# HIERARCHICAL STRUCTURE FROM MOTION COMBINING GLOBAL IMAGE ORIENTATION AND STRUCTURELESS BUNDLE ADJUSTMENT

A. Cefalu*, N. Haala, D. Fritsch

Institute for Photogrammetry, University of Stuttgart, 70174 Stuttgart, Germany
alessandro.cefalu@ifp.uni-stuttgart.de

**Commission I, WG I/9**

**KEY WORDS:** Structure from Motion, Global Image Orientation, Structureless Bundle Adjustment, Hierarchical Image Orientation

**ABSTRACT:**

Global image orientation techniques aim at estimating camera rotations and positions for a whole set of images simultaneously. One of the main arguments for these procedures is an improved robustness against drifting of camera stations in comparison to more classical sequential approaches. Usually, the process consists of computation of absolute rotations and, in a second step, absolute positions for the cameras. Either the first or both steps rely on the network of transformations arising from relative orientations between cameras. Therefore, the quality of the obtained absolute results is influenced by tensions in the network. These may e.g. be induced by insufficient knowledge of the intrinsic camera parameters. Another reason can be found in local weaknesses of image connectivity. We apply a hierarchical approach with intermediate bundle adjustment to reduce these effects. We adopt efficient global techniques which register image triplets based on fixed absolute camera rotations and scaled relative camera translations but do not involve scene structure elements in the fusion step. Our variant employs submodels of arbitrary size, orientation and scale, by computing relative rotations and scales between - and subsequently absolute rotations and scales for - submodels and is applied hierarchically. Furthermore we substitute classical bundle adjustment by a structureless approach based on epipolar geometry and augmented with a scale consistency constraint.

## 1. INTRODUCTION

Research on Structure from Motion (SfM), i.e. automatically solving the task of image orientation and scene structure computation for sets of visually connected images, has made large progress over the past decades. Generally speaking, the process consists of three steps: (1) visual connections are found, usually using distinct image features which are matched between images based on local texture descriptors and subsequently validated geometrically, (2) provide approximate solutions for scene structure and/or camera orientation and (3) optimize the approximate solution. The first step is crucial for success and is one the main bottlenecks in terms of time consumption. However, in this paper we will focus on the latter two steps, which are of no less importance for the overall efficiency.

The majority of approaches are of incremental nature, i.e. starting from a pair or triplet of images, a number of images is added to the current model (2) and the result is optimized (3). The process repeats until the full model is solved. This approach benefits from careful progression and can be successful in complicated and large scenarios (Agarwal et al. 2011; Snavely et al. 2006; Snavely et al., 2008; Jeong et al. 2012). However, the optimization is carried out by bundle adjustment. The resulting system is highly nonlinear and solving it is computationally intensive. Its' repetitive use on a growing number of images is one of the drawbacks of incremental approaches. The problem of loop closure poses another challenge especially for larger image sequences. The orientation error built up during reconstruction may become large enough to prevent the geometric establishment of a connection between images. This also directly induces a dependency on the order in which images are added to the model. Hierarchical approaches separate the data into smaller subsets, which are solved separately and fused in subsequent steps following the tree like structure of connected images clusters. These approaches work on problems of moderate size during intermediate reconstructions and thereby reduce the overall workload. Furthermore, the orientation error build-up is reduced, leading to better loop closing behaviour.

Global approaches to SfM on the other hand solve step (2) for all cameras simultaneously by exploiting relative transformation information from step (1), aiming at a single and final execution of the optimization (3). The majority of approaches solves absolute rotations in a first step and separately estimates scene structure and / or camera positions in a subsequent step. Apart from efficiency, an advantage is the even distribution of orientation errors over the whole data, which effectively eliminates the loop closing problem. On the other hand, the main difficulty lies in finding robust solutions, as the underlying set of relative relations may be corrupted by inaccuracies and erroneously established connections.

We present our SfM-approach which bases on the use of structureless bundle adjustment in combination with global image orientation techniques which are applied hierarchically to submodels instead of cameras.

## 2. RELATED WORK

An overview on classical bundle adjustment can be found in (Triggs et al., 2000) which covers many efficiency related topics. A very efficient structureless approach has been presented by (Rodriguez et al., 2011a, 2011b). Only two-view epipolar constraints, encapsulated in a measurement matrix (Hartley, 1998) are used. Factorization reduces the system to

*Corresponding author

9x9 per image pair, resulting in remarkable speed up. In (Steffen et al., 2010) epipolar constraints are combined with trifocal constraints to overcome the problem of distance ambiguity for collinear camera stations which is induced by the exclusive use of epipolar constraints. (Indelman, 2012a, Indelman et al., 2012b) derives a scale consistency constraint from addition of scaled observation vectors and camera baselines in a three-view scenario by reformulating rank conditions of the equation system. (Rodriguez et al., 2011a, 2011b) and (Indelman et al., 2012b) demonstrate that accurate solutions can be achieved without introduction of additional unknowns, which is also the case for our implementations (Cefalu et al., 2016). Apart from a reduced number of unknowns, a further advantage is found in the absence of numerical problems due to points at far distance.

Early work on global image orientation has been conducted by (Govindu, 2001). In the majority of existing approaches absolute camera rotations are estimated and held fix for a subsequent separate estimation of camera poses. Thorough discussion on rotation averaging and distance measures in SO(3) is given in (Hartley et al., 2013). Based on quaternions (Govindu, 2001) suggests a linear least squares solution which in practice suffers from missing orthonormality constraints. In (Martinec & Pajdla, 2007) the constraints are enforced in a subsequent step. (Arie-Nachimson et al., 2012) include these constraints while using semi-definite programing. The approach is adopted and enhanced in (Reich & Heipke, 2016). The Weiszfeld algorithm under $L_1$ norm is applied in (Hartley et al., 2011) to increase robustness. (Chatterjee & Govindu., 2013) refine their $L_1$ solution in an iteratively reweighted least squares process in which they incorporate the Huber estimator (Huber, 1964), i.e. an M-estimator. Other approaches aim at detecting and eliminating incorrect relative rotations based on the concept of cycle consistency. The Bayesian inference approach of (Zach et al., 2010) is picked up in (Moulon et al., 2013) and combined with the cycle length weighting of (Enqvist et al., 2011). The latter can be categorized as a spanning tree approach in which paths are drawn from the camera graph following some heuristic and used to identify inconsistent graph edges (Govindu, 2006; Olsson & Enqvist, 2011, Cefalu & Fritsch, 2014b). A survey on various methods is presented in (Tron et al., 2016). (Wilson et al., 2016) investigates problem sources in rotation averaging and suggests reducing to smaller sub-problems, very much in the sense of our solution to model fusion using absolute rotations for models.

Given known absolute rotations for cameras, various approaches exist to estimate absolute positions. Some approaches simultaneously estimate scene structure. A reformulation of the collinearity constraint is used in (Kahl, 2005) to simultaneously estimate camera poses and scene structure under the $L_\infty$ norm. The concept is picked up by e.g. (Olsson & Enqvist, 2011) and (Martinec & Pajdla, 2007). The latter suggest a reduction of the number of reconstructed points by selecting four representative points per image pair to reduce the computational workload. (Arie-Nachimson et al., 2012) makes use of point observations, but reformulates the epipolar constraint by replacing relative orientations with absolute ones. With fixed rotations, the resulting system is linear and can be solved efficiently. Essentially, this approach equals structureless bundle adjustment without additional constraints. The number of unknowns is drastically reduced, but collinear camera distributions cannot be handled satisfyingly. The same holds for early methods using only baseline estimates (Govindu, 2001; Govindu, 2006). Implicit use of point triangulation is made in (Cui et al., 2015) which overcomes this problem. (Reich & Heipke, 2016) further robustify the approach. The problem can also be avoided by the use of trifocal tensors (oriented image triplets) as relative scales are encapsulated (Moulon et al., 2013; Jiang et al., 2013; Özyesil & Singer, 2015). In (Jiang et al., 2013) a linear model is used which is closely related to the scale consistency constraint used in our structureless bundle adjustment.

Work on hierarchical reconstruction is given in (Nister, 2000; Farenzena et al., 2009; Gherardi et al., 2010; Ni & Dellaert, 2012; Havlena et al. 2009, Chen et al., 2016), to name a few. Approaches differ in how the tree structure is defined and in strategies for balancing the tree, i.e. maintaining a homogenous size of partial reconstructions and avoidance of unnecessary workload. However, usually 3D points are used for registration of models. We are not aware of approaches within this category which rely on exterior camera orientation.

Our approach builds up on trifocal tensor based methods and makes a step towards combining some of the previously mentioned strategies. Scene structure is only taken into account when solving single image triplets. Our variant fuses models of arbitrary image number, scale and orientation and is applied repeatedly to locally neighbouring models. The resulting process creates a hierarchy of overlapping models and allows intermediate optimization including camera calibration. Global rotation computation is applied to models instead of cameras and camera rotation computation is broken down to a single rotation averaging problem. To refine fused models we apply structureless bundle adjustment. Section 3.5 explains the scale consistency constraint used in our current implementations, which to our knowledge has not been used in this context. The development of this strategy was mainly driven by the desire for intermediate adjustment. Insufficient knowledge on calibration parameters may result in drifts which in some cases are hard to compensate with a single final adjustment after a fully global orientation of images. Examples demonstrating such cases are given in section 4, which also contains a numerical evaluation on public benchmark data sets.

## 3. APPROACH

### 3.1 Process Chain

We start by detecting image features using SIFT (Lowe, 2004) and match over image pairs based on the descriptors. Geometric validation is carried out using parallax based pre-filtering (a histogram based variant of Cefalu et al., 2014a), followed by a standard RANSAC (Fischler & Bolles., 1981) process. Here, the 5-Point-algorithm (Nister, 2004) is applied, using approximate information on the camera, e.g. from EXIF headers. In order to stabilize relative orientation estimates for pairs with strong perspectives, we establish point tracks throughout the complete data, thereby creating additional point matches which have not been revealed during descriptor matching. The tracks are furthermore used to select well distributed stable points in every image to reduce the total amount of points. The results are revalidated in a second run of the 2-view RANSAC to make use of new established matches. Weak pairs are refused based on minimal point count (20) and overlap of the convex hull of matching features (5%). The connectivity graph is thinned out further by reducing to the 10 strongest connections per image in terms of point support. A test on rotational consistency follows. In contrast to (Zach et al., 2010), we consider only rotational errors of image triplets, which makes the process fast since no longer loops have to be found. Every triplet is evaluated by a simple 3σ outlier test

against the error budget of its' local triplet neighbourhood. The edges of remaining triplets define the new image connectivity graph. The test is repeated until no further edges are filtered out. The remainder of image triplets is ranked based on the number of threefold point observations and, in contrast to (Moulon et al., 2013), intersection angles between observed points. The latter measure influences the ranking towards a preference of triplets with longer base lines. Following the rank order (best first), we reconstruct triplet models (see section 3.2) only if they consist of at least one edge which has not been part of an already successfully solved triplet. Suspicious triplets are rejected. If rejections occurred, the process is repeated on the remainder of untested triplets. Edges of the connectivity graph which are not part of a successfully reconstructed triplet are removed. Depending on the overall connectivity, this sub-selection reduces the number triplets to roughly 30% - 60%.

