# Robust Incremental Structure from Motion

Manfred Klopschitz, Arnold Irschara, Gerhard Reitmayr, Dieter Schmalstieg

Graz University of Technology

{klopschitz,irschara,reitmayr,schmalstieg}@icg.tugraz.at

## Abstract

*We present a novel method to solve structure and motion problems robustly and incrementally from unordered sets of input images. The proposed method can build large reconstructions without depending on global structure for outlier rejection and starts from the most reliable parts of the data sets. The two main ideas are a strategy to identify reliable subsets of images that have the highest mutual compatibility and an ordering of the reