Non-Sequential Structure from Motion

Olof Enqvist Fredrik Kahl Carl Olsson
Centre for Mathematical Sciences, Lund University, Sweden
{olofe,fredrik,calle}@maths.lth.se

Abstract

Prior work on multi-view structure from motion is dominated by sequential approaches starting from a single two-view reconstruction, then adding new images one by one. In contrast, we propose a non-sequential methodology based on rotational consistency and robust estimation using convex optimization. The resulting system is more robust with respect to (i) unreliable two-view estimations caused by short baselines, (ii) repetitive scenes with locally consistent structures that are not consistent with the global geometry and (iii) loop closing as errors are not propagated in a sequential manner. Both theoretical justifications and experimental comparisons are given to support these claims.[1]

1. Introduction

Given a set of images with known calibration data, we want to estimate scene structure as well as camera positions. Although this problem has been studied extensively over the years, no fully satisfactory solution exists. Among the things that make this problem so challenging one can mention the high dimension of the space of unknowns and the difficulty in correctly matching features between views. Yet another challenge is the existence of repetitive or planar structures, short baselines between views or moving objects in the scene. Unlike ordinary mismatches that will cause random outliers in the data, repetitive structures can cause locally consistent geometries that do not agree with the global geometry. This can lead to two-view geometries supported by a large number of point correspondences, but not reflecting the underlying true geometry; cf. Figure 1.

1.1. Structure from Motion Approaches

Many methods for multi-view structure from motion start by estimating the geometry of two views. Often, a minimal solver is applied in combination with RANSAC, for example [14]. The three-dimensional geometry estimated from these two views is used to estimate the pose of another camera which in turn improves the quality of the reconstruction. More cameras are thus added, essentially one by one. The reconstructions are often improved using local optimization, so called bundle adjustment [18]. We will refer to this approach as sequential structure from motion. An apparent weakness of these methods is that the quality of the reconstruction might depend heavily on the choice of the initial pair. This is addressed in [20] where a heuristic approach based on covariance estimation of the structure and the CIRC criterion [22] is presented. Another weak point is the iterative process of adding new cameras. It might be that the final reconstruction is affected by the order in which cameras are added. Had the cameras been added in a different order we might have found a better reconstruction. Furthermore, due to their sequential nature these methods suffer from drift (error build-up) [2] as the error is not evenly distributed over the sequence. An automated system using this technique was presented in [16] showing impressive results of large-scale reconstructions. The system, known as BUNDLER, will be compared to our work.

A different approach is taken by the methods based on factorization. In [21] a solution for the affine camera model is provided and [17] gives an extension to perspective cameras. The missing data problem and sensitivity to outliers are major concerns for this approach and it has been the object of study in subsequent papers, e.g. [19]. Hierarchical

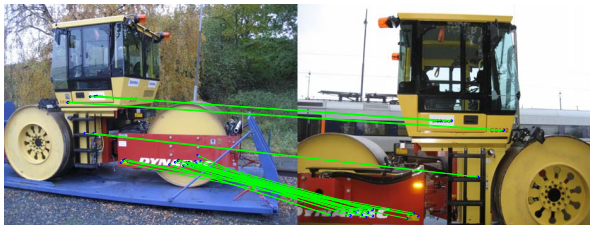

Figure 1. ROAD ROLLER. In this image pair, 28 seemingly correct correspondences (green lines) are obtained in the estimation of the epipolar geometry. Even though the epipolar geometry is plausible and perfectly valid, it does not correspond to the true geometry.

[1]This work was supported by the European Research Council (grant 209480) and the Swedish Foundation for Strategic Research.

methods [13, 5] organize images in a hierarchical cluster tree, and do the reconstruction from root to leafs.

In this paper, a non-sequential method for estimating the geometry of multiple views is suggested. Unlike sequential and hierarchical methods, our approach does not depend on a good initial solution of a partial 3D reconstruction. The method consists roughly of three parts. First the orientations of all cameras are estimated with a method being robust to low-level noise as well as completely inconsistent outlier rotations. Then the robust algorithm from [4] is used to solve the structure and motion with known orientations and finally the reconstruction is improved using bundle adjustment. Note that all views are used in the computation of camera positions and scene structure in order to get a global solution. Only the orientations are taken from the pairwise estimations and as we shall see, these are often quite accurately estimated.

