Multi-Camera Structure from Motion with Eye-to-Eye Calibration

Sandro Esquivel and Reinhard Koch

Christian-Albrechts-University, Kiel, Germany

Abstract. Imaging systems consisting of multiple conventional cameras are of increasing interest for computer vision applications such as Structure from Motion (SfM) due to their large combined field of view and high composite image resolution. In this work we present a SfM framework for multi-camera systems w/o overlapping camera views that integrates on-line extrinsic camera calibration, local scene reconstruction, and global optimization based on combining hand-eye calibration methods with standard SfM. For this purpose, we propose a novel method for extrinsic calibration based on rigid motion constraints that uses visual measurements directly instead of motion correspondences. Only a single calibration pattern visible within the view of one camera is needed to provide an accurate reconstruction with absolute scale.

1 Introduction

During the recent years, camera systems with large visual field coverage have proved useful to solve a variety of practical computer vision problems such as surveillance tasks, pose tracking, scene reconstruction, and Augmented Reality. Omnidirectional cameras with a 360° field of view in the horizontal plane are commonly used in robotics for visual odometry and simultaneous localization and mapping, e. g., for advanced driver assistant systems, autonomous vehicle navigation, and urban scenes modeling, while wide-angle fisheye lens cameras are often used for panorama imaging, edificial inspection, and site measuring.

While omnidirectional cameras made up from specific lenses or cameras imaging mirror surfaces are still very common for these tasks, rigs composed of multiple off-the-shelf cameras have gained popularity during the recent years. Major advantages of such devices are often lower costs, flexible configuration, less complex mathematical models and intrinsic calibration, and considerably higher resolution of the virtual composite field of view. In order to maximize the visual field it is beneficial to assemble the individual cameras so that their fields of view have minimal overlap. However, extrinsic camera calibration (i. e., determining the locations and orientations of all cameras within a common reference coordinate frame) is complicated by this setup since conventional calibration methods such as [26] rely on jointly observed patterns or objects with known geometry.

Previous Work. Common approaches for extrinsic multi-camera calibration without overlapping views require very specific calibration objects such as large patterns [15] or planar mirrors [13, 11, 21] to supply global image correspondences.

Finding correspondences between cameras over time during motion of the rig [9] poses difficult matching problems. Also, all these methods can be impractical due to occlusions or large camera offsets. Attempts based on *per-camera* image or pose correspondences only were first proposed in [3] for cameras with coinciding projection centers and in [5] for general setups. In [7], a flexible method for extrinsic camera calibration from rigid motion constraints was described that utilizes simultaneous Structure from Motion (SfM) to estimate camera motion correspondences. This approach – denoted as *eye-to-eye calibration* here – is based on the classical hand-eye calibration problem from the robotics community [25], in particular on extended methods using SfM for camera localization [1]. Since publication, it has been developed further, most notably towards vehicle-based camera systems [20], and improved by global optimization using joint bundle adjustment [14] or including partial rigid motion constraints in the SfM step [6].

Our Contribution. In this paper we will propose a multi-camera SfM pipeline integrating the aforementioned approaches to provide a reconstruction with absolute scale from rigidly coupled cameras without overlapping views with known intrinsics but a priori unknown extrinsic parameters. Only a single calibration pattern visible for the first camera is needed. The eye-to-eye calibration problem is solved with a novel method minimizing image errors instead of motion differences and is further refined via the bundle adjustment approach from [14].

2 Rigidly Coupled Motion Constraints

Each *pose transformation* $\mathbf{T} \in \mathrm{SE}(3)$ is described by a rotation matrix $\mathbf{R} \in \mathrm{SO}(3)$ and translation vector $\boldsymbol{t} \in \mathbb{R}^3$. Rotations with angle α around axis $\boldsymbol{r} \in S^2$ are parametrized by unit quaternions $\mathbf{q} \in S^3$ in the following (see [24], Sec. 2.4):

$$\mathbf{q} = (\boldsymbol{q}, q) = (\sin(\tfrac{\alpha}{2})\boldsymbol{r}, \cos(\tfrac{\alpha}{2})) \quad \text{and} \quad \mathbf{R}_{\mathbf{q}} = (q^2+1)\mathbf{I} + 2q[\boldsymbol{q}]_\times + 2[\boldsymbol{q}]_\times^2 \tag{1}$$

Given $n+1$ rigidly coupled cameras at $m+1$ different positions as illustrated in Fig. 1, the relative coordinate transformations $\mathbf{R}_k^i, \boldsymbol{t}_k^i$ for the i-th camera at the k-th position with respect to the *reference pose* at $k = 0$ ("local" measurements) are given by some pose measuring process. Denoting the *reference camera* by $i = 0$, the *eye-to-eye transformations* $\Delta\mathbf{T}_i, \Delta\lambda_i$ describe the coordinate transfer from the i-th camera to the reference camera for each $i = 0, \ldots, n$. Due to the rigid coupling, for each $k = 0, \ldots, m$ holds:

$$\mathbf{R}_k^0 \Delta\mathbf{R}_i = \Delta\mathbf{R}_i \mathbf{R}_k^i \quad \text{and} \quad \mathbf{R}_k^0 \Delta\boldsymbol{t}_i + \boldsymbol{t}_k^0 = \Delta\lambda_i \Delta\mathbf{R}_i \boldsymbol{t}_k^i + \Delta\boldsymbol{t}_i \tag{2}$$