This set of triplets represents the lowest level of our submodel hierarchy. The following hierarchical fusion process consists of three repeatedly applied steps. First, target models are defined by local growth of all source models (models of the current highest level, section 3.3). Second, the target models are fused using the corresponding source models as described in section 3.4 and eventually refined using our structureless bundle adjustment implementation (section 3.5).

### 3.2 Triplet Reconstruction

Every Triplet selected for reconstruction is solved by first computing absolute rotations for the three cameras and their local neighbourhood (similarly to our approach for absolute model rotations in section 3.4.2). The inclusion of the neighbourhood is actually not necessary but can smooth possible inaccuracies inherent in a single triplet. Similar to (Moulon et al., 2013) we solve the position of the cameras within a RANSAC framework, while keeping the rotations (of the triplet) fixed and forcing correct chirality, i.e. points must triangulate to the viewing direction of the cameras. However, we use a different model. Let $q = RK^{-1}p$ denote the ray cast in viewing direction from a projection centre $T$ to an observed point, where $p$ is the point's projection on the image plane in homogeneous coordinates, $R$ is the camera rotation and $K$ the camera matrix (see also section 3.5). Further, let $\lambda$ be a scale factor which scales the ray $q$ to end in the observed point in object space (Figure 1). For three cameras $i$, $j$, and $k$, vector addition yields:

$$\begin{aligned} T_k - T_i &= \lambda_i q_i - \lambda_k q_k \\ T_j - T_i &= \lambda_i q_i - \lambda_j q_j \\ T_k - T_j &= \lambda_j q_j - \lambda_k q_k \end{aligned} \quad (1)$$

In every RANSAC iteration a solution is generated from two sampled threefold observations, solving the above system using an interior point method, forcing positivity of $\lambda$. The position estimates are used to triangulate all points of the triplet and carry out the consensus test on chirality and reprojection errors. The final RANSAC result is further refined using structureless bundle adjustment. Triplets resulting in low inlier rates or high error budget are rejected. Taking scene structure into account in this step significantly improves robustness of the reconstruction process, as it eases identification and removal of outlier correspondences and therefore significantly improves the quality of orientation estimates.

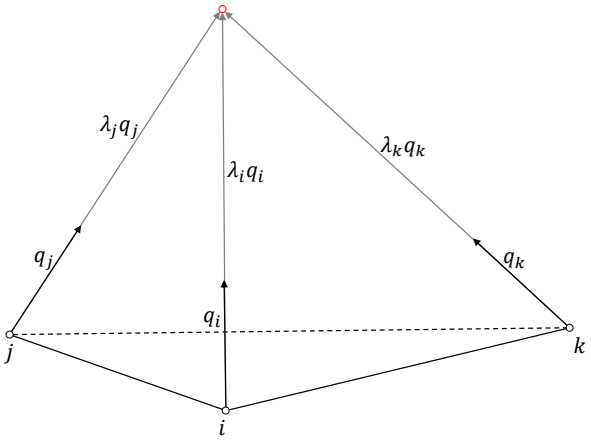


Figure 1. Three cameras $i$, $j$ and $k$ casting rays $q$ towards an observed object point (red). When scaled by factors $\lambda$, the rays should intersect in this point.

### 3.3 Model Growth

The set of triplets which have been solved in the previous step are the starting level of our iterative fusion of models. We refer to models of the current level as source models and to the models of the next level as target models. For every source model we perform a growth step, which defines a corresponding target model. All source models having at least one edge in common with the currently growing source model define the set of images of the resulting target model and will be used for fusion. Further source models are added if they consist solely of this set of images. The latter step ensures that all possible triplets are used in the following fusion step, but is of minor importance on higher hierarchy levels. We remove duplicate target models and those fully being part of others (e. g. at borders of the reconstruction). The resulting set of target models is a set locally grown models, overlapping in the sense of sharing identical cameras. As every edge of the camera graph remains represented, loop closings are preserved.

The defined target models are created by fusion of the corresponding source models and eventually optimized via bundle adjustment in a subsequent step. The results serve as source models for the next stage of local growth, while models of a size of 15 or more images are not grown further. Instead they are passed through to the next level, unchanged. If a single model is created, a final adjustment terminates the process. Otherwise, when all source models have reached maximum size, the local growth and fusion is replaced by a global fusion of all source models and a final adjustment is carried out.

### 3.4 Model Fusion

Given a set of source models which are to be fused to a target model, we apply a three step procedure to fuse the models. First, relative rotations and scales between the models are computed from groups of camera rotations and base length estimates respectively. In a second step, the relative estimates are used to derive absolute rotations and scales for the models. Finally, the third step separately solves absolute camera rotations and positions.

**3.4.1 Relative Model Rotations and Scales:** In the same way as images may be represented by a graph of relative orientations, we can build a graph of relative orientations between source models being fused (Figure 2). We first compute relative rotations between source models using the rotations of all cameras shared between models. Let $R_n^i$ denote the rotation of a camera $n$ in a model $i$. With $Q$ being the set of $N$ cameras shared by the models $i$ and $j$, the corresponding rotations are stacked to $\bar{R}^i = stack(\{R_n^{i'}\}), n \in Q$ and $\bar{R}^j$ accordingly. The two resulting matrices are of size $3N \times 3$. We apply Kabsch's algorithm (Kabsch, 1976), which is also used in Procrustes analysis. With

$$C = \bar{R}^{j'}\bar{R}^i \tag{2}$$

and $C = UDV'$ (singular value decomposition), we compute the relative rotation as:

$$R_{j,i} = Udiag(1,1,sign(\det(UV')))V' \tag{3}$$

Further, we compute a relative scale between source models. For every possible image pair in Q, an estimate can be derived from the ratio of the corresponding base lengths in the two models. We use the mode of a kernel density estimation (KDE) as the final relative scale $s_{j,i}$.