Our approach belongs to the same category as [12], where camera orientations are estimated using an over-parameterized linear least-squares formulation. As is well-known, linear least squares can be sensitive to outliers in the data. Then, various heuristics for identifying 4 inlier points are applied and finally, these 4 points are fed to the convex optimization scheme to recover camera translations and the 3D coordinates of the 4 points. The requirement of identifying 4 correct matches in multiple views makes this step of their algorithm sensitive to outliers. In contrast, we remove incorrect pairwise rotations before estimating camera orientations and do not rely on identifying 4 good matches. Our system also has clear similarities to [23]. They set up a rather involved Bayesian model to detect outlier rotations based on cycle errors. This leads to an intractable optimization problem, and hence they restrict the approach to cycles of length at most six. We will show examples where this is not sufficient (see Section 4.2). Another method to discard erroneous rotation estimates is given in [7], where a random sampling strategy over spanning trees is suggested. In [15], cycles are also used as a means to estimate camera orientations. From a spanning tree they generate a set of fundamental cycles in the camera graph. For each of these cycles they compute the rotational deviation from the identity. This error is distributed over the respective rotations in the cycle to form a consistent cycle. Unlike our approach they cannot handle outliers among the relative rotations.

1.2. Contributions and System Overview

Our main contribution is a robust structure from motion system. For all parts of the pipeline, the ability to cope with outliers is in focus. From a practical point of view, this improves state-of-the-art for handling short baselines, repetitive structures and closed-loop sequences. From a theoretical point of view, our main technical contributions are (i) showing that even though the estimation of two-view epipolar geometry may be ill-posed due to a short baseline, the relative rotation can still be reliably estimated, and (ii) a new mathematical model for robust estimation of orientations based on cycles and the notion of consistency. The theoretical results are of general nature and provide motivations not only for this paper, but also for other systems based on similar concepts such as [15, 7, 23].

The outline of our system pipeline is as follows.

1. Feature extraction using SIFT [10] and matching between pairs of views.
2. Estimation of the relative orientation for pairs of views. A standard 5-point solver [14] is used in a RANSAC loop in combination with bundle adjustment.
3. Detection and removal of large errors among the relative rotations.
4. Estimation of camera orientations using the remaining relative rotations.
5. 3D reconstruction using the estimated camera orientations. The reconstruction is computed using SOCP as described in [4, 9]. Auxiliary variables are used to handle outliers.
6. Bundle adjustment to improve the 3D reconstruction.

2. Estimation with Short Baseline

Geometries with short baselines are problematic for most approaches to structure from motion. Even though matching is particularly simple, the problem of estimating the 3D structure is ill-posed. In this section, we will show that the relative rotation for a two-view geometry can still be reliably estimated. The theoretical findings are accompanied with experimental validations. The results provide strong motivations for our approach as well as similar methods to structure from motion. In particular, the mathematical model of rotational consistency presented in the next section is based on this fact.

Let us first look at estimating the relative orientation of two cameras with the same camera center when there is no noise. Allowing points at infinity, any translation direction will do so the problem is not well-posed. However, as the main theorem of this section shows, it is only in some rare degenerate configurations that the rotation is not uniquely determined.

Before stating the theorem, we recall the definition of an essential matrix. Consider two views of the same scene. Let x and y be unit vectors representing the projections in two different images of the same 3D point. Then, these points satisfy the epipolar constraint

$$x^T [t]_\times R y = 0, \tag{1}$$

where R is a rotation and $E = [t]_\times R$ is the essential matrix. We will use the following property for essential matrices. It is proven as Lemma 5.6 in [11].

Lemma 1. *Let $E = [t]_\times R$ be an essential matrix. If E is skew-symmetric then $R = I$ or $R = e^{\pi[t]_\times}$.*

Theorem 1. *If the image points x_i and y_i are related by a pure rotation*

$$y_i = Qx_i \tag{2}$$

then for a unit vector t, every solution of the form

$$\gamma_i x_i = [I\ 0]U_i, \quad \gamma_i' y_i = [R\ t]U_i \tag{3}$$

where $\gamma_i, \gamma_i' > 0$ has $R = Q$ unless the image points lie on a 2nd order surface $y_i^T A y_i = 0$, where $A \neq 0$ with eigenvalues λ_1, λ_2 and $\lambda_1 + \lambda_2$ and $\lambda_1\lambda_2 \leq 0$.

Proof. From (3) we get

$$\gamma_i' y_i = \gamma_i R x_i + at, \quad a \in \mathbb{R}. \tag{4}$$

We form an essential matrix $E = [t]_x RQ^T$ and note that

$$y_i^T [t]_x RQ^T y_i \overset{(2)}{=} y_i^T [t]_x R x_i \overset{(4)}{=} \frac{(\gamma_i R x_i + at)^T}{\gamma_i'} [t]_x R x_i = 0.$$

Hence with $A = E + E^T$ we get $y_i^T A y_i = 0$. It was shown in [8] that such a matrix has eigenvalues λ_1, λ_2 and $\lambda_1 + \lambda_2$, with $\lambda_1\lambda_2 \leq 0$. It remains to consider the case $A = 0$. By Lemma 1, this implies that $S = I$ or $e^{\pi[t]_\times}$. In any case $St = t$. Using (4)

$$\gamma_i' y_i = \gamma_i R x_i + at \overset{(2)}{=} \gamma_i RQ^T y_i + at = \gamma_i S y_i + at. \tag{5}$$

Taking cross products with t yields,

$$\gamma_i' t \times y_i = \gamma_i t \times (S y_i) = \gamma_i (St) \times (S y_i) = \gamma_i S(t \times y_i) \tag{6}$$

which means that $t \times y_i$ is an eigenvector of S with positive eigenvalue. Unless all y_i are parallel with t this excludes the solution $S = e^{\pi[t]_\times}$. Hence $S = I$ and thus $R = Q$. □

The theorem states that in general there are no other solutions than R with $t \neq 0$. It is easy to see that the same holds if $t = 0$. This shows that in the noise-free case the rotation can generally be determined even with zero baseline. To see what happens when there is image noise a simple synthetic experiment was performed. A set of 50 3D-points were randomly generated in the unit cube $[-1, 1]^3$ and two cameras with unit focal lengths were placed in the points $(0, 0, 10) \pm r$ where r is a random vector of a specified length. Gaussian noise with standard deviation 0.0001 was added to the image projections. Then the relative orientation of the cameras was estimated using the minimal solver from [14] in a 100-iteration RANSAC loop. The solution having most inliers was picked as a starting point for bundle adjustment. This was repeated for different camera distances and the results were averaged over 50 runs; see Figure 2. Note that since the scale of the reconstruction is arbitrary we can only compare the translation direction. Still the figure shows clearly that, as the camera distance decreases, the rotation estimates remain stable whereas the translation errors increase drastically. This will also lead to very poor accuracy in the 3D structure estimation.

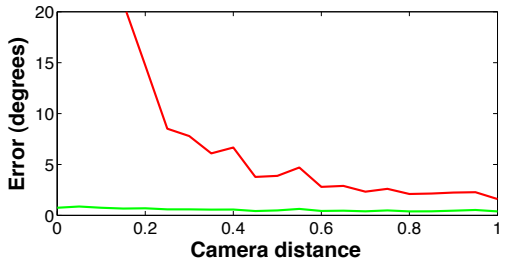

Figure 2. *Stability of the orientation estimate*. Average error in degrees of the estimates for the translation (red line) and rotation (green line) vs. different baseline distances.

This experiment shows that there is no reason not to use view pairs with short baselines. In fact, when the baseline is short the number of correspondences is generally large and hence the orientation estimate will often be more accurate than with a larger baseline.

3. Cycles and Consistency

In this section, we introduce and motivate a mathematical model for rotational consistency in structure from motion problems.

3.1. Problem Formulation

The relative orientations that are estimated from pairs of views induce a camera graph. The graph has a vertex for each camera and edges between cameras i and j if a relative rotation, denoted by $\tilde{R}_{ij}$, is given between these views. By analyzing the graph we can estimate the absolute camera rotations, R_i, with respect to a global coordinate system.

Our discussion will require some characteristics of the group of 3D-rotations, often referred to as $SO(3)$. The usual metric on $SO(3)$ can be defined

$$d(R, S) = \max_{x \in \mathbb{R}^3 \setminus \{0\}} \angle(Rx, Sx) \tag{7}$$

where $\angle(x, y)$ denotes the angle between vectors x and y taking values in $[0, \pi]$. If R is given by an axis of rotation and an angle $\alpha \in [-\pi, \pi]$ then $d(R, I) = |\alpha|$.

Ideally there would exist rotations R_i such that

$$R_i = \tilde{R}_{ij} R_j \text{ for all } (i, j) \in E. \tag{8}$$

Due to uncertainty in the matching process and the camera model, this will not be the case. Instead we will have to deal with low-level noise as well as completely inconsistent rotations. Low-level noise can be handled with the method in [6] but first, all outlier rotations must be removed. The goal is to find a set of consistent rotations.

Definition 1. *Given a camera graph $G = (V, E)$ and error tolerance ϵ, we say that G is consistent if there exist rotations $R_1, \ldots, R_N$ where $|V| = N$ satisfying*

$$d(R_i, \tilde{R}_{ij} R_j) \leq \epsilon \text{ for all } (i, j) \in E. \tag{9}$$

If there are outlier rotations, the whole camera graph will not be consistent. In these cases we want to find a large consistent subgraph. More precisely, we want a subgraph that contains as reliable orientations as possible. Let p_{ij} be the probability that the estimated relative orientation $\tilde{R}_{ij}$ is an outlier. In Section 2 we saw that the accuracy of an estimated rotation does not depend directly on the baseline distance. Thus it is reasonable to model p_{ij} as a decreasing function of the number of inliers, denoted by w_{ij}. For simplicity, we choose to optimize the sum of w_{ij}'s rather than trying to estimate the log probabilities. We seek those camera orientations which are supported by the maximum number of point correspondences. However, the same approach can be used when estimates of the probabilities exist.