Each scalar $\Delta\lambda_i > 0$ describes an isometric scaling between the local coordinate frames of the i-th and the reference camera while $\Delta\mathbf{R}_i, \Delta\boldsymbol{t}_i$ describe the pose of camera i within its reference coordinate frame. Note that $\mathbf{R}_0^i = \mathbf{I}, \boldsymbol{t}_0^i = \mathbf{0}$ for all $i = 0, \ldots, n$ and $\Delta\mathbf{R}_0 = \mathbf{I}, \Delta\boldsymbol{t}_0 = \mathbf{0}, \Delta\lambda_0 = 1$ are fixed in (2).

If poses of the reference camera are measured within the world coordinate frame instead, a similar equation is derived:

$$\tilde{\mathbf{R}}_k^0 \Delta\mathbf{R}_i = \Delta\tilde{\mathbf{R}}_i \mathbf{R}_k^i \quad \text{and} \quad \tilde{\mathbf{R}}_k^0 \Delta\boldsymbol{t}_i + \tilde{\boldsymbol{t}}_k^0 = \Delta\lambda_i \Delta\tilde{\mathbf{R}}_i \boldsymbol{t}_k^i + \Delta\tilde{\boldsymbol{t}}_i \tag{3}$$

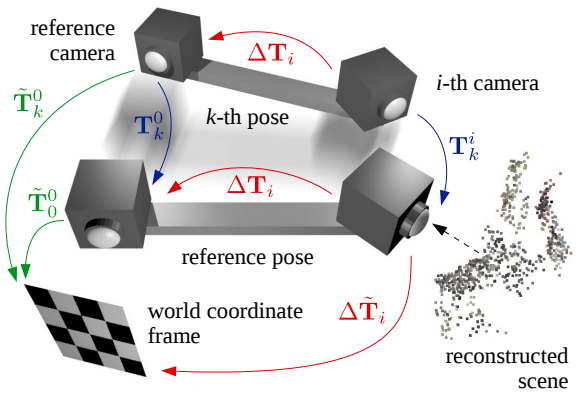

Fig. 1. Overview of coordinate frames and transformations for two rigidly coupled cameras at reference location and k-th location as used in eye-to-eye SfM.

where $\Delta\tilde{\mathbf{T}}_i = \tilde{\mathbf{T}}_0^0 \Delta\mathbf{T}_i$ describes the *eye-to-world transformation* (in accordance to the *hand/eye and world/base calibration* problem from robotics). To distinguish local poses and 3d points from measurements within the world coordinate frame ("global" measurements), we will use a tilde for the latter.

Partial Rigid Motion Constraints. Following from (2), all rigidly coupled motions $\mathbf{R}_k^i, \boldsymbol{t}_k^i$ with non-zero rotation have the same absolute rotation angle α_k^i and amount of translation along the rotation axis $p_k^i = \boldsymbol{r}_k^{iT} \boldsymbol{t}_k^i$ (see [4], Sec. 4.1). Using the latter constraint, the scaling $\Delta\lambda_i$ can be derived for non-planar motion as $\Delta\lambda_i = p_k^0 \,/\, p_k^i$ for any pose with $\mathbf{R} \neq \mathbf{I}$ and $p_k^i \neq 0$. Both constraints can be used to robustify simultaneous SfM for rigidly coupled cameras as described in [6].

Geometric Eye-to-Eye Calibration. Similar to hand-eye calibration where the reference camera is replaced by a robotic gripper providing absolute poses, (2) can be solved for the eye-to-eye transformation parameters from $m \geq 2$ motion correspondences with sufficient rotation and translation and distinct rotation axes. A standard approach is to solve the first part of (2) for $\Delta\mathbf{R}_i$ first, e. g., using the unit quaternion parametrization [7] (solved via SVD):

$$\min_{\Delta\mathbf{q}_i} \sum_{k=0}^{m} \|\mathbf{q}_k^0 \cdot \Delta\mathbf{q}_i - \Delta\mathbf{q}_i \cdot \mathbf{q}_k^i\|^2 \quad \text{s. t. } \|\Delta\mathbf{q}_i\| = 1 \tag{4}$$

Then solve the linear equation system resulting from the second part of (2) for $\Delta\boldsymbol{t}_i, \Delta\lambda_i$ and refine all parameters jointly via nonlinear optimization [23] (using the reduced unit quaternion parametrization from [24] to avoid constraints):

$$\min_{\Delta\theta_i} \sum_{k=0}^{m} d_{\text{rot}}(\mathbf{R}_k^0 \Delta\mathbf{R}_i, \Delta\mathbf{R}_i \mathbf{R}_k^i)^2 + d_{\text{pos}}(\mathbf{R}_k^0 \Delta\boldsymbol{t}_i + \boldsymbol{t}_k^0, \Delta\lambda_i \Delta\mathbf{R}_i \boldsymbol{t}_k^i + \Delta\boldsymbol{t}_i)^2 \tag{5}$$

where $\Delta\theta_i$ are the eye-to-eye transformation parameters describing $\Delta\mathbf{R}_i, \Delta\boldsymbol{t}_i, \Delta\lambda_i$ for the i-th camera, and $d_{\text{rot}}, d_{\text{pos}}$ are appropriately weighted error measures between rotations (e. g., quaternion distance or residual angle measure $d_\angle$) and translations (e. g., Euclidean distance). This approach is denoted as *geometric eye-to-eye calibration* (E2E-GEOM) in the following since the error function (5) describes differences between pose transformations. As pointed out in [23], weighting of the rotational and translational error terms has a crucial impact on the estimation results. The authors advise to use statistical weights derived from the input pose accuracy, accessed for instance via covariance propagation from the prior pose estimation process.