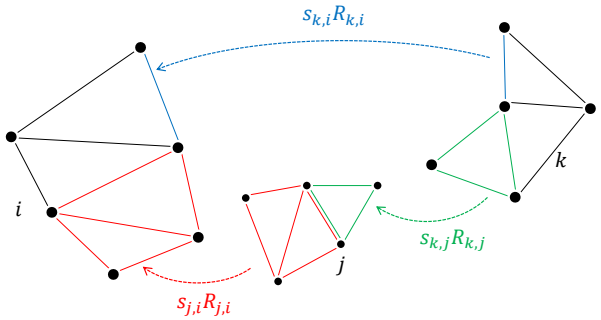


Figure 2. Sketch of a group of three models $i$, $j$ and $k$. Dots indicate cameras. Bases shared by different models are marked by colour. Relative rotations and scales between the models are indicated by arrows.

**3.4.2 Absolute Model Rotations and Scales:** The desired absolute rotations and scales of the source models are derived from multiple paths (kinematic chains) drawn at random from the source model graph. We start at random nodes (models) and guide the selection of edges (relative relations between models) to be roughly evenly distributed. We establish a common rotational frame for the paths, by setting the mean rotation (see section 3.4.3) of every path to identity, i.e. applying the inverse mean rotation. In a similar way the mean scale of a path is set to 1. As a result we obtain multiple estimates for the absolute rotation $R_i$ and scale $s_i$ of every source model $i$. Single solutions can be computed independently. We use KDE again for the scales and search for the mode. For the rotation $R_i$, we use $L_2$ chordal mean rotation computation with iterative reweighing.

**3.4.3 Mean Shift Single Rotation Averaging:** A $L_2$ mean rotation $R$ under chordal metric (Hartley et al., 2013) for a set of $n$ rotations can be computed as in (3), except:

$$C = \sum_n R_n \tag{4}$$

In order to robustify the resulting estimate against outliers, we compute an initial guess of $R$ and use the chordal distance $d_n$ between $R$ and $R_n$ to derive weights for subsequent iterations. As weighting function we use a Gaussian kernel with a bandwidth of the current weighted rms σ.

$$w_n = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{d_n}{\sigma}\right)^2} \tag{5}$$

Accordingly, we rewrite $C = \sum_n w_n R_n$ and apply (3). We iterate the process until convergence or at most ten times. The process essentially resembles a mean shift algorithm. A three sigma cut-off is added to fully suppress strong outliers.

**3.4.4 Absolute Camera Positions:** Given an absolute rotation $R_i$ and an absolute scale $s_i$ are known for every source model $i$, the absolute poses $T_n$ and $T_m$ of two cameras $n$ and $m$ can be expressed as (6). The full equation system is formed by considering all visually connected image pairs in all source models participating in the fusion of the target model.

$$s_i R_i\left(T_n^i - T_m^i\right) = T_n - T_m \tag{6}$$

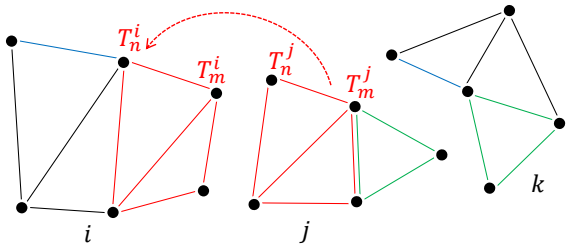


Figure 3. Camera positions in a target model are derived from corresponding base lines in rotated and scaled source models $i$, $j$ and $k$. For simplification only one camera pair is included in the sketch.

In order to robustify against incorrect bases we apply iterative reweighting using three weighting functions. A first weight $w_m$ is derived from the number of matches between the images and stays fix during iterations. A second weight evaluates the a posteriori Euclidian position residual using the Hampel estimator (Hampel et al., 1986) with thresholds at $1\sigma$, $2.5\sigma$ and $4\sigma$. A third weight for directional discrepancy is computed as the scalar product between the (normalized) left- and right-hand sides of (6), with negative values set to zero. Multiplication yields the final weights. Standard weighted least squares is used to solve the system.

**3.4.5 Absolute Camera Rotations and Intrinsic Camera Parameters:** By applying the absolute source model rotations to the single cameras, we can obtain one or more solutions for every camera rotation. In the latter case, we again apply the single rotation averaging described in section 3.4.3. As initial values for the subsequent bundle adjustment, intrinsic camera parameters are 'fused' by taking the median of every parameter over the set of source models which have been used for fusion.

### 3.5 Structureless Bundle Adjustment

The results of the source model fusion step are utilized to initialize a structureless bundle adjustment. Here, the adjustment takes place without use of 3D structure elements by founding the system of equations on epipolar constraints instead of collinearity equations. We have improved our approach presented in (Cefalu et al., 2016), the major difference being the

additional use of a scale consistency constraint. Though satisfying results could be achieved with our previous implementation for many datasets, this augmentation significantly improved the robustness of our approach for two reasons. First, collinear camera positions form a degenerate case for epipolar geometry, as distances between cameras become ambiguous. As our approach to image orientation relies on the quality of base line estimation, the exclusive use of epipolar geometry may induce disadvantageous instabilities, though this is rarely the case. Second, and to our experience of much greater importance, human made objects often exhibit regularly distributed repetitive structures. As a result, mismatches of feature descriptors occur more frequently and may often not be identified by epipolar distance. The scale consistency constraint is violated if rays do not intersect in a single point and thereby helps, not only in disambiguating distances between cameras, but also in identifying outliers.

As in section 3.2, let $p$ be the projection of a point onto an image plane in homogenous coordinates. We define its' pendant corrected for lens distortions using the Brown model (Brown, 1966) as $\bar{p} = p + \Delta p_{rad} + \Delta p_{tan}$. We compute the epipolar distance in image $i$ for the same point observed in the images $i$ and $j$ as:

$$d_{j,i} = \frac{\bar{p}_i' l_{j,i}}{\sqrt{{l_{j,i}}_{(1)}^2 + {l_{j,i}}_{(2)}^2}} \tag{7}$$

Where $l_{j,i} = K_i^{-\prime} R_i' [T_j - T_i]_\times R_j K_j^{-1} \bar{p}_j$ is the formulation for the epipolar line in image $i$ using absolute camera rotations $R$ and positions $T$. The camera matrix $K$ encapsulates the camera constant and the principal point. Normalization is necessary to avoid scale dependency, which would cause the camera stations to collapse to a single point. Moreover, the variant used above expresses the residual in pixel metric. However, it results in $d_{j,i} \neq d_{i,j}$ and we observe increased quality of the results when both projection directions are used.

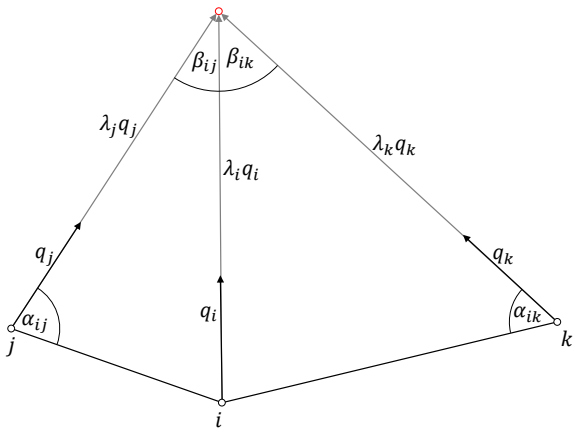


Figure 4. Three cameras $i$, $j$ and $k$ casting rays $q$ towards an observed object point as in Figure 1. Here, angles $\alpha$ between base lines and image rays and angles $\beta$ for intersection angles between observations are added.

We further redefine $q = RK^{-1}\bar{p}$ to include distortion corrections. For readability we write $B$ for the base between two cameras in the further context. Moreover, we introduce angles $\alpha$ between base lines and image rays and $\beta$ for intersection angles between rays (Figure 4). For a triangle formed by two images $i$ and $j$ and a point observed in both images, we may express the distance from projection centre of image $i$ to the point using the law of sines:

$$\|\lambda_i q_i\| = \|B_{j,i}\| \frac{\sin(\alpha_{j,i})}{\sin(\beta_{j,i})} \tag{8}$$

The relation between the sine and the cross product allows the substitutions:

$$\sin(\alpha_{j,i}) = \frac{\|B_{j,i} \times q_j\|}{\|B_{j,i}\|\|q_j\|}, \sin(\beta_{j,i}) = \frac{\|q_i \times q_j\|}{\|q_i\|\|q_j\|} \tag{9}$$

The distance must be equal when computed using an observation of a third image $k$, which leads to the constraint:

$$\|\lambda_i q_i\| = \frac{\|B_{j,i} \times q_j\|}{\|q_i \times q_j\|} \|q_i\| = \frac{\|B_{k,i} \times q_k\|}{\|q_i \times q_k\|} \|q_i\| \tag{10}$$

Setting the ratio of the two right hand side expressions to one, formulates our functional model, which is free of the unknown $\lambda$. Subtraction is also possible, but suffers from dependency of the overall scale. In any case, the errors are not in pixel metric and are therefore handled in the sense of a variance component analysis. Moreover, the number of equations which could be used may be very high for long point tracks. Therefore we use only one equation for every ray $q$, derived from two other cameras in the order of appearance. As (Rodriguez et al., 2011a, 2011b) and (Indelman, 2012b) we solve the system without additional unknowns. Our implementation uses weighted (Hampel estimator) nonlinear least squares in a Levenberg-Marquardt framework.

## 4. EXPERIMENTAL RESULTS

We numerically evaluate the accuracy of our overall approach on the six benchmark datasets published in (Strecha et al., 2008) and compare our results to those reported by other authors. For these datasets ground truth is available for image orientation as well as two camera constants and principal point. The images have been corrected for radial distortion. However, we do not use the given intrinsic parameters but initialize our parameters from the EXIF headers of the dataset *FountainR25* (for which unfortunately no ground truth orientation is available) and solve for all parameters implemented in our camera model. Furthermore, we have deactivated the hierarchical procedure for this test to assess the quality of our fusion approach. I.e. image triplets are fused directly to a final model and refined by a single final adjustment.

For evaluation we transform our resulting camera stations onto the ground truth camera stations using a seven parameter transformation. Table 1 and Table 2 summarize the results before adjustment, while Table 3 and Table 4 present the final

| Mean position error [mm] before final adjustment | | | |
|---|---|---|---|
| | Ours EXIF | [1] EXIF | [2] EXIF |
| HJ P8 | 48.0 | | |
| E P10 | 141.3 | | |
| F P11 | 38.0 | 35.0 | 53.0 |
| C P19 | 602.3 | 428.0 | |
| HJ P25 | 94.7 | 83.0 | 106.0 |
| C P30 | 617.8 | 1312.0 | 1158.0 |

Table 1. Comparison of mean position errors before final bundle adjustment: [1] Reich & Heipke, 2016, TE-SI, [2] Jiang et al., 2013.

| Mean rotation error [°] before final adjustment | | | |
|---|---|---|---|
| | Ours EXIF | [1] EXIF | [2] EXIF |
| HJ P8 | 0.310 | | |
| E P10 | 0.671 | | |
| F P11 | 0.730 | 0.249 | 0.517 |
| C P19 | 1.554 | 0.647 | |
| HJ P25 | 0.491 | 0.206 | 0.573 |
| C P30 | 1.390 | 0.583 | 1.651 |