Problem 1. *Given a connected camera graph $G = (V, E)$ and edge weights w_{ij}, we want to find a consistent subgraph $G_c = (V, E_c)$ that maximizes*

$$\sum_{(i,j) \in E_c} w_{ij}. \tag{10}$$

This formulation involves finding both the absolute rotations R_i and the set of consistent rotations E_c. One way to attack this difficult optimization problem is to consider cycles in the camera graph. This gives means to detect and remove incorrect relative rotations prior to the continuous optimization. Computing the product of rotations along a cycle in the camera graph should give roughly the identity matrix. Large deviations, inconsistencies, could indicate an incorrectly estimated geometry. The next section contains some theoretic results concerning the connection between cycles and consistency.

3.2. Cycles and Consistency

The first result gives a necessary constraint on cycles for a graph to be consistent. Essentially the same result can also be found in [15] but we give a simpler proof.

Theorem 2. *Consider a camera graph G that consists of a single simple cycle $i_1, i_2, i_3 \ldots i_n, i_1$. If the estimated rotations along this cycle satisfy*

$$d(\tilde{R}_{i_1 i_2} \tilde{R}_{i_2 i_3} \ldots \tilde{R}_{i_n i_1}, I) = \omega, \tag{11}$$

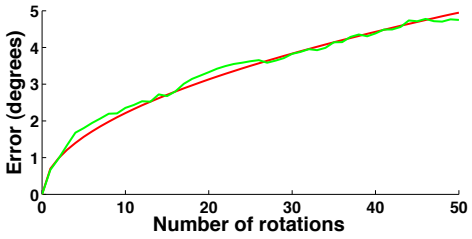

Figure 3. *Cycle error vs. cycle length.* The green curve gives the average over 50 trials and the red curve the function $y = c\sqrt{x}$.

then $\epsilon = \omega/n$ is the smallest ϵ such that G is consistent with Definition 1.

Proof. To prove that G is consistent with $\epsilon = \omega/n$, it is sufficient to find rotations R_{ij} such that $d(R_{ij}, \tilde{R}_{ij}) \leq \omega/n$ and $R_{12} R_{23} \ldots R_{n1} = I$. To find R_{12}, let $D = \tilde{R}_{12} \ldots \tilde{R}_{n1}$. This is a rotation ω radians around some axis. Let $D_{\omega/n}$ be a rotation around the same axis but $-\omega/n$ radians and set $R_{12} = D_{\omega/n} \tilde{R}_{12}$. Then

$$d(R_{12} \tilde{R}_{23} \ldots \tilde{R}_{n1}, I) = \omega - \omega/n. \tag{12}$$

By repeating this scheme with $D = \tilde{R}_{23} \ldots \tilde{R}_{n1} R_{12}^T$ we can compute R_{23} such that the error decreases to $(n-2)\omega/n$ and the result follows by induction. □

The above theorem states that if the cycle error is larger than $n\epsilon$, than the cycle must contain an outlier rotation. However, we want to know when it is probable that a cycle contains an outlier. The following experiment illustrates how the error in a cycle depends on the cycle length if there are no outliers. If the error in a cycle is significantly larger than this, then there is a high probability that the cycle contains at least one outlier.

Fifty pairs of views were generated in the exact same manner as in the experiment of rotation stability (Section 2). For each pair a rotation was estimated using RANSAC followed by bundle adjustment. Let R_i be the ground truth rotation for the ith view pair and let $\tilde{R}_i$ be the estimated rotation. For each n the error $d(R_1 R_2 \ldots R_n, \tilde{R}_1 \tilde{R}_2 \ldots \tilde{R}_n)$ was measured. The error for different values of n, is shown in Figure 3 (averaged over 50 runs). The conclusion is that comparing the cycle error to $\sqrt{n}\epsilon$ gives a good indication of the presence of outliers.

4. Robust Estimation of Orientations

We will now describe how to robustly estimate the camera orientations with respect to a global coordinate system.

4.1. Handling Low-Level Noise

For low levels of noise we will use the method proposed in [6]. It is based on representing rotations as unit

quaternions. Let us first remark that unit quaternions q and $-q$ correspond to the same rotation and that the quaternion representation gives us a way to compute the distance between rotations. Let $\langle p, q \rangle$ denote scalar multiplication with quaternions seen as 4-vectors. If $\langle p, q \rangle \geq 0$ we have

$$d(R_p, R_q) = 2 \arccos \langle p, q \rangle. \tag{13}$$

Using quaternions, (8) can be expressed as

$$\tilde{Q}_{ij} q_j - q_i = 0, \text{ for all } (i,j) \in E. \tag{14}$$

Here q_i,q_j are the quaternions of the absolute rotations of cameras i and j and $\tilde{Q}_{ij}$ is a 4×4 matrix corresponding to quaternion multiplication by $\tilde{q}_{ij}$. In [6] the camera orientations are determined by solving these equations in a least squares sense. To motivate this, look at two unit quaternions p and q corresponding to rotations R_p and R_q such that $d(R_p, R_q) = \alpha$. Assuming that α is small, (13) yields,

$$|p - q|^2 = \langle p - q, p - q \rangle = 2 - 2 \cos(\alpha/2) \tag{15}$$

$$= 4 \sin^2(\alpha/4) \approx \alpha^2/4. \tag{16}$$

Thus what we are minimizing is approximately the sum of squared angular errors.

