Relative Bundle Adjustment based on Trifocal Constraints

Richard Steffen, Jan-Michael Frahm and Wolfgang Förstner

Department of Computer Science, The University of North Carolina at Chapel Hill
{rsteffen,jmf}@cs.unc.edu

Department of Photogrammetry, University of Bonn
wf@ipb.uni-bonn.de

Abstract. In this paper we propose a novel approach to bundle adjustment for large-scale camera configurations. The method does not need to include the 3D points in the optimization as parameters. Additionally, we model the parameters of a camera only relative to a nearby camera to achieve a stable estimation of all cameras. This guarantees to yield a normal equation system with a numerical condition, which practically is independent of the number of images. Secondly, instead of using the classical perspective relation between object point, camera and image point, we use epipolar and trifocal constraints to implicitly establish the relations between the cameras via the object structure. This avoids the explicit reference to 3D points thereby handling points far from the camera in a numerically stable fashion. We demonstrate the resulting stability and high convergence rates using synthetic and real data.

1 Introduction

Motivation. Bundle adjustment has become the workhorse of structure from motion estimation, triggered by the review by Triggs et al. [1] and the first public domain software by Lourakis and Argyros [2] and more recently made fully aware by the software *bundler* [3]. The generality of the concept and the optimality of the achieved solution cause bundle adjustment to serve as a reference and to be of broad interest.

Despite these advantages, some problems still exist: The stability of large systems is sensitive to the arrangement of images as in classical photogrammetric mapping applications, tend to show instabilities. These instabilities are difficult to identify, and to date, there are still no tools for giving recommendations how to cure the situation by deliberately taking additional images, a precondition to make bundle adjustment usable by non-specialists. In real time applications, identifying and resolving so-called loop closures, where after long image strips one reaches positions visited in the past, requires careful storage management for fast access and proper representation of the geometry taken up to that point.

This paper proposes a novel model for bundle adjustment, especially useful for dealing with weak configurations due either to long motion sequences or due to the existence of points far from the cameras. First of all, long image sequences accumulate drift leading to a random walk, which decreases the accuracy

and increases the numerical condition number. The condition number increases super-linear with time. Secondly, points far from the cameras cause problems when determining approximate values, especially in the case where the camera positions are not yet determined with enough stability. This can happen for example, when the parallactic angle between two viewing rays is too small to reliably identify the depth of a point.

The proposed concept integrates two remedies: (1) camera parameters are not represented w. r. t. a common world system, but relative to a well chosen set of cameras distributed over the complete set of all cameras. This leads to a tree type structure with kinematic chains linking the cameras with the reference cameras. This increases the numerical stability of the bundle adjustment, increases the speed of convergence and the robustness with respect to bad initial values. (2) The object structure is not represented explicitly but using pairs and triplets of geometric constraints between cameras. This avoids the handling of 3D points far off the cameras, which in turn avoids any problem with determining approximate values. On the contrary, points far away from the cameras can be used advantageously to stabilize the rotation information. The cost for using this advantage is the slightly increased complexity of the Jacobians.

Related work. There is a lot of work on hierarchically representing large sets of images, in order to partition the bundle estimation into smaller better conditioned subsystems, for example partitioning an image sequence hierarchically [4], applying a spectral decomposition of the connection graph [5], [6] or building a tree based on the overlap of pairs of images [7], building a hierachical map during simultaneous localisation and mapping [8] and performing an effcient, close to optimal estimation. These approaches may also be coupled with our setup. The most closely related work is the setup in [9], where camera parameters are related to reference views which are related to a world system. However, the individual parts are connected in a second step, which altogether does not lead to a statistically optimal solution. Similarily, non-Euclidian object point representation like the inverse depth representation proposed by [10] can be applied to bundle adjustment to model points at infinity. In contrast to our proposed method, [10] still has to include the points as parameter.

Using trifocal constraints has been proposed to avoid the explicit representation of 3D points within the estimation of an image triplet [11]. To use trifocal constraints within bundle adjustment already has been proposed in [4], however, only for chaining within an image sequence and deriving approximate values. Trifocal constraints have been used within an extended Kalman filter approach in [12]. No approach is known to the authors that (1) only uses constraints between the image observations and the cameras; (2) uses a relative representation of the camera positions for improving the numerical condition; (3) performs a statistically optimal estimation equivalent to classical bundle adjustment.

The paper is structured the following way: we first describe the estimation procedure base on epipolar and trifocal constraints, give an insight into the modelling of the relative camera poses and then demonstrate the strengths of the approach with synthetic and real data.

2 Model for Relative Bundle-Adjustment with Image Triplet Constraints

This section describes the approach in more detail, first contrasting the classical bundle adjustment model using direct observations equations with the estimation procedure with implicit constraints, then deriving the constraints and their use in the bundle adjustment and finally introducing the representation with relative camera poses.

2.1 Classical bundle adjustment

Classical bundle adjustment simultaneously estimates the 3D structure of the environment and the camera parameters by minimizing the reprojection error in a weighted least square manner. The co-linearity constraint relates the parameters $\boldsymbol{q}$ describing the object, usually being a set of 3D points, the parameters $\boldsymbol{z}$ describing the 2D image observations, usually the image points, and the extrinsic, possibly also the intrinsic camera parameters $\boldsymbol{p}$ using an explicit observation model $\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{p}, \boldsymbol{q})$, often called a non-linear Gauss-Markov model. Additional constraints $\boldsymbol{h}(\boldsymbol{p}, \boldsymbol{q}) = \boldsymbol{0}$ on the object or camera parameters may be used to fix the gauge and to enforce certain properties of the object to be recovered. Assuming the image measurements have a covariance that is denoted in matrix form as C_{zz}, the classical approach [1] minimizes the reprojection errors or residuals $\boldsymbol{v}(\boldsymbol{p}, \boldsymbol{q}) = \boldsymbol{f}(\boldsymbol{p}, \boldsymbol{q}) - \boldsymbol{z}$ weighted with the inverse covariance matrix under the given constraints leading to the energy function

$$E(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{\mu}) = \boldsymbol{v}^{\mathsf{T}}(\boldsymbol{p}, \boldsymbol{q})\, \mathsf{C}_{zz}^{-1}\, \boldsymbol{v}(\boldsymbol{p}, \boldsymbol{q}) + \boldsymbol{\mu}^{\mathsf{T}} \boldsymbol{h}(\boldsymbol{p}, \boldsymbol{q}) \tag{1}$$

to be minimized, where $\boldsymbol{\mu}$ are the corresponding Lagrangian parameters for the constraints. The iterative solution typically exploits the sparsity of the structure of the normal equation system and is optimized by a marginalization to the usually much smaller number of camera parameters $\boldsymbol{p}$ using the *Schur complement*.

A novel model for bundle adjustment to gain efficiency and numeric stability by first deploying implicit constraints between the observations and the camera parameters to eliminate the object points as estimated parameters and second, by instead of referring the camera parameters $\boldsymbol{p}$ to a common world system representing the camera poses by the relative poses $\boldsymbol{p}_{rs}$ between neighbouring cameras $\boldsymbol{p}_r$ and $\boldsymbol{p}_s$ is presented.

2.2 Estimation with implicit constraints

We replace the classical reprojection model using the epipolar and trifocal constraints, well known from the two and three view epipolar geometry [13]. The epiolar and trifocal constrains are implicit functions of both the camera parameters and the observations, and do not allow to express the observations as a function of the parameters. Thus, instead of using the explicit observation functions $\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{p}, \boldsymbol{q})$, we use constraints $\boldsymbol{g}(\boldsymbol{p}, \boldsymbol{z})$ between the camera parameters $\boldsymbol{p}$ and the observed image observations $\boldsymbol{z}$. Constraint optimization is known in the

classical least square estimation technique as the Gauss-Helmert model and can be solved by minimizing

$$E(\boldsymbol{p}, \boldsymbol{v}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{v}^\mathsf{T} \mathsf{C}_{zz}^{-1} \boldsymbol{v} + \boldsymbol{\lambda}^\mathsf{T} \boldsymbol{g}(\boldsymbol{p}, \boldsymbol{z} + \boldsymbol{v}) + \boldsymbol{\mu}^\mathsf{T} \boldsymbol{h}(\boldsymbol{p}) \tag{2}$$

w. r. t. the parameters $\boldsymbol{p}$, the residuals $\boldsymbol{v}$ and the Lagrangian parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. The optimal estimates $\widehat{\boldsymbol{p}}$ for the parameters and the fitted observations $\widehat{\boldsymbol{z}}$ should fulfill the model constraints $\boldsymbol{g}(\widehat{\boldsymbol{p}}, \widehat{\boldsymbol{z}}) = \mathbf{0}$ and $\boldsymbol{h}(\widehat{\boldsymbol{p}}) = \mathbf{0}$. Minimizing the energy function (2) can be iteratively achieved by determining the corrections $\widehat{\Delta \boldsymbol{p}}$ from the linear equation system

$$\begin{bmatrix} J_p^\mathsf{T}(J_z^\mathsf{T} C_{zz} J_z)^{-1} J_p & H \\ H^\mathsf{T} & 0 \end{bmatrix} \begin{bmatrix} \Delta \boldsymbol{p} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} J_p^\mathsf{T}(J_z^\mathsf{T} \mathsf{C}_{zz} J_z)^{-1} \boldsymbol{c}_g \\ \boldsymbol{c}_h \end{bmatrix}, \tag{3}$$

with

$$\boldsymbol{c}_g = -\boldsymbol{g}(\widehat{\boldsymbol{p}}, \widehat{\boldsymbol{z}}) + J_z^\mathsf{T}(\widehat{\boldsymbol{z}} - \boldsymbol{z}) \qquad \text{and} \qquad \boldsymbol{c}_h = -\boldsymbol{h}(\widehat{\boldsymbol{p}}), \tag{4}$$

starting at approximate values $\widehat{\boldsymbol{p}}^{(\nu)}$ for the estimated parameters and the fitted observations $\widehat{\boldsymbol{z}}$. The matrices J_p and J_z^T are the Jacobian of the constraints $\boldsymbol{g}$ with respect to the parameter vector $\boldsymbol{p}$ and the observations $\boldsymbol{z}$ and H^T is the Jacobian of $\boldsymbol{h}$ with respect to the parameters evaluated at the approximate values. The residuals can be determined from $\boldsymbol{v}^{(\nu)} = -\mathsf{C}_{zz} J_z (J_z^\mathsf{T} \mathsf{C}_{zz} J_z)^{-1} (\boldsymbol{c}_g - J_p \Delta \boldsymbol{p})$. We iteratively find new approximate values for the estimated parameters $\widehat{\boldsymbol{p}}^{(\nu+1)} = \widehat{\boldsymbol{p}}^{(\nu)} + \Delta \boldsymbol{p}$ and the fitted observations $\widehat{\boldsymbol{z}}^{(\nu+1)} = \boldsymbol{z} + \boldsymbol{v}^{(\nu)}$. This will be very useful to reconstruct the structure of the environment in our approach simply by intersecting two estimated projection rays, as the rays derived from the fitted observations $\widehat{\boldsymbol{z}}$ intersect and no optimization needs to be performed anymore.

Conceptionally, the explicit estimation and the implicit estimation model both minimize the residuals $\boldsymbol{v}$. It has been shown in [14, 15], that the implicit model is a generalization of the least square estimation framework. Our proposed implicit model minimizes the backprojection errors (residuals) in a least square manner, using the constraint that the projection rays intersect in a single point as the explicit model does. The used epipolar and trifocal constraints can be derived from the explicit reprojection model by eliminating the object point [16]. Therefore, the results of the classical formulation and our proposed formulation are equal w.r.t the solution and its precision. Additionally, robustification can be achieved by reweighting the residuals as used in the classical model as a standard enhancement.

2.3 Epipolar and Trifocal Constraint Bundle

In all cases we rely on the classical partitioning of the projection $\mathbf{x}_t = \mathsf{P}_t \mathbf{X}$ of a 3D point with homogeneous coordinates $\mathbf{X}$ into the t-th camera and the partitioning of the projection matrix P_t into its internal part, containing the intrinsic parameters in K and its external part, containing its motion M_t w. r. t. to a reference system

$$\mathsf{P}_t = \mathsf{P}_{0t} \mathsf{M}_t \qquad \text{with} \qquad \mathsf{P}_{0t} = [\mathsf{K}_t \mid \mathbf{0}] \qquad \text{and} \qquad \mathsf{M}_t = \begin{bmatrix} R_t & \boldsymbol{T}_t \\ \mathbf{0}^\mathsf{T} & 1 \end{bmatrix} \tag{5}$$

This decomposition will be deployed in the next section for introducing the kinematic chains.

We now propose to replace the classical projection model by using the epipolar and trifocal constraints in order to achieve two goals: (1) avoid the provision of approximate values for the 3D object points, which in case of bad approximate values, may be far off the true values and hinder the estimation process to converge and (2) to allow for points which are very far from the cameras or even at infinity. The epipolar constraint for two cameras, r and s, can be written as

$$\mathbf{0} = g_E(\boldsymbol{p}, \boldsymbol{z}) = \mathbf{x}_r^\mathsf{T} \mathsf{K}_r^{-\mathsf{T}} R_r S(\boldsymbol{B}_{r,s}) R_s^\mathsf{T} \mathsf{K}_s^{-1} \mathbf{x}_s \tag{6}$$

with $S(\cdot)$ indicating the skew matrix of the base line vector $\boldsymbol{B}_{r,s} = \boldsymbol{T}_s - \boldsymbol{T}_r$ between the two projection centres of camera r and s. For calibrated cameras the observed homogeneous image coordinates $\mathbf{x}$ can be normalized by applying K. Equation (6) constrains the parameters of the two cameras, which themselves will in general depend on all relative motions which connect the two cameras indexed with r and s. The trifocal constraint can be interpreted as the intersection of 4

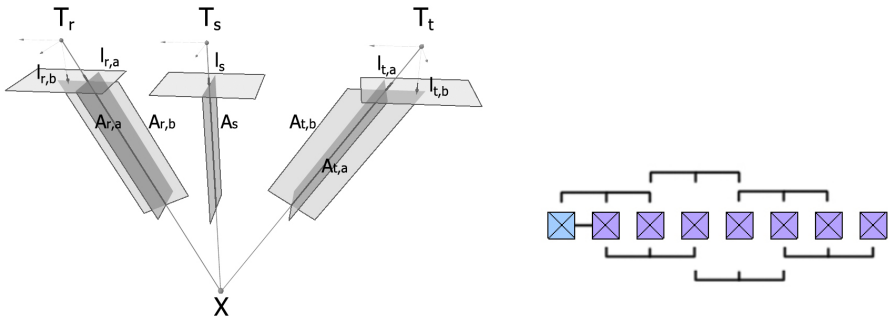

Fig. 1. Left: Scheme of the trifocal constraint. Four planes have to intersect in a single object point. The planes are inverse projections of 2D lines intersecting the observed image points. Right: Chaining the trifocal constraint using consecutive images. An epipolar constraint is introduced between the first (blue) and second image. Two trifocal constraints are introduced for every new image the point is observed

planes in a single point. As for instance outlined in [17], it can be written as

$$0 = g_T(\boldsymbol{p}, \boldsymbol{z}) = \det\left[\mathbf{A}_{r,i_a}, \mathbf{A}_{r,i_b}, \mathbf{A}_{s,i}, \mathbf{A}_{t,i}\right] \tag{7}$$

with the projection plane

$$\mathbf{A}_i = \mathsf{P}^\mathsf{T} \mathbf{l}_i \tag{8}$$

for the line $\mathbf{l}_i$ in camera t. The 2d line in each image has to intersect the observed image point $\mathbf{x}_i$. Our method ensures this by choosing an arbitrary direction α

and computing the line by

$$\mathbf{l}_i = \begin{bmatrix} \sin(\alpha) \\ -\cos(\alpha) \\ \cos(\alpha)y_i - \sin(\alpha)x_i \end{bmatrix}. \tag{9}$$

After introducing the basic constraints (6) and (7) used by our method we will now detail the constraint selection for each observation of an observed image point $\mathbf{x}_i$.

Assume an object point i is observed in a set of cameras $t_1 ... t_k$. A point observed for the first time provides no constraint. A point observed in two cameras delivers one epipolar constraint (6). The observation of a point in more than two cameras provides two constraints based on the trifocal tensor. We introduce two trifocal constraints in the following manner. We randomly select two lines $\boldsymbol{l}_{1,a}$ and $\boldsymbol{l}_{1,b}$, one line $\boldsymbol{l}_{2,a}$ and two lines $\boldsymbol{l}_{k,a}$ and $\boldsymbol{l}_{k,b}$ (a and b are the indices of two lines in the image), which have different directions and hence provide two constraints through Equation 1, c. f. Figure 1.

The scheme for chaining trifocal constraints for consecutive images is outlined in Figure 6. When using the epipolar and trifocal constraints in the bundle adjustment, we do not need 3D object point coordinates, which have to be optimized. Instead, the object points are encoded implicitly in the trifocal constraints. Furthermore, the trifocal constraints are accountable for the transition of the scale through the chain of images ensuring a reconstruction with a consistent scale.

The choice of the lines for the trifocal constraint in Equation (8) directly influences the numerical stability of the system. We use this to our advantage by determining a proper combination of five lines that leads to the smallest condition of $J_z^\mathsf{T} J_z$ for the camera triplets in concern. This enhances the numerical stability of our bundle adjustment. The choice has five degrees of freedom, corresponding to the rotations of the planes $\boldsymbol{A}$ around the projection rays. Obtaining the best configuration is a non-trival optimization problem in itself.

For efficiency we opted for a simple random sampling strategy to obtain an acceptable set of lines. First, we choose random line directions and then we evaluate the condition number for our particular choice of lines. In case the condition number is too high to obtain a numerically stable solution, we randomize again. We empirically found that the space of acceptable configurations in order to achieve numerical stability is significantly larger than the space of the weak configurations. We leave a formal proof of this fact to future work.

2.4 Relative Camera Representation with Kinematic Chains

Modeling camera poses. One of the main problems of traditional bundle adjustment is that the condition of the information matrix (normal equation matrix) for large scale environments becomes huge leading to numerical instabilities of the linear solvers used. This is one of the reasons why hierarchical representations of large sets of images and the relative representation of camera positions are

used. The camera orientations are represented locally, depending on an arbitrary local coordinate system.

The idea of [9] we are following here is to choose some reference cameras, say with pose M_t and model the pose of its k-th neighbour M_{t+k} using the relative pose $\mathsf{M}_{t,k} = \mathsf{M}_{t+k}\mathsf{M}_{t+k-1}^{-1}$ and estimate the rotation and translation parameters of this relative motion $\mathsf{M}_{t,k}$. This leads to the recursive relation

$$\mathsf{M}_{t+k} = \mathsf{M}_{t,k}\mathsf{M}_{t+k-1} \tag{10}$$

or when modeling the complete kinematic chain from t to $t+k$

$$\mathsf{M}_{t+k} = \mathsf{M}_{t,k}\mathsf{M}_{t,k-1}\cdot\ldots\cdot\mathsf{M}_{t,1}\mathsf{M}_t. \tag{11}$$

The projection matrix P_{t+k} thus refers to the reference camera using

$$\mathsf{P}_{t+k} = \mathsf{P}_{0,t+k}\mathsf{M}_{t+k} = \mathsf{P}_{0,t+k}\mathsf{M}_{t,k}\mathsf{M}_{t,k-1}\cdot\ldots\cdot\mathsf{M}_{t,1}\mathsf{M}_t \tag{12}$$

In case one has a constraint between two or three cameras, one needs to identify the path between these two via the reference cameras. Obviously, the sparseness of the Jacobian J_p now depends on the length of these chains of cameras observing the same individual object point.