Table 2. Comparison of mean rotation errors before final bundle adjustment: [1] Reich & Heipke, 2016, GO-SDP-ls, [2] Jiang et al., 2013.

| Mean rotation error [°] after final adjustment | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Ours EXIF | [1] GT | [1] EXIF | [2] EXIF | [3] EXIF | [4] EXIF | [5] EXIF |
| HJ P8 | 0.344 | | | | | | |
| E P10 | 0.118 | | | | | | |
| F P11 | 0.065 | 0.024 | 0.420 | 0.027 | 0.195 | 0.035 | 0.035 |
| C P19 | 0.207 | | | 0.076 | | | |
| HJ P25 | 0.203 | 0.045 | 0.348 | 0.021 | 0.188 | 0.127 | 0.128 |
| C P30 | 0.227 | | | 0.039 | 0.480 | 0.139 | 0.158 |

Table 3. Comparison of mean rotation errors after final bundle adjustment: [1] Arie-Nachimson et al., 2012; [2] Reich & Heipke, 2016, GO-SDP-ls; [3] Jiang et al., 2013; [4] Chen et al., 2016; [5] Gherardi et al., 2010 as reported by [4].

results. Entries are left empty when no result is reported on a dataset. For authors reporting different variants of their approaches, a representative one is selected and indicated in the table. The tables also state whether EXIF information or ground truth (GT) was used for camera initialization. The number of images is given in the names of the datasets (full names are given in Table 4).

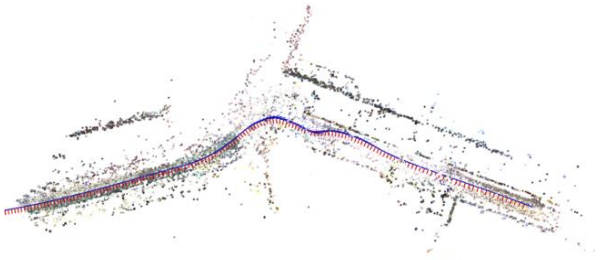

Figure 5. Partial reconstruction of an image based mobile mapping dataset (first 156 images, Cavegn et al., 2016). The approach is able to handle collinear and forward looking camera constellations.

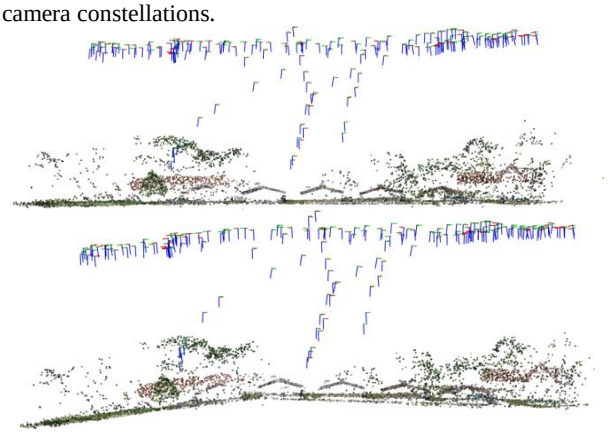

Figure 6. Top: Result for a UAV flight over a farm (185 images) computed with intermediate camera calibration. Bottom: Result with camera calibration activated only during the final adjustment. An area with weak point connectivity causes a part of the reconstruction to drift off.

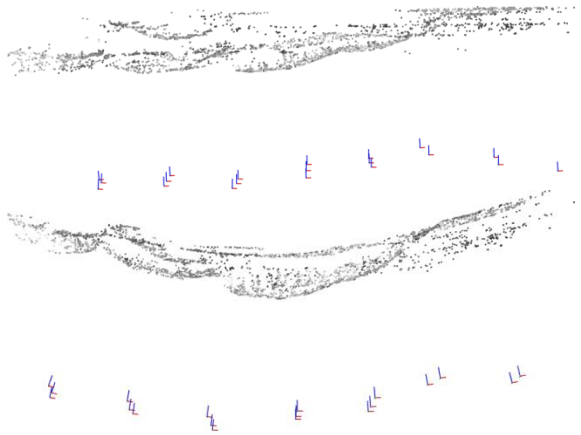

Figure 7. Partial reconstruction results for 20 images of a granite relief captured with an industrial camera. The used lens causes radial distortion of ~250 pixels in the image corners at a sensor size of 1600x1200 pixels. Top: Result with intermediate camera calibration. Bottom: Result with camera calibration activated only during the final adjustment. The adjustment converged before lens distortion effects could be fully compensated.

The examples in Figures 5 to Figure 8 have been computed with the hierarchical approach as described in the paper. The sparse point clouds have been triangulated after final adjustment for visualization purpose. Camera stations are depicted as coordinate frames (red, green and blue as X, Y and Z, the latter points in viewing direction). An example for a relatively large dataset with many narrow camera stations is given in Figure 8 (Farenzena et al., 2009). Figure 5 exemplarily demonstrates the ability of our approach to handle collinear and forward looking cameras at the example of an image based mobile mapping sequence (Cavegn et al., 2016). Figures 6 and Figure 7 show two cases in which the effect of intermediate camera calibration is evident.

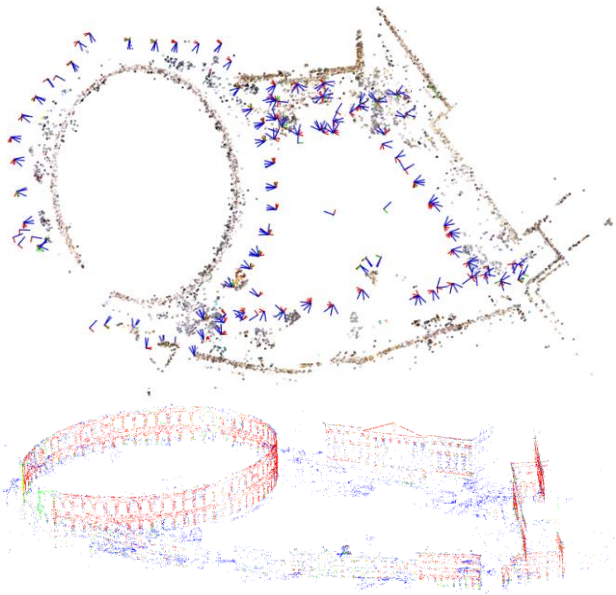

Figure 8. Result for the Piazza Bra dataset (Farenzena et al., 2009). The largest of five resulting models is displayed (287 of 329 images). Image connectivity cut offs and drop off in reconstruction quality occur in areas with weak observation redundancy, as indicated by the colors in the lower image (blue, green, yellow and red for 2, 3, 4 and ≥5 observations). However, narrow baselines are handled succesfully.

| Mean position error [mm] after final adjustment | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ours EXIF | [1] GT | [2] GT | [3] GT | [4] GT | [3] EXIF | [4] EXIF | [5] EXIF | [6] EXIF | [7] EXIF | [8] EXIF |
| HerzJesu P8 | 4.3 | 3.5 | 3.9 | | | | | | | | |
| Entry P10 | 10.1 | 5.9 | 6.9 | | | | | | | | |
| Fountain P11 | 2.6 | 2.5 | 2.2 | 4.8 | 2.5 | 27.0 | 7.0 | 14.0 | 7.0 | 5.4 | 5.4 |
| Castle P19 | 28.5 | 25.6 | 76.2 | | | | | | 81.0 | | |
| HerzJesu P25 | 5.4 | 5.3 | 5.7 | 7.8 | 5.0 | 52.0 | 26.2 | 64.0 | 13.0 | 15.6 | 15.6 |
| Castle P30 | 36.2 | 21.9 | 66.8 | | 21.2 | | 166.7 | 235.0 | 44.0 | 143.8 | 126.7 |

Table 4. Comparison of mean position errors after final bundle adjustment: [1] Moulon et al., 2013; [2] Olsson & Enqvist, 2011 as reported by [1]; [3] Arie-Nachimson et al., 2012; [4] Cui et al., 2015, $L_1$; [5] Jiang et al., 2013; [6] Reich & Heipke, 2016, TE-SI; [7] Chen et al., 2016; [8] Gherardi et al., 2010 as reported by [7].

## 5. CONCLUSION

We have presented our approach to Structure from Motion which combines hierarchical reconstruction, global image orientation techniques and structureless bundle adjustment. Though initialized with EXIF information only (and hierarchical processing turned off), the final position results achieved on the benchmark datasets are comparable to results of other authors who initialized with given camera parameters. Our rotation results are reasonably close to the ground truth, considering the different underlying camera models. The process successfully handles narrow baselines as well as collinear and forward looking scenarios. Since scene structure is not utilized in the majority of our approach, it is insensitive to noisy surface reconstructions or numerical problems induced by far distant points. The hierarchical progression and intermediate camera calibration help bridging areas of weaker connectivity. However, the overall robustness will remain subject to future work. Currently, our work focuses on reducing the computational effort by a model selection strategy, as a balancing of the hierarchy is not tackled at present. An inclusion of the absolute source model orientations in the adjustment, as well as local optimization techniques, could be beneficial to the convergence behaviour and may become topics of near future work.

## REFERENCES

Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S. M., & Szeliski, R., 2011. Building rome in a day. *Communications of the ACM*, *54*(10), 105-112.

Arie-Nachimson, M., Kovalsky, S. Z., Kemelmacher-Shlizerman, I., Singer, A., & Basri, R., 2012. Global motion estimation from point matches. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT)*, Second International Conference on (pp. 81-88). IEEE.

Brown, Duane C., 1966. Decentering distortion of lenses. *Photometric Engineering* 32(3), pp. 444-462.

Cavegn, S., & Haala, N., 2016. Image-Based Mobile Mapping for 3D Urban Data Capture. *Photogrammetric Engineering & Remote Sensing*, *82*(12), 925-933.

Cefalu, A., Fritsch, D., & Haala, N., 2014a. Multivariate Kerndichteschätzung zur Filterung automatischer Punktzuordnungen. In *Gemeinsame Tagung der DGfK, der DGPF, der GfGI und des GiN (DGPF Tagungsband 23 / 2014)*

Cefalu, A., & Fritsch, D., 2014b. Non-Incremental Derivation of Scale and Pose from a Network of Relative Orientations. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *40*(3), 53.

Cefalu, A., Haala, N., & Fritsch, D., 2016. Structureless Bundle Adjustment with Self-Calibration Using Accumulated Constraints. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3-9.

Chatterjee, A., & Madhav Govindu, V., 2013. Efficient and robust large-scale rotation averaging. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 521-528).

Chen, Y., Chan, A. B., Lin, Z., Suzuki, K., & Wang, G., 2017. Efficient tree-structured SfM by RANSAC generalized Procrustes analysis. *Computer Vision and Image Understanding*, *157*, 179-189.

Cui, Z., Jiang, N., Tang, C., & Tan, P., 2015. Linear global translation estimation with feature tracks. In *Proceedings of the British Machine Vision Conference*

Enqvist, O., Kahl, F., & Olsson, C., 2011. Non-sequential structure from motion. In *Computer Vision Workshops (ICCV Workshops)*, IEEE International Conference on (pp. 264-271). IEEE.

Farenzena, M., Fusiello, A., & Gherardi, R., 2009. Structure-and-motion pipeline on a hierarchical cluster tree. In *Computer Vision Workshops (ICCV Workshops), IEEE 12th International Conference on* (pp. 1489-1496). IEEE.

Fischler, M. A., & Bolles, R. C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381-395.

Gherardi, R., Farenzena, M., & Fusiello, A., 2010. Improving the efficiency of hierarchical structure-and-motion. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on* (pp. 1594-1600). IEEE.

Govindu, V. M., 2001. Combining two-view constraints for motion estimation. In *Computer Vision and Pattern Recognition*, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on (Vol. 2, pp. II-218). IEEE.

Govindu, V. M., 2006. Robustness in motion averaging. In *Computer Vision–ACCV 2006* (pp. 457-466). Springer Berlin Heidelberg.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A., 1986. *Robust statistics: the approach based on influence functions* (Vol. 114). John Wiley & Sons.

Hartley, R., 1998. Minimizing Algebraic Error in Geometric Estimation Problem. *Proceedings of Sixth International Conference on Computer Vision*.

Hartley, R., Aftab, K., & Trumpf, J., 2011. L1 rotation averaging using the Weiszfeld algorithm. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 3041-3048). IEEE.

Hartley, R., Trumpf, J., Dai, Y., & Li, H., 2013. Rotation averaging. *International journal of computer vision*, 103(3), p267-305.

Havlena, M., Torii, A., Knopp, J., & Pajdla, T., 2009. Randomized structure from motion based on atomic 3d models from camera triplets. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (pp. 2874-2881). IEEE.

Huber, Peter J., 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics* 35(1), pp. 73-101.

Indelman, V., 2012a. Bundle adjustment without iterative structure estimation and its application to navigation. *Position Location and Navigation Symposium (PLANS), 2012 IEEE/ION*.

Indelman, V., Roberts, R., Beall, C., & Dellaert, F., 2012b. Incremental light bundle adjustment. *Proceedings of the British Machine Vision Conference (BMVC 2012)*, pp. 3-7.

Jeong, Y., Nister, D., Steedly, D., Szeliski, R., & Kweon, I. S., 2012. Pushing the envelope of modern methods for bundle adjustment. *IEEE transactions on pattern analysis and machine intelligence*, *34*(8), 1605-1617.

Kabsch, W., 1976. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, *32*(5), 922-923.

Kahl, F., 2005. Multiple view geometry and the L∞-norm. In *Computer Vision*, 2005. ICCV 2005. Tenth IEEE International Conference on (Vol. 2, pp. 1002-1009). IEEE.

Martinec, D., & Pajdla, T., 2007. Robust rotation and translation estimation in multiview reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE Conference on (pp. 1-8). IEEE.

Moulon, P., Monasse, P., & Marlet, R., 2013. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3248-3255.

Ni, K., & Dellaert, F., 2012. HyperSfM. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on* (pp. 144-151). IEEE.

Nistér, D., 2000. Reconstruction from uncalibrated sequences with a hierarchy of trifocal tensors. *Computer Vision-ECCV 2000*, 649-663.

Nistér, D., 2004. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, *26*(6), 756-770.

Olsson, C., & Enqvist, O., 2011. Stable structure from motion for unordered image collections. In: *Image Analysis* (pp. 524-535). Springer Berlin Heidelberg.

Özyesil, O., & Singer, A., 2015. Robust camera location estimation by convex programming. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2674-2683).

Triggs, B., McLauchlan, P. F., Hartley, R. I., & Fitzgibbon, A. W., 2000. Bundle adjustment—a modern synthesis. In *Vision algorithms: theory and practice.* Springer Berlin Heidelberg. pp. 298-372.

Reich, M., & Heipke, C., 2016. Convex image orientation from relative orientations. In *23rd ISPRS Congress, July 12-19 2016, Prague, Czech Republic*. Göttingen: Copernicus GmbH.

Rodriguez, A. L., López-de-Teruel, P. E., & Ruiz, A., 2011a. Reduced epipolar cost for accelerated incremental SfM. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*.

Rodriguez, A. L., López-de-Teruel, P. E., & Ruiz, A., 2011b. GEA Optimization for live structureless motion estimation. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*.

Snavely, N., Seitz, S. M., & Szeliski, R., 2006. Photo tourism: exploring photo collections in 3D. In *ACM transactions on graphics (TOG)* (Vol. 25, No. 3, pp. 835-846). ACM.

Snavely, N., Seitz, S. M., & Szeliski, R., 2008. Skeletal graphs for efficient structure from motion. In *CVPR* (Vol. 1, p. 2).

Steffen, R., Frahm J.-M., & Förstner W., 2010. Relative bundle adjustment based on trifocal constraints. *Trends and Topics in Computer Vision*. Springer Berlin Heidelberg, 2012. pp. 282-295.

Strecha, C., von Hansen, W., Gool, L. V., Fua, P., & Thoennessen, U., 2008. On benchmarking camera calibration and multi-view stereo for high resolution imagery. *Computer Vision and Pattern Recognition (CVPR), 2008 IEEE Conference on*.

Tron, R., Zhou, X., & Daniilidis, K., 2016. A survey on rotation optimization in structure from motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 77-85).

Wilson, K., Bindel, D., & Snavely, N., 2016. When is Rotations Averaging Hard?. In *European Conference on Computer Vision* (pp. 255-270). Springer International Publishing.

Zach, C., Klopschitz, M., & Pollefeys, M., 2010. Disambiguating visual relations using loop constraints. In *Computer Vision and Pattern Recognition (CVPR)*, IEEE Conference on (pp. 1426-1433). IEEE.