3 Integrating Eye-to-Eye Calibration into SfM

In the following we describe how to integrate eye-to-eye calibration into the classical SfM pipeline and provide an algorithm for incremental eye-to-eye calibration of multi-camera systems based on errors in the image domain, relieving the problem of weighting geometric error terms.

Pose Transfer. Given an estimate for the k-th pose of the reference camera $\mathbf{R}_k^0, \boldsymbol{t}_k^0$ relative to the reference pose, the corresponding pose for the i-th camera within its reference frame is inferred from (2) as:

$$\mathbf{R}_k^i = \Delta\mathbf{R}_i^T \mathbf{R}_k^0 \Delta\mathbf{R}_i \quad \text{and} \quad \boldsymbol{t}_k^i = \Delta\lambda_i^{-1} \Delta\mathbf{R}_i^T \left((\mathbf{R}_k^0 - \mathbf{I}) \Delta\boldsymbol{t}_i + \boldsymbol{t}_k^0 \right) \tag{6}$$

where $\mathbf{R}_0^0 = \mathbf{I}, \boldsymbol{t}_0^0 = \mathbf{0}$ are fixed. The corresponding global pose given the initial global pose $\tilde{\mathbf{R}}_0^0, \tilde{\boldsymbol{t}}_0^0$ of the reference camera is inferred by:

$$\tilde{\mathbf{R}}_k^i = \tilde{\mathbf{R}}_0^0 \mathbf{R}_k^0 \Delta\mathbf{R}_i \quad \text{and} \quad \tilde{\boldsymbol{t}}_k^i = \tilde{\mathbf{R}}_0^0 (\mathbf{R}_k^0 \Delta\boldsymbol{t}_i + \boldsymbol{t}_k^0) + \tilde{\boldsymbol{t}}_0^0 \tag{7}$$

Visual Eye-to-Eye Calibration. Given N_i 3d points for the i-th camera within its local coordinate frame and corresponding projections $\boldsymbol{x}_{k,\ell}^i$ of the ℓ-th 3d point $\boldsymbol{X}_\ell^i$ into the k-th image with known camera functions $\mathcal{K}_i$, the i-th eye-to-eye transformation is obtained by minimizing the reprojection error using the pose transfer function (6):

$$\min_{\Delta\theta_i} \sum_{k=0}^{m} \sum_{\ell=1}^{N_i} V_{k,\ell}^i \, d_i(\boldsymbol{x}_{k,\ell}^i, \mathbf{R}_k^{iT} (\boldsymbol{X}_\ell^i - \boldsymbol{t}_k^i))^2 \tag{8}$$

where $V_{k,\ell}^i \in \{0, 1\}$ describes the visibility of 3d point $\boldsymbol{X}_\ell^i$ in the k-th image. The reprojection error is described by a generic function $d_i : \mathbb{R}^2 \times \mathbb{R}^3 \to \mathbb{R}$ for the i-th camera which is commonly chosen as $d_i(\boldsymbol{x}, \boldsymbol{X}) = \|\boldsymbol{x} - \mathcal{K}_i(\boldsymbol{X})\|$ assuming that the camera function $\mathcal{K}_i$ is known (e. g., from previous intrinsic calibration). For 2d point observations $\boldsymbol{x}$ with non-isometric errors described by covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, the Mahalanobis distance $\|\boldsymbol{x} - \mathcal{K}_i(\boldsymbol{X})\|_{\boldsymbol{\Sigma}_{\boldsymbol{x}}}$ can be used instead. This novel approach will be denoted as *visual eye-to-eye calibration* (E2E-VIS). Note that

the scaling parameter $\Delta\lambda_i$ can be encoded implicitely in (8) by parametrizing the scaled rotation matrix $\Delta\lambda_i^{-1}\Delta\mathbf{R}_i$ used in the prediction function (6) with a non-unit quaternion $\Delta\mathbf{q}_i$, i. e. $\Delta\lambda_i^{-1}\Delta\mathbf{R}_i = \mathbf{R}_{\Delta\mathbf{q}_i}$ with $\Delta\lambda_i^{-1} = \|\Delta\mathbf{q}_i\|^2$ as defined in (1), leading to an unconstrained optimization problem.