It was noted in [3] that the ambiguity when representing rotations with quaternions may cause this method to fail. Say for example that the camera has moved around a building, while rotating an angle 2π. Let $q_1 = (1, 0, 0, 0)$. If we choose quaternion representation in the standard way, we will constrain the orientation quaternions q_i to move smoothly on the unit sphere of quaternions. This means that when we are back where we started the orientation has just moved halfway around the sphere of unit quaternions. So for the linear equations to hold we have to represent q_1 with $(1, 0, 0, 0)$ in some equations and $(-1, 0, 0, 0)$ in others.

The approach presented in the next subsection to remove inconsistent relative rotations will also provide estimates $\bar{q}_i$ for the camera orientations. This gives us a way to resolve the ambiguity problem. For all $(i,j) \in E$:

1. Represent $\tilde{R}_{ij}$ with a quaternion $\tilde{q}_{ij}$.
2. Compute $\tilde{Q}_{ij}$ as the matrix representation of $\tilde{q}_{ij}$.
3. If $|\bar{q}_i - \tilde{Q}_{ij} \bar{q}_j| > |\bar{q}_i + \tilde{Q}_{ij} \bar{q}_j|$, set $\tilde{Q}_{ij} = -\tilde{Q}_{ij}$.

4.2. Handling Large Errors

The results in Section 3.2 showed that cycles can be used to detect inconsistent rotations, but considering all cycles is rarely feasible. Instead, we start from a spanning tree. If there are cycles in the graph there are also multiple ways to choose spanning trees. Following Problem 1, it seems natural to seek a maximum-weight spanning tree. Such a tree is easily found using a greedy algorithm [1].

Assuming (for a moment) that the generated spanning tree contains no outlier rotations, we now have means to detect outliers among the other epipolar geometries. It easy to see that adding any edge to a spanning tree will generate a cycle. Let R_C be the composition of all rotations along the cycle. As shown in Section 3.2, for a cycle of length $|C|$, if

$$d(R_C, I) > \sqrt{|C|}\epsilon, \tag{17}$$

then with high probability, the cycle contains an outlier rotation and should not be used; see Algorithm 1.

Algorithm 1 Rotational consistency

Compute a maximum spanning tree, T, with weights w_{ij}.
Set $E_c = T$.
for each $e \in E \setminus T$
  Let C be the cycle formed by e and T.
  if the error in C is less than $\sqrt{|C|}\epsilon$
    $E_c = E_c \cup e$
Estimate absolute orientations from E_c (see Section 4.1).
Apply additional search heuristics (see text).

The absolute orientations yielded by this approach can be improved by repeating steps 2 to 7 to give a better solution to Problem 1. If the initial spanning tree did contain an actual outlier rotation, then it would prevent us from adding rotations that should be regarded as inliers. Let $E_{outlier}$ be the set of outlier edges after the for-loop of Algorithm 1. We also use some other search heuristics for improving the solution:

- For each $e \in E_{outlier}$, estimate absolute rotations for $E_c \cup e$ and check for consistency.

- For each $e \in E_{outlier}$, create a new spanning tree that contains e and repeat steps 2 to 7 of Algorithm 1.

These simple heuristics work remarkably well. Since the spanning tree will consist of those relative rotations that had the highest number of inliers, large errors in the spanning tree are unlikely. In Table 1 we show the effects of applying our heuristics to the datasets of the experimental section. The total number of epipolar pairs after the RANSAC stage, after the initial spanning tree and for the final reconstruction stage, respectively, are presented. Note that the local heuristics of improving the initial spanning tree solution is an important step. The fact that we are able to reconstruct plausible 3D reconstructions for all the data sets, also show that the rotational consistency algorithm does a good job.

Another way to check the validity of the rotational consistency estimation is to apply the algorithm to closed-loop sequences, that is, a sequence of images where the camera trajectory is a long closed loop. Note that this fact is *not* used in the inference. Figure 4 shows two such examples. The camera graphs, before and after applying the algorithm,

Sequence name	Images	Epipolar pairs	Initial tree	Final result
CASTLE	83	667	509	519
ROAD ROLLER	41	214	138	148
RAILROAD	34	126	112	114
STREET	99	1413	1044	1242
APARTMENT	222	1570	1097	1201
CATHEDRAL	544	8168	7868	7868

Table 1. Results of rotational consistency: Number of images, total number of two-view geometries, the initial number of consistent pairs and the final number of consistent pairs are tabled.