Sparsity of the normal equation system. We now analyze, how our representation influences the structure of the linear solver and propose a strategy to increase the sparseness given an image sequence containing large loops. We are aware of the fact that this method may not always achieve optimal sparsity for example for image collections. Here we demonstrate that the sparsification is an important property to solve the unknown parameters more efficiently. While there are structural differences between the classical formulation of bundle adjustment and our method we will demonstrate how to take advantage of the same set of methods to improve the computational performance. We start with an example of a simulated environment illustrated on the left hand of Figure 2 and consisting of two loops. The simulated sequence contains of 71 images and approximately 170 object points on the planar surface. The structures of the normal equation matrices are shown in Fig. 3. The classical structure leads to the sparsest structure. A naive choice of the relative motions between cameras would follow the numbering of the cameras. Here, it leads to a nearly full normal equation matrix, as the first loop is the one from image 1 to image 41, and the second one is the large loop, containing all images except 1 to 16, resulting in the overlay of two square blocks. Therefore, one needs to analyze the effect of a certain numbering onto the structure of the normal equation matrix for the new type of representation

The matrix $\mathsf{J}_z^\mathsf{T} C_{zz} \mathsf{J}_z$ is a block-diagonal matrix. Every block represents the set of constraints involving an observed object point. The determination of $(\mathsf{J}_z^\mathsf{T} C_{zz} \mathsf{J}_z)^{-1} \mathsf{J}_p$ in (3) then can be done block-wise by solving a linear equation system, exploiting the sparsity of $\mathsf{J}_z^\mathsf{T} C_{zz} \mathsf{J}_z$. For structure from motion scenes, where an object point is only visible in a small subset of all cameras used, this

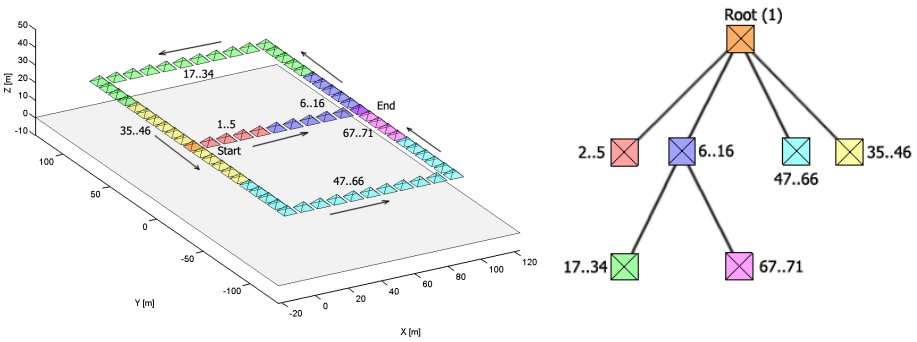

Fig. 2. Simulated example of a kinematic chain and spanning tree. The simulated sequence contains 71 images and 170 object points on a planar surface. Left: A camera trajectory at a birds eye view. Right: Simplified graph shown as a tree. The computed six subtrees are colored in a different manner.

is usually not computationally expensive. The resulting information matrix has the size of the number of the camera parameters plus the number of constraints for the gauge only. As we can see in our example in Figure 3 the resulting information matrix is sparse too and therefore the equation system can be solved efficiently. The sparseness of the information matrix varies with the choice of

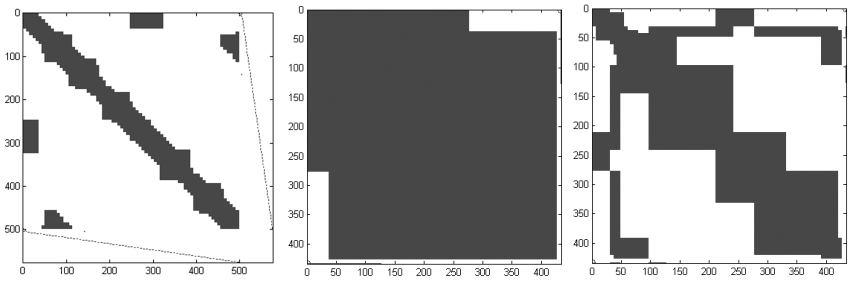

Fig. 3. Structure of the normal equation system. Left: Classical absolute representation. Middle: relative representation, naively taking all images in the order of appearance. Right: relative representation using our algorithm

the reference cameras, and thus the choice of the relative representations. We therefore need to select a representation which is optimal in some sense. This leads to a trade off between the condition of the information matrix and the sparseness and therefore the computational cost of the solution of the linear solver, taking the fill-in into account. In addition an optimal solver has to resort the information matrix to reduce the computational costs.

We have developed a scheme which aims at finding a good compromise. We can represent the whole set of relative cameras using a connection graph. As a camera has a unique reference system, we have to choose one of the possible

spanning trees of the graph. For example, this can be easily achieved by enumerating all cameras in an arbitrary order. For image sequences an ordering by acquisition time is reasonable. We first sort the cameras in an ascending order. Then we obtain the connection graph for all connected cameras observing a common object point represented by an adjacency matrix. Next we search the subgraph that connects any camera to the camera with the smallest identifier. In a final step we merge the branches of the resulting tree to achieve a number of connected cameras larger than a threshold ($G_N = 5$) in each branch.

Figure 2 illustrates the simulated environment with two loops and the corresponding connection graph computed by our proposed method. The computed six sub-branches are colored differently in Figure 2. Images 1 and 6 are connected as they observing at least one common object point. Therefore our approach connects image *6* to the root node. In case a loop closing is detected, in our example in image *35*, a new branch is generated that connects it to the root node in the graph. This can be done for all used cameras incrementally.

3 Experimental Results

After the detailed description of the algorithm and the structure of the solution, we will verify its feasibility on synthetic and real datasets. The implementation has been done in Matlab™.

3.1 Simulation results

We use two synthetic datasets to demonstrate the usefulness and practicality of our novel approach. The first dataset is a long linear camera motion, for instance acquired by an aerial vehicle or a mobile camera for urban scenes of facades. Using this dataset we will analyze the behaviour of the condition number of the linear solver and the demonstrate the benefit of including points at infinity. The second dataset has already been shown in Figure 2. This dataset is used to show the convergence behaviour and the applicability of our approach for datasets with loop closure.

Both datasets are generated using a synthetic camera setup with an image resolution of 800×600 pixel, a principal point in the middle of the image and a focal length of 400 pixel. In both datasets the distance between consecutive frames is $b = 10$ m and the distance of the camera centers to the plane of the observed object points is $h_g = 30$ m. The average number of observed object points per image is approximately $N \approx 20$.

In the first experiment we compare the the condition number of the information matrix between the classical and the novel approach. This issue will be noteworthy to solve the task of structure from motion in the presence of large-scale loops. On the left hand side of Figure 4 the simulated trajectory is outlined. We varied the length of the linear path in the experiment from 100 m to a maximal length of 1000 m. We assumed Gaussian noise of 1 pixel for the observations. The right hand side of Figure 4 shows the computed logarithmic

condition numbers of the information matrix for the classical bundle adjustment and the newly proposed approach. We can observe that the condition number for the classical approach steadily increases. This increase is proportional to the increase of the uncertainty of the camera parameters toward the end of the strip. Due to the relative representation, the condition number is practically independent of the length of the trajectory in our approach. The peak at a strip length of 200 m is caused by a badly chosen direction to generate the trifocal constraint (see Section 2.3). Another important evaluation is the usefulness of incorporat-

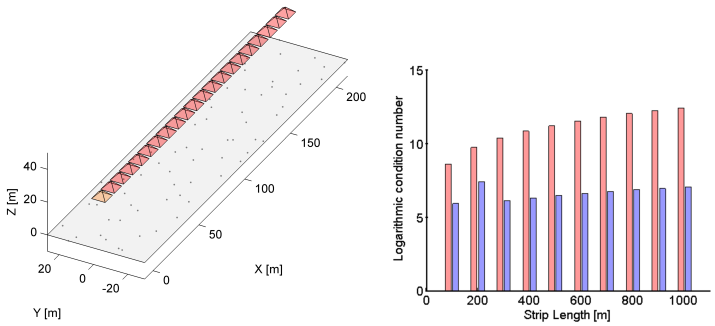

Fig. 4. Left: Long strip of consecutive cameras. The gauge is fixed to the first camera. The scale is introduced by the known true base length to the second camera. Right: Condition number of the linear equation system (Information-Matrix) for a long strip. Classical bundle adjustment (red), new method (blue). Observe, that the condition number of the experiment with a strip length of 1000 m differ by a factor of $\approx 10^5$.

ing points at infinity. In this experiment we added just three additional points at infinity. In Figure 5 the expected standard deviations of the camera parameter in

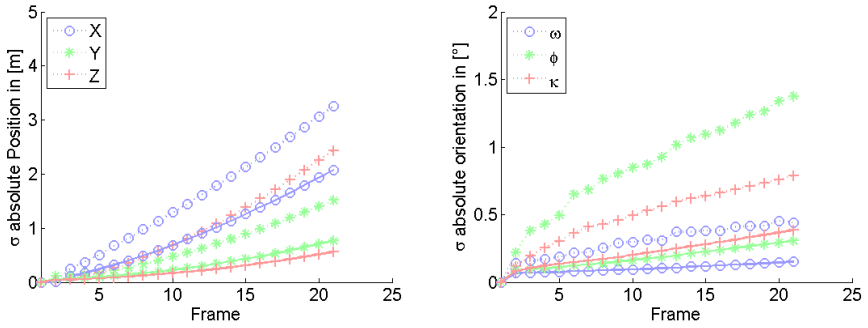

Fig. 5. Expected accuracy of the camera position (left) and rotation (right) without (dotted) and with (solid line) 3 additional points at infinity. The shown absolute uncertainty is computed by error propagation through the chain of relative representations.

the global coordinate system with and without the points at infinity are shown. The uncertainty is computed throughout all cameras by variance propagation. We can observe that the points at infinity have a significant influence on the determination of the rotation as well as the translation due to the correlation to the rotation parameters. As our approach can deal with points at infinity, the solution of a bundle will be significantly improved if points at infinity are available using the novel method.

In the last experiment using synthetic data our method is able to perform loop closures and it is robust to corrupted approximate values and large uncertainty of the observations. Approximate values for the exterior camera parameters are obtained in general computing a robust estimation of the essential matrix [13]. The rotation parameter can be usually determined very accurately, however the baseline vector can not be. Therefore, for image sequences the approximate values are chained, which leads to a random walk. In our example presented in Figure 6 (left), we generated approximate values chaining relative orientations with a randomized accuracy of 3 m for the translation parameters and 0.5 degree for the rotation parameters. On the right side of Figure 6 the mean of the

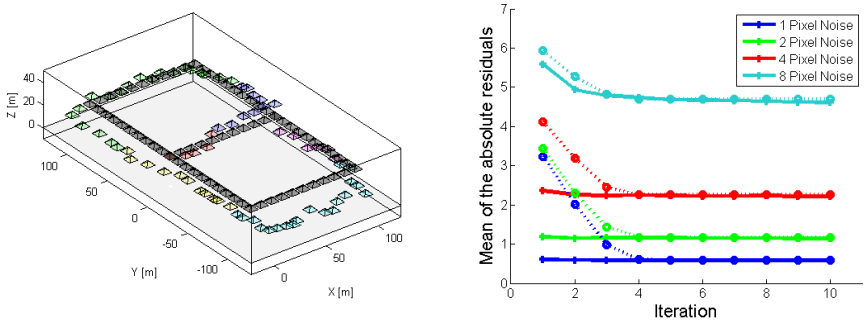

Fig. 6. Left: Simulated example with two loops. The approximated values are computed using a random walk with $3m$ Gaussian noise for the translation components and 0.5 degree for the rotation parameter, Right: Mean of the absolute sum of the residuals for 11 iterations. The dashed lines are the mean residuals for the classical bundle adjustment, the solid line for our approach.

absolute sum of the residuals for 11 iterations are presented. The dashed lines are the results using the classical model, the solid lines are the results of our approach. For the classical model the object points were initialized at the first projection ray with the known distance. Both simulations are run with the same observations and initial values. We can observe that in presence of small noise the residuals become significantly smaller in the first iterations in our method compared to the classical approach, since the object points act in the classical approach as anchor. Our proposed method does not show this disadvantageous behavior. The convergence behavior has to be examined in more detail in the future, when integration robustification methods is completed.

3.2 Real Data

We also tested our method on a real datasets. The first dataset consists of an image sequence of 624 images of the left camera of a stereo system. A feature detection and flow computation system does tracking using a graphics processing unit implementation. Additionally SIFT features are extracted and descriptor matching is performed [18]. A keyframe dataset of the whole sequence using 20 images and 55 randomly selected feature tracks has been taken. A reference trajectory was computed using the stereo tracking system of [19] including a huge number of observations. In Figure 7 sample keyframe with detected im-

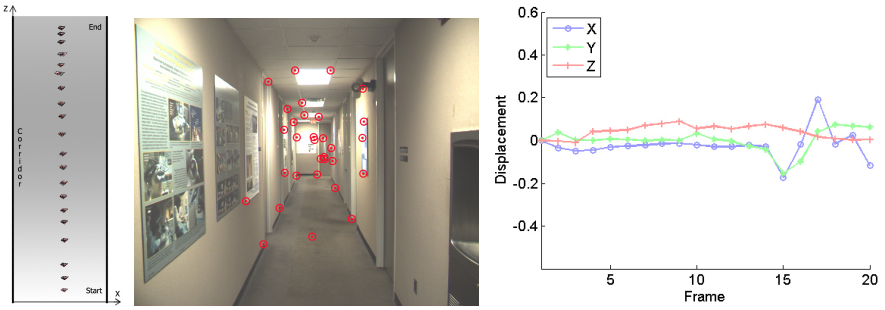

Fig. 7. Left: Birds-eye view of the corridor and estimated as well as reference camera position and orientation. Middle: Single frame extracted from an image sequence with tracked features. Right: Differences between reference camera and estimated camera parameter, X, Y, Z.

age features is shown. To the left a schematic birds-eye view of the estimated camera trajectory derived by error propagation is presented. To the right the differences of the estimated camera position to the high accuracy reference trajectory is shown. We remark, that the present implementation is not robustified and optimized for speed yet.

The second dataset consists of an image collection of the *Brandenburger Tor* containing 100 images with 1600 3d-points and roughly 24000 trifocal constraints taken from a photo-sharing website like *Flickr.com*. The focal length initally is taken from the image header and the principal point is fixed to the image center. An existing robust classical bundle adjustment incorporating intrinsic parameters as unknowns determined intrisic parameter as references. Again SIFT features are extracted and descriptor matching is performed, then pairwise relative orientations are computed using a RANSAC based scheme and outliers are rejected. To the left of Figure 8 the estimated camera orientations, as well as the reconstructed 3d object points determined by intersection 2 estimated projection rays are shown. To the right the absolute sum of the residuals thought 20 iterations are presented. The novel method decrease the residuals constantly and seems to have converged for this real dataset.

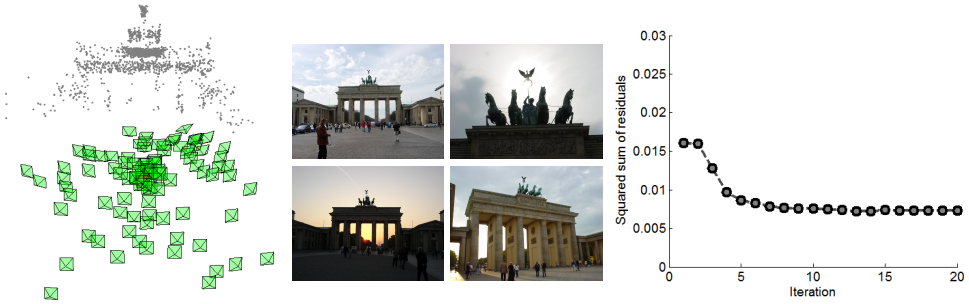

Fig. 8. Left: 3d-view of estimated camera parameters and reconstructed object points using the novel method. Middle: Example images of an image collection. Right: Absolute sum of the residuals for 20 iterations.