Eye-to-Eye Bundle Adjustment. Including 3d points and reference camera poses within the world coordinate frame according to (7) as parameters provides the *eye-to-eye bundle adjustment* (E2E-BA) problem similar to [14]:

$$\min_{\substack{\Delta\theta_i,\tilde{\theta}_0,\ldots,\tilde{\theta}_m \\ \tilde{\chi}_1^i,\ldots,\tilde{\chi}_{N_i}^i}} \sum_{k=0}^{m} \sum_{j\in\{0,i\}} \sum_{\ell=1}^{N_j} V_{k,\ell}^j \, d_j(\boldsymbol{x}_{k,\ell}^j, \tilde{\mathbf{R}}_k^{j\,T}(\tilde{\boldsymbol{X}}_\ell^j - \tilde{\boldsymbol{t}}_k^j))^2 \tag{9}$$

where $\tilde{\theta}_k$ are the k-th global pose parameters for the reference camera and $\tilde{\chi}_\ell^i$ are the parameters of the ℓ-th 3d point $\tilde{\boldsymbol{X}}_\ell^i$ for the i-th camera transformed into the world coordinate frame, initialized by $\tilde{\boldsymbol{X}}_\ell^i = \tilde{\mathbf{R}}_0^0(\Delta\lambda_i\Delta\mathbf{R}_i\boldsymbol{X}_\ell^i + \Delta\boldsymbol{t}_i) + \tilde{\boldsymbol{t}}_0^0$. The scaling parameter $\Delta\lambda_i$ is dropped from $\Delta\theta_i$ since it is encoded by the 3d points. Note that 3d points for the reference camera are already expressed within the world coordinate frame associated with some calibration object here. Gauge freedoms are avoided since the 3d points of the reference camera are fixed. Depending on the given application, the E2E-BA error function can be modified in order to fix either all 3d points (calibration objects used for both cameras) or none (SfM used for both cameras). The first case is equivalent to adding the reference camera poses as parameters to E2E-VIS, in the latter case gauge freedoms must be taken care of (in general by fixing $\mathbf{T}_0^0 = \mathbf{I}$ and $\|\boldsymbol{t}_1^0\| = 1$).

Pairwise E2E-BA as defined in (9) can be extended to cover several coupled cameras at the same time in a straightforward way, leading to large-scale sparse optimization problems. Common sparse bundle adjustment implementations such as `sba` cannot be applied to solve (9) since the Jacobian matrix of the error function has not the distinct block structure needed to compute the Schur complement [17], due to the fact that $\Delta\theta_i$ appears in all residuals for the i-th camera. We use `sparseLM` instead, a sparse implementation of the Levenberg-Marquardt algorithm [16].

Eye-to-Eye Structure from Motion. The proposed algorithm for interactive online eye-to-eye calibration via SfM (E2E-SFM) is outlined as follows. First, camera functions $\mathcal{K}_1, \ldots, \mathcal{K}_n$ are obtained by individual intrinsic camera calibration (e.g., following [26]). A calibration object (e. g., a checkerboard pattern) is placed within viewing range of the reference camera. Images of the calibration pattern are captured with the reference camera during motion of the camera rig, and images for the i-th camera are captured simultaneously (start with $i := 1$):

- Add initial keyframe with poses $\mathbf{T}_0^0 = \mathbf{I}$ and $\mathbf{T}_0^i = \mathbf{I}$.
- Compute global reference pose $\tilde{\mathbf{T}}_0^0$ for reference camera from 2d/3d matches.
- Detect feature points in reference image of the i-th coupled camera.
- For each subsequently captured image:

- Set $k :=$ number of keyframes for each camera.
- Compute current global pose $\tilde{\mathbf{T}}_k^0$ for reference camera from 2d/3d matches.
- Find feature matches from reference to current image of i-th camera.
- If $k = 1$ ($\rightarrow$ *SfM initialization stage*):
  - If $\|\boldsymbol{t}_1^0\| = \|(\tilde{\mathbf{R}}_0^0)^T(\tilde{\boldsymbol{t}}_1^0 - \tilde{\boldsymbol{t}}_0^0)\| > t_{\min}$:
    - Estimate essential matrix $\mathbf{E}_i$ from 2d/2d correspondences and initialize SfM for i-th camera (see [10], Part II).
    - Refine and scale relative pose $\mathbf{T}_1^i$ derived from essential matrix $\mathbf{E}_i$ using partial rigid motion constraints as described in [6].
    - Add keyframe with poses $\mathbf{T}_1^0 = (\tilde{\mathbf{T}}_0^0)^{-1}\tilde{\mathbf{T}}_1^0$ and $\mathbf{T}_1^i$.
- Else ($\rightarrow$ *SfM tracking stage*):
  - Estimate current pose $\mathbf{T}_k^i$ for i-th camera from 2d/3d matches.
  - Refine pose $\mathbf{T}_k^i$ using partial rigid motion constraints [6].
  - Triangulate new 3d points for i-th camera from 2d/2d matches.
  - If $d_{\angle}(\tilde{\mathbf{R}}_{k-1}^0, \tilde{\mathbf{R}}_k^0) > \alpha_{\min}$ and $\angle(\tilde{\boldsymbol{r}}_{k-1}^0, \tilde{\boldsymbol{r}}_k^0) > \beta_{\min}$:
    - Add keyframe with poses $\mathbf{T}_k^0 = (\tilde{\mathbf{T}}_0^0)^{-1}\tilde{\mathbf{T}}_k^0$ and $\mathbf{T}_k^i$.
    - Compute initial eye-to-eye transformation $\Delta\mathbf{T}_i$ from corresponding motions in keyframes via E2E-GEOM.
    - Refine eye-to-eye transformation $\Delta\mathbf{T}_i$ from 2d/3d matches in keyframes via E2E-VIS.
    - Compute E2E-BA with fixed 3d points for the reference camera.
- If $k = k_{\max}$ (or other termination criterion holds):
  - Clear keyframes and start over with $i := i + 1$ unless $i = n$ holds.

The main pipeline is illustrated in Fig. 2. SfM requires some minimal initial translation defined by the threshold $t_{\min}$ (here: $t_{\min} = 25\,\text{cm}$). Keyframes for eye-to-eye calibration are added according to the criteria suggested in [1] for on-line hand-eye calibration, i. e., sufficiently large rotation angle and rotation axis difference w. r. t. the previous keyframe pose using thresholds $\alpha_{\min}$, $\beta_{\min}$ (here: $\alpha_{\min} = 10°$, $\beta_{\min} = 15°$). The termination criterion can be based on the covariance matrix $\boldsymbol{\Sigma}_{\Delta\theta_i}$ of the estimated eye-to-eye transformation parameters $\Delta\theta_i$ resulting from E2E-BA given some accuracy requirement for the solution, or maximal keyframe number $k_{\max}$. Further details on the basic SfM algorithms can be found in [10]. E2E-BA can be computed in a separate thread for efficiency.

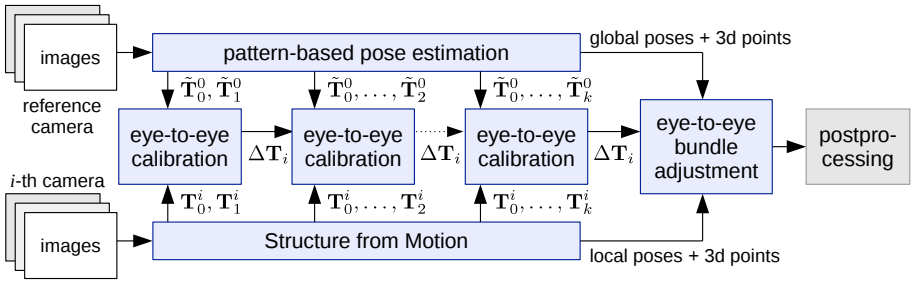

Fig. 2. Overview of eye-to-eye Structure from Motion pipeline.

Post-processing. After all eye-to-eye transformations have been estimated, multi-camera SfM using all cameras jointly as described in [9] or [12] can be applied. The resulting reconstruction and extrinsic parameters can be optionally refined via E2E-BA using all coupled cameras at the same time.

4 Tests and Evaluation

4.1 Evaluation of Visual Eye-to-Eye Calibration

First, geometric and visual eye-to-eye calibration as described above were implemented in C/C++ (using MINPACK [18] and `sparseLM` [16]) in order to compare both methods with synthetic data. For each test case, N_i random 3d points with uniform distribution were created in front of 2 virtual cameras with random spatial arrangement set apart by $\Delta\alpha_1 = 60°$ and $\|\Delta\boldsymbol{t}_1\| = 25\,\text{cm}$, image size $800 \times 600\,\text{px}$ and $60° \times 46.8°$ field of view (FOV). m random poses of the reference camera with max. rotation angle $\alpha_{\max} = 30°$ and distance $d_{\max} = 1\,\text{m}$ w. r. t. the original location were created, providing up to N_i 2d projections into the virtual image of the i-th camera per keyframe. Zero-mean Gaussian noise with standard deviation $\sigma_{\boldsymbol{x}}$ was added to all 2d points prior to pose estimation.

In the first test, all 3d points are supposed to be known, resembling the case of using a calibration object for each camera. In the second test, only 3d points for the reference camera are known, corresponding to the proposed scenario. In the third test, all 3d points are assumed unknown ($\Delta\boldsymbol{t}_1$ can only be recovered up to scale here). Camera poses are computed from 2d/3d matches for known 3d

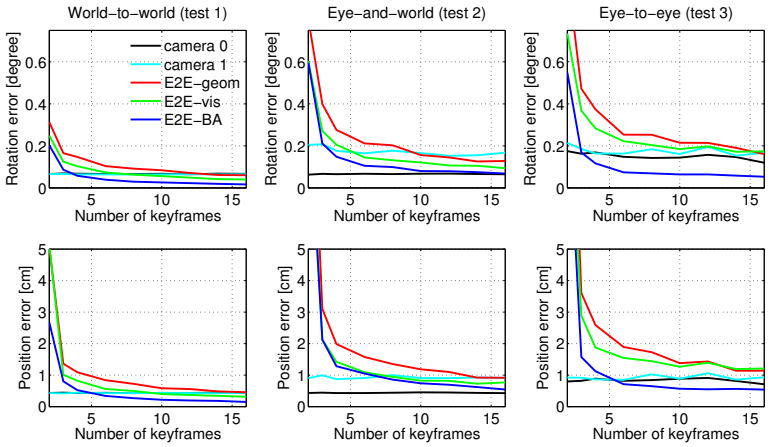

Fig. 3. Evaluation of pose estimation and eye-to-eye calibration accuracy with respect to number of keyframes m (left column: known 3d points for both cameras [test 1], middle column: known 3d points for reference camera [test 2], right: known 3d points for none [test 3]; upper row: rotation errors, lower row: position errors).