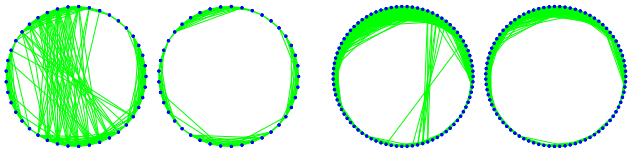

Figure 4. Left: Camera graphs before and after running Algorithm 1 for the ROAD ROLLER sequence. Right: Camera graphs before and after running Algorithm 1 for the CASTLE sequence. Each edge corresponds to an estimated relative rotation. For this illustration the cameras have been placed on a circle using prior knowledge of the true geometry, but this prior knowledge has *not* been used to detect the erroneous rotations.

are plotted as circles. Only two-view geometries that passed the RANSAC stage with at least 10 image correspondences are used as input. An edge in the (circular) camera graph corresponds to a two-view geometry. As can be seen, there are many false edges occurring for cameras far away from each other. By enforcing rotational consistency, hence computing a solution for Problem 1, the resulting camera graphs do not have any false edges.

In [23] a method for removing inconsistent cycles is proposed. This method is however limited to cycles of length 6 (and uses a prior that is independent of cycle length). For the sequences in Figure 4, many of the erroneous edges cannot be detected without analyzing considerably longer cycles. Therefore many of the incorrect edges would not be removed by that method.

5. Reconstructions

We have tested the developed approach on a collection of real image sequences. Figures 5-10 show some screenshots. This section presents some performance statistics and a comparison to the state-of-the-art software BUNDLER.

Data. The set of image collections has been obtained by taking photos with standard digital cameras. Image sizes vary from a couple of hundred to up to 3000 pixels in width. The number of images in a sequence ranges from 34 to 544 images; see Table 1. Only images where it has been possible to extract the focal length from the EXIF tag has been processed. The principal point is assumed to be in the middle of the image and the skew is set to zero and the aspect ratio to one.

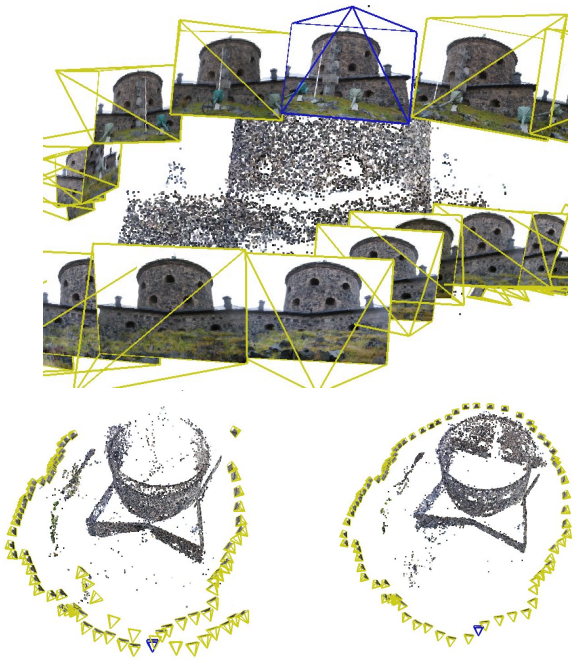

Figure 5. CASTLE. The top and left figures shows the 3D result from BUNDLER and on the right, our result is given. By careful inspection, one can see that the top and bottom image rows of the top figure display *different* facades of the castle. (A window is blocked by stairs in the top row.) This confusion of facades yields an incomplete and false reconstruction. Using rotational consistency, a complete trajectory is obtained.

Implementation details. For the feature extraction and matching, as for the extraction of focal lengths, we use exactly the same setting as in BUNDLER. More precisely, the matching stage is based on standard SIFT matching with default settings. Note that the input to our system and BUNDLER is identical since the same software is used. For the estimation of pairwise epipolar geometries, we allow 1000 RANSAC iterations. The threshold is set to 3 pixels for a point correspondences (measured from the epipolar line in each image). If more than 10 point correspondences are obtained for an image pair, the two-view geometry is kept for later processing. We always check for cheirality (positive depths). When computing camera translation and 3D points, we allow a larger error, namely 10 pixels, as the initial estimates of rotations may be slightly off.

All our algorithms have been implemented in MATLAB. Given estimated two-view geometries, running times are

typically between 5-10 minutes depending on the number of images, and the number of extracted feature points. The same parameter settings have been used for all the data sets.[2]

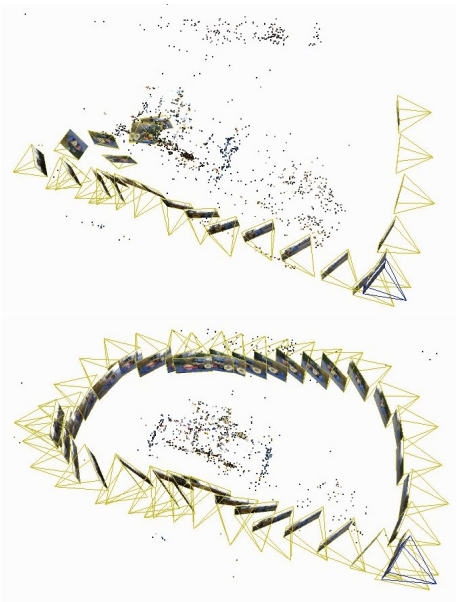

Figure 6. ROAD ROLLER. Top: BUNDLER. Bottom: Proposed method.