4 Conclusions

This paper introduced a new approach to circumvent the limitations of classical bundle adjustment by changing the observation model and the camera representation of the least square solution. The results of the classical bundle and the novel approach are equal as proved in [14, 15]. We focus in the paper on the structural differences of the normal equation system and proved the usefulness of the proposed concept on simulated data and real data. The main advantages can be summarized as follows:

- No approximate values for the object points are necessary any more. The new algorithm is therefore able to handle points at infinity. This can improve the solution of a structure from motion task significantly. In addition the prefiltering of the observations can be neglected and there is no need of the reduction of the normal equation system using the Schur-Complement.
- Due to the relative representation the condition number of the information matrix seems to be independent of the length of a camera trajectory. This is very useful for structure from motion tasks on mobile platforms.
- We observed a faster convergence and robustness in present of corrupted approximate values in our experiments compared to a classical bundle adjustment. We are aware that this fact should be investigated in more detail in future experiments.

Although our algorithm shows significant positive properties, the computation of the Jacobians using kinematic chains is computationally more complex compared to the classical formulation. We have yet to examine how this interacts with the speed up due to faster convergence.

We leave it to future work to demonstrate the performance of the new method using large image sets along with applying robustification techniques to the parameter estimation. While not demonstarted the approach can be extended

to a more general approach to accomodate uncalibrated cameras. We also plan to implement an online version, where images can be incrementally added.

References

1. Triggs, B., McLauchlan, P., Hartley, R., Fitzgibbon, A.: Bundle Adjustment A Modern Synthesis. In Triggs, B., Zisserman, A., Szeliski, R., eds.: Vision Algorithms: Theory and Practice. Volume 1883 of LNCS., Proc. of the Intl. Workshop on Vision Algorithms: Theory and Practice, Springer-Verlag (2000) 298–372
2. Lourakis, M.I.A., Argyros, A.A.: SBA: A Software Package for Generic Sparse Bundle Adjustment. ACM Trans. Math. Software **36** (2009) 1–30
3. Agarwal, S., Snavely, N., Simon, I., Seitz, S.M., Szeliski, R.: Building Rome in a Day. In: Proc. ICCV, IEEE (2009)
4. Fitzgibbon, A., Zisserman, A.: Automatic camera recovery for closed or open image sequences. (1998) I: 311
5. Steedly, D., Essa, I., Delleart, F.: Spectral partitioning for structure from motion. In: ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision, Washington, DC, USA, IEEE Computer Society (2003) 996
6. Ni, K., Steedly, D., Dellaert, F.: Out-of-core bundle adjustment for large-scale 3d reconstruction. In: ICCV07. (2007)
7. Farenzena, M., Fusiello, A., Gherardi, R.: Stucture-and-motion pipeline on a hierarchical cluster tree. In: Proceedings of the IEEE Internation Workshop on 3-D Digital Imaging and Modeling. (2009) 1489–1496
8. Estrada, C., Neira, J., Tardos, J.D.: Hierarchical SLAM: real-time accurate mapping of large environments. IEEE Transactions on Robotics (2005) 588–596
9. Sibley, G., Mei, C., Reid, I., Newman, P.: Adaptive relative bundle adjustment. In: Proceedings of Robotics: Science and Systems, Seattle, USA (2009)
10. Montiel, J., Civera, J., Davison, A.: Unified inverse depth parametrization for monocular slam. In: Proceedings of Robotics: Science and Systems, Philadelphia, USA (2006)
11. Hartley, R.I.: Lines and points in three views and the trifocal tensor. IJCV **22** (1997) 125–140
12. Pagel, F.: Robust monocular egomotion estimation based on an iekf. In: CRV '09: Proceedings of the 2009 Canadian Conference on Computer and Robot Vision, Washington, DC, USA, IEEE Computer Society (2009) 213–220
13. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Second edn. Cambridge University Press, ISBN: 0521540518 (2004)
14. Koch, K.R.: Parameter estimation and hypothesis testing in linear models. Springer (1988)
15. Mikhail, E.M., Ackermann, F.: Observations and Least Squares. University Press of America (1976)
16. Faugeras, O.D., Mourrain, B.: On the geometry and algebra of the point and line correspondences between n images. In: ICCV. (1995) 951–956
17. Heuel, S.: Uncertain Projective Geometry: Statistical Reasoning for Polyhedral Object Reconstruction. Volume 3008 of LN in Computer Science. Springer (2004)
18. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision **60** (2004) 91–110
19. Clipp, B., Zach, C., Frahm, J.M., Pollefeys, M.: A New Minimal Solution to the Relative Pose of a Calibrated Stereo Camera with Small Field of View Overlap. (2009)