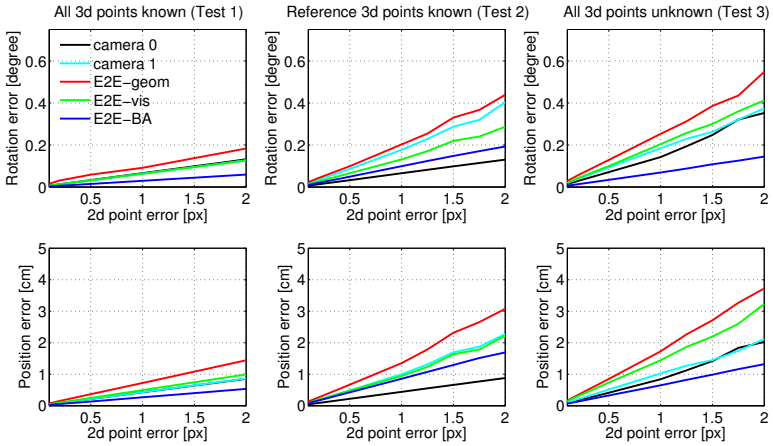

Fig. 4. Evaluation of pose estimation and eye-to-eye calibration accuracy with respect to 2d point error $\sigma_{\boldsymbol{x}}$ (see Fig. 3 for description).

points (use $N_i = 100$), otherwise via SfM initialized with the first two keyframes and extended via triangulation for each subsequent keyframe (use $N_i = 1000$).

Methods E2E-GEOM, E2E-VIS, and E2E-BA were evaluated for 1000 random samples with respect to the number of keyframes m for fixed $\sigma_{\boldsymbol{x}} = 1\,\text{px}$ resp. 2d point error $\sigma_{\boldsymbol{x}}$ with fixed $m = 8$. The resulting average pose estimation errors for both cameras and eye-to-eye calibration errors for all methods are shown in Fig. 3 and Fig. 4. In all cases, E2E-VIS is capable of improving the results from E2E-GEOM. This becomes most significant when SfM and absolute pose estimation from known 3d points are combined (2nd test). In general, calibration accuracy increases with rising number of keyframes and 2d point accuracy.

4.2 Eye-to-Eye Structure from Motion Application

The complete eye-to-eye SfM pipeline including image preprocessing, feature detection and matching (using methods from the OpenCV library [2]) was evaluated with rendered and real image sequences. In order to achieve robustness against erroneous feature point matches, RANSAC is used in the SfM initialization and tracking stages, and triangulated 3d points are pruned by evaluating their reprojection errors using the *X84* outlier rejection rule [8].

In the first test, a sequence consisting of 87 images ($800 \times 600\,\text{px}$) viewed by a virtual rig composed of 4 cameras was rendered (see Fig. 5). The scene size is $8 \times 8 \times 3\,\text{m}$. Cameras $\mathcal{C}_1$ and $\mathcal{C}_2$ are yawed 81° left and right w. r. t. reference camera $\mathcal{C}_0$, camera $\mathcal{C}_3$ is tilted 30° upwards. The distance to $\mathcal{C}_0$ is 57.4 cm for $\mathcal{C}_{1/2}$ and 70.1 cm for $\mathcal{C}_3$. $\mathcal{C}_0$ has $60° \times 46.8°$ FOV, the other cameras are limited to $53.1° \times 41.1°$. SfM initialization succeeded after 8 images. For each camera, 10 keyframes were used for eye-to-eye calibration. The pose estimation errors for

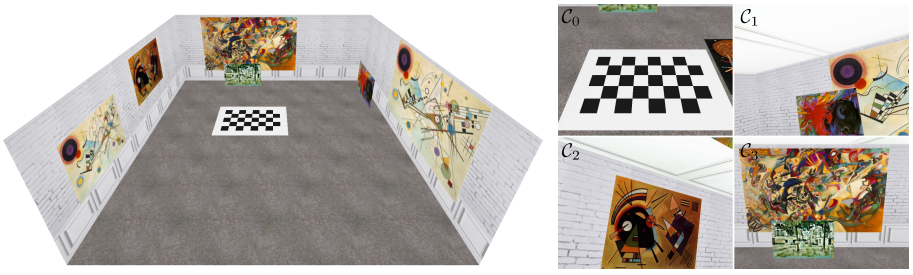

Fig. 5. Scene and example images of 4 cameras from virtual test dataset.

each camera during eye-to-eye SfM are shown in Fig. 6. While the pattern-based pose estimation for $\mathcal{C}_0$ has constant error, pose estimation errors of $\mathcal{C}_{1-3}$ via SfM vary depending on the visible scene and motion. However, the plots show that intermediate rigid motion constraint enforcement is capable of preventing drift and reducing the average pose estimation error over time. The calibration error is $< 0.3°$ in rotation and $1.1\% - 1.7\%$ in translation (Table 1), improving comparable test results from [21] ($\Delta\alpha_{\text{err}} \approx 0.8°$, $\Delta t_{\text{err}} \approx 1.8\%$ for 10 views).

