground plane for each camera individually. We assume the camera rig to be moving in a plane and rotating only about the plane normal. The corresponding homography associating frame $k-1$ and frame $k$ for camera $c$ is then

$$H_{c,k} = R(n_c, \beta_{c,k}) - v_{c,k} t(n_c, \alpha_{c,k}) n_c^T,$$

(9)
where \( \mathbf{n} \) is the ground plane normal, \( \beta \) is the angular velocity about the plane normal, \( \mathbf{t}(\mathbf{n}, \alpha) \) is the unit length translation vector in the plane, parameterized by the angle \( \alpha \), and \( \mathbf{v} \) is the relative velocity, i.e. the velocity divided by the camera height. We can estimate the parameter vector \( \mathbf{p}_{c,k} = (\hat{\beta}_{c,k}, \alpha_{c,k}, \hat{\mathbf{v}}_{c,k}, \hat{\mathbf{n}}_{c,k}) \) in each successive pair of frames for each camera, respectively. In order to obtain initial relative pose estimates, we treat the estimated parameter vectors as samples from an unknown density function and divide the set of samples into linear motion and turning, based on angular velocity.

A. ORIENTATION AND RELATIVE HEIGHT

For linear motion, the translation direction \( \alpha_{c,k} \) and hence the unit translation vector \( \mathbf{t} \) do not change. In combination with the plane normal vector, they form two axes of an orthonormal basis. We can interpret this as the relative orientation between the camera and a reference coordinate system if we assume the axes of the reference coordinate system to be aligned with the dominant motion direction and the ground plane normal. We parameterize the orientation matrices using angle-axis representation [8]. Each orientation is then represented as a point in 3-space\(^3\). A mean shift algorithm is used to find the maximum sample density and hence an estimate of the reference-to-camera orientation. The relative heights can be estimated from the ratios of relative velocities. For a pair of adjacent cameras \( l \) and \( m \), the relative height ratios are given by \( \hat{h}_{m,k} = \hat{v}_{l,k}/\hat{v}_{m,k} \). Again, the mean shift algorithm is employed to find the maximum density and thus the relative height estimate \( \hat{h}_{m} \). The height of the first camera is set to \( \hat{h}_{1} = 1 \) (cf. Figure 1).

B. 2D RELATIVE POSITION

As we already obtained the orientation between reference coordinate system and camera, the problem of relative position estimation is transformed into a two dimensional problem in the plane. Using the hand-eye calibration problem formulation we can write the linear estimation problem as

\[
(R^{2D} \left( \frac{\hat{\beta}_{l,k} + \hat{\beta}_{m,k}}{2} \right) - I_{2 \times 2}) \Delta \mathbf{c}_{lm} = \hat{\mathbf{v}}_{l,k} \mathbf{t}_{l,k}^{2D} - \hat{\mathbf{v}}_{m,k} \mathbf{t}_{m,k}^{2D}
\]

where \( l \) and \( m \) are the indices of two adjacent cameras, \( \mathbf{t}_{l,k}^{2D} \) is the 2D translation vector, and \( \Delta \mathbf{c}_{lm} \) is the relative 2D position. As the left side of equation (10) vanishes for linear motion, hypotheses for the relative position can only be determined while turning. Once again a mean shift algorithm is used to find the maximum density among the relative position hypotheses \( \Delta \mathbf{c}_{lm} \). Without loss of generality, we assume the reference coordinate system to be located beneath the first camera \( l \). The relative poses of pairs of cameras with respect to local reference coordinate systems in the ground plane are then given by combining the above results.

\[^{3}\text{Caution has to be paid if } \alpha \approx \pm\pi. \text{ In this case, an additional point with angle } \alpha \mp 2\pi \text{ is inserted.}\]

References