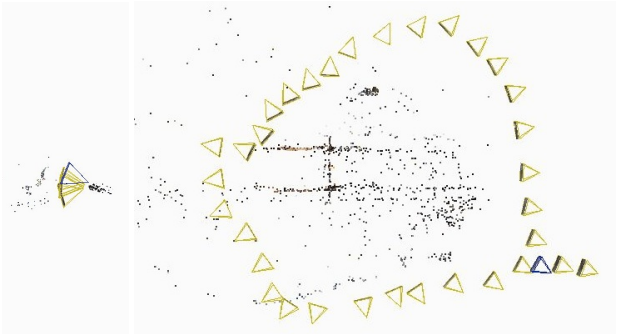

Figure 7. RAILROAD. Left: BUNDLER. Right: Proposed method.

Closed-loop sequences. For our first three experiments we have chosen to use closed-loop sequences. For these datasets it is easy to detect if the method fails by investigating the ability to close the loop. Note that since there is no independent system that is guaranteed to give the true reconstruction it is very difficult to obtain ground truth. Hence the only way that we can really determine the quality of the reconstruction is by visual inspection. Note that, in our system, the ordering of the images is *nowhere* used. However, the images are taken in order so this gives us a way to check if correct epipolar geometries have been computed.

[2]The data sets and code will be made publicly available.

The closed-loop sequences are the CASTLE (see Figure 5), ROAD ROLLER (see Figures 1 and 6) and RAILROAD (see Figure 7). In the first two sequences there are repeated textures introducing false two-view geometries. Because of these false geometries BUNDLER fails to reconstruct the loop. Note that the front and the back of the castle are confused. In case of the RAILROAD data set, it is a bit unclear why BUNDLER fails, but we believe it is due to too unreliable two-view epipolar estimates.

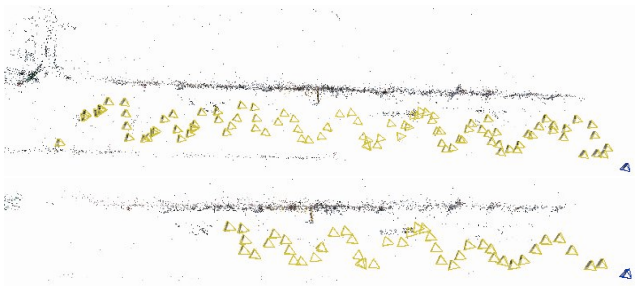

Figure 8. STREET. Solutions from the leave-one-out test. Top: Complete reconstruction using the proposed method. Bottom: Incomplete reconstruction using BUNDLER; see text for details.

Leave-one-out test. STREET. In our method, all images are handled in a uniform manner. This is in contrast to BUNDLER which selects an initial epipolar geometry to base the reconstruction on and then sequentially adds new images. To test this dependency, we tried to reconstruct the same sequence with one image removed. This was repeated for all 99 images. The reconstructions were validated by registering to the original reconstruction. While our method is unaffected by removing one image in all cases, there are two cases for which BUNDLER fails to reconstruct the whole scene (and in these cases, only 45 and 57 cameras are reconstructed, respectively); see Figure 8.

Regular scenes. APARTMENT. In this two bedroom apartment, there are (natural) weak geometric links between different rooms. It is difficult to detect any difference from the two overview images in Figure 9. Both methods provide satisfactory results, and the differences are minor, however it turns out that there are 6 images of the bathroom that BUNDLER is not able to incorporate into the reconstruction. The final data set is the CATHEDRAL, see Figure 10. Both methods produce satisfactory results for this data set.

6. Conclusions

We have presented a system for large-scale 3D reconstructions from unordered images. Compared to the standard sequential approaches, we have shown that short baselines can be used as reliable building blocks, since rotations

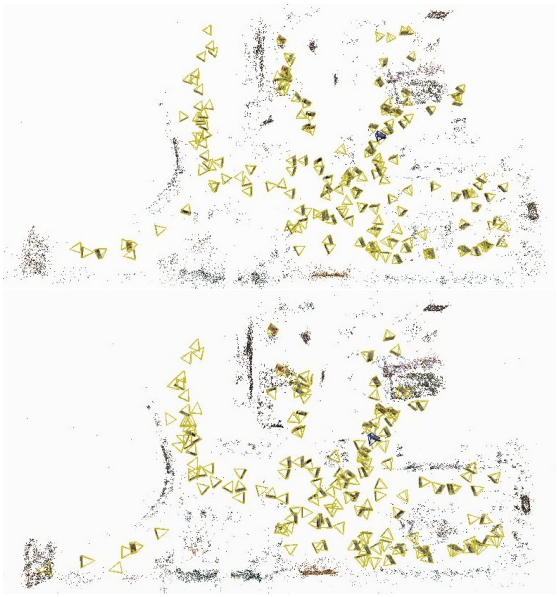

Figure 9. APARTMENT. Top: BUNDLER. Bottom: Proposed method.