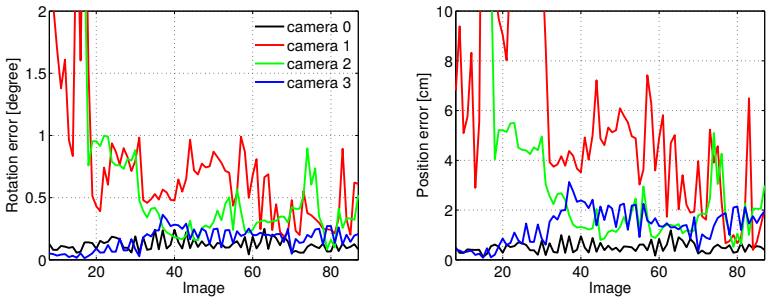

Fig. 6. Pose estimation errors for virtual test dataset (left: rotation, right: position).

Table 1. Eye-to-eye parameters and calibration results for virtual test dataset INDOOR (rotation angles in XYZ order, symbols with $*$ indicate ground truth values).

	$\Delta\alpha_x^*$	$\Delta\alpha_y^*$	$\Delta\alpha_z^*$	Δt_x^*	Δt_y^*	Δt_z^*	$\Delta\alpha_x$	$\Delta\alpha_y$	$\Delta\alpha_z$	Δt_x	Δt_y	Δt_z	$\Delta\alpha_{\text{err}}$	Δt_{err}
$\mathcal{C}_1$	60°	60°	−5°	50	20	−20	60.1°	60.0°	−4.99°	49.9	20.8	−20.5	0.07°	1.0 cm
$\mathcal{C}_2$	60°	−60°	5°	−50	20	−20	59.8°	−60.1°	4.65°	−49.5	19.3	−20.4	0.21°	0.98 cm
$\mathcal{C}_3$	30°	0°	0°	0	−50	−50	30.0°	−0.01°	0.01°	0.03	−49.2	−49.9	0.03°	0.8 cm

In the second test, a video sequence was captured with a real setup consisting of two *Point Grey Grasshopper®* (GRAS-20S4C-C) cameras equipped with *Schneider-Kreuznach Cinegon 1.8/4.8* lenses with $70° \times 56°$ FOV. Camera $\mathcal{C}_1$ is mounted approx. 25 cm to the right of $\mathcal{C}_0$ and is rotated towards the upper left direction (Fig. 7). Note that the cameras have partially overlapping fields of view. However, this is used only for validation of the calibration results. An

image sequence of 320 images (800×600 px) captured during handheld motion was used for the eye-to-eye SfM pipeline, providing 24 keyframes in total.

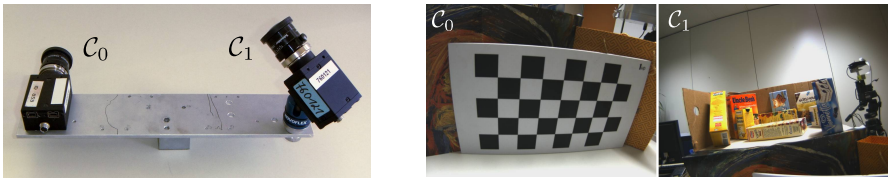

Fig. 7. Camera setup and example images from real test dataset.

The calibration results are shown in Table 2. For comparison, the results from classical stereo calibration according to [26] were used instead of ground truth data. The translational part of $\Delta\mathbf{T}_1$ differs by 1.4% which is slightly better than comparable results from [21] ($\Delta\alpha_{\text{err}} \approx 0.7°$, $\Delta t_{\text{err}} \approx 1.6\%$ for > 12 views).

Table 2. Eye-to-eye parameters and calibration results for real test dataset BOXES (rotation angles in XYZ order, symbols with $*$ indicate stereo calibration results).

	$\Delta\alpha_x^*$	$\Delta\alpha_y^*$	$\Delta\alpha_z^*$	Δt_x^*	Δt_y^*	Δt_z^*	$\Delta\alpha_x$	$\Delta\alpha_y$	$\Delta\alpha_z$	Δt_x	Δt_y	Δt_z	$\Delta\alpha_{\text{err}}$	Δt_{err}
$\mathcal{C}_1$	38.5°	−36.2°	22.1°	24.8	−8.5	−7.4	38.4°	−35.9°	22.3°	24.6	−8.7	−7.2	0.42°	0.37 cm

5 Conclusion

In this paper we proposed a Structure from Motion framework with integrated eye-to-eye calibration that is capable of estimating the extrinsics of a multi-camera system with non-overlapping views stepwise for each camera with respect to a designated reference camera capturing images of a default calibration pattern. We proposed a novel method for eye-to-eye calibration (E2E-VIS) based on reprojection errors instead of pose-based error functions as used in existing methods (E2E-GEOM) adopted from hand-eye calibration. It was demonstrated that E2E-VIS improves the results from E2E-GEOM and can be used a preprocessing step for advanced optimization methods such as eye-to-eye bundle adjustment (E2E-BA). Accurate calibration results could be obtained in experiments with both synthetic data and real image sequences.

Future work. A remaining disadvantage of E2E-BA as final optimization step is the large problem size for systems consisting of several cameras. This problem could be solved by either pruning the resulting 3d point clouds prior to joint bundle adjustment or by removing explicit 3d point parameters from the error function entirely as proposed in [22] for monocular bundle adjustment. Furthermore, real-time processing of the proposed algorithm should be achieved by further parallelization and usage of GPU accelerated algorithms as present in more recent real-time SfM applications such as DTAM [19].

References

1. Andreff, N., Horaud, R., Espiau, B.: On-line hand-eye calibration. In: 2nd International Conference on 3D Digital Imaging and Modeling. pp. 430–436 (1999)
2. Bradski, G.: The OpenCV library. Dr. Dobb's Journal of Software Tools (2000)
3. Caspi, Y., Irani, M.: Alignment of non-overlapping sequences. International Journal of Computer Vision 48(1), 39–51 (2002)
4. Chen, H.H.: A screw motion approach to uniqueness analysis of head-eye geometry. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 145–151 (1991)
5. Dornaika, F., Chung, R.: Stereo geometry from 3d ego-motion streams. IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics 33(2), 308–323 (2003)
6. Esquivel, S., Koch, R.: Structure from motion using rigidly coupled cameras without overlapping views. In: 35th German Conference on Pattern Recognition. Lecture Notes in Computer Science, vol. 8142, pp. 11–20 (2013)
7. Esquivel, S., Woelk, F., Koch, R.: Calibration of a multi-camera rig from non-overlapping views. In: 29th DAGM Symposium on Pattern Recognition. Lecture Notes in Computer Science, vol. 4713, pp. 82–91 (2007)
8. Farenzena, M., Fusiello, A., Gherardi, R.: Structure-and-motion pipeline on a hierarchical cluster tree. In: IEEE International Conference on Computer Vision Workshops. pp. 1489–1496 (2009)
9. Frahm, J.M., Köser, K., Koch, R.: Pose estimation for multi-camera systems. In: 26th DAGM Symposium on Pattern Recognition. Lecture Notes in Computer Science, vol. 3175, pp. 286–293 (2004)
10. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, 2nd edn. (2004)
11. Hesch, J.A., Mourikis, A.I., Roumeliotis, S.I.: Mirror-based extrinsic camera calibration. In: Workshop on the Algorithmic Foundations of Robotics. pp. 285–299 (2008)
12. Kim, J.H., Chung, M.J.: Absolute motion and structure from stereo image sequences without stereo correspondence and analysis of degenerate cases. Pattern Recognition 39(9), 1649–1661 (2006)
13. Kumar, R.K., Ilie, A., Frahm, J.M., Pollefeys, M.: Simple calibration of non-overlapping cameras with a mirror. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–7 (2008)
14. Lébraly, P., Royer, E., Ait-Aider, O., Deymier, C., Dhome, M.: Fast calibration of embedded non-overlapping cameras. In: IEEE International Conference on Robotics and Automation. pp. 221–227 (2011)
15. Li, B., Heng, L., Köser, K., Pollefeys, M.: A multiple-camera system calibration toolbox using a feature descriptor-based calibration pattern. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1301–1307 (2013)
16. Lourakis, M.I.A.: Sparse non-linear least squares optimization for geometric vision. In: European Conference on Computer Vision. vol. 2, pp. 43–56 (2010)
17. Lourakis, M.I.A., Argyros, A.A.: The design and implementation of a generic sparse bundle adjustment software package based on the Levenberg-Marquardt algorithm. Tech. Rep. #340, Institute of Computer Science, Foundation for Research and Technology – Hellas (FORTH) (2004)
18. Moré, J.J., Garbow, B.S., Hillstrom, K.E.: User guide for MINPACK-1. Tech. Rep. ANL-80-74, Argonne National Laboratory (1980)

19. Newcombe, R.A., Lovegrove, S., Davison, A.J.: DTAM: Dense tracking and mapping in real-time. In: IEEE International Conference on Computer Vision. pp. 2320–2327 (2011)
20. Pagel, F.: Calibration of non-overlapping cameras in vehicles. In: IEEE Intelligent Vehicles Symposium. pp. 1178–1183 (2010)
21. Rodrigues, R., Barreto, J.P., Nunes, U.: Camera pose estimation using images of planar mirror reflections. In: European Conference on Computer Vision. Lecture Notes in Computer Science, vol. 6314, pp. 382–395 (2010)
22. Rodríguez, A.L., de Teruel, P.E.L., Ruiz, A.: GEA optimization for live structureless motion estimation. In: IEEE International Conference on Computer Vision. pp. 715–718 (2011)
23. Strobl, K.H., Hirzinger, G.: Optimal hand-eye calibration. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 4647–4653 (2006)
24. Terzakis, G., Culverhouse, P., Bugmann, G., Sharma, S., Sutton, R.: A recipe on the parameterization of rotation matrices for non-linear optimization using quaternions. Tech. Rep. MIDAS.SMSE.2012.TR.004, Marine and Industrial Dynamic Analysis School of Marine Science and Engineering, Plymouth University (2012)
25. Tsai, R.Y., Lenz, R.K.: A new technique for fully autonomous and efficient 3d robotics hand/eye calibration. IEEE Transactions on Robotics and Automation 5(3), 345–358 (1989)
26. Zhang, Z.: A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1330–1334 (2000)