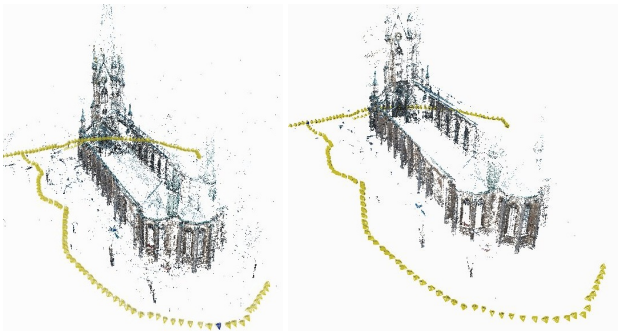

Figure 10. CATHEDRAL. Left: BUNDLER. Right: Proposed method.

can be estimated accurately and matching between such views is relatively simple. Moreover, in our non-sequential system, we have demonstrated improvements with respect to state-of-the-art regarding loop-closing, detecting repetitive structures and obtaining a good global solution without depending on a specific base pair of views. These features have been supported by theoretical results as well as experimental comparisons on real data.

Another possible weakness of the proposed method is that repetitive structures with consistent rotations, for example two parallel billboards would not be detected by rotational consistency and would have to be handled by the final robust estimation stage, but this stage is not designed to cope with large rates of outlier points. Although we have not encountered such problems in practice, we view it as an interesting topic for future work.

References

[1] J. A. Bondy and U. S. R. Murty. *Graph Theory*. Springer-Verlag, 2008. 5

[2] K. Cornelis, F. Verbiest, and L. Van Gool. Drift detection and removal for sequential structure from motion algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004. 1

[3] Y. Dai, J. Trumpf, H. Li, N. Barnes, and R. Hartley. Rotation averaging with application to camera-rig calibration. In *Asian Conf. Computer Vision*, 2009. 5

[4] A. Dalalyan and R. Keriven. L1-penalized robust estimation for a class of inverse problems arising in multiview geometry. In *Neural Information Processing Systems*, 2009. 2

[5] R. Gherardi, M. Farenzena, and A. Fusiello. Improving the efficiency of hierarchical structure-and-motion. In *Conf. Computer Vision and Pattern Recognition*, 2010. 2

[6] V. M. Govindu. Combining two-view constraints for motion estimation. In *Conf. Computer Vision and Pattern Recognition*, 2001. 4, 5

[7] V. M. Govindu. Robustness in motion averaging. In *Asian Conf. Computer Vision*, 2006. 2

[8] F. Kahl and R. Hartley. Critical curves and surfaces for Euclidean reconstruction. In *European Conf. on Computer Vision*, 2002. 3

[9] F. Kahl and R. Hartley. Multiple view geometry under the L_∞-norm. *IEEE Transations on Pattern Analysis and Machine Intelligence*, 2008. 2

[10] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 2004. 2

[11] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An Invitation to 3-D Vision: From Images to Geometric Models*. Springer Verlag, 2003. 3

[12] D. Martinec and T. Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *Conf. Computer Vision and Pattern Recognition*, 2007. 2

[13] D. Nistér. Reconstruction from uncalibrated sequences with a hierarchy of trifocal tensors. In *European Conf. on Computer Vision*, 2000. 2

[14] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transations on Pattern Analysis and Machine Intelligence*, 2004. 1, 2, 3

[15] G. Sharp, S. Lee, and D. Wehe. Multiview registration of 3d scenes by minimizing error between coordinate frames. *IEEE Transations on Pattern Analysis and Machine Intelligence*, 2004. 2, 4

[16] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from Internet photo collections. *Int. Journal of Computer Vision*, 2008. 1

[17] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *European Conf. on Computer Vision*, 1996. 1

[18] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag, 2010. 1

[19] J.-P. Tardif, A. Bartoli, M. Trudeau, N. Guilbert, and S. Roy. Algorithms for batch matrix factorization with application to structure-from-motion. In *Conf. Computer Vision and Pattern Recognition*, 2007. 1

[20] T. Thormaehlen, H. Broszio, and A. Weissenfeld. Keyframe selection for camera motion and structure estimation from multiple views. In *European Conf. on Computer Vision*, 2004. 1

[21] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *Int. Journal of Computer Vision*, 1992. 1

[22] P. Torr, A. Fitzgibbon, and A. Zisserman. The problem of degeneracy in structure and motion recovery from uncalibrated image sequences. *Int. Journal of Computer Vision*, 1999. 1

[23] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. In *Conf. Computer Vision and Pattern Recognition*, 2010. 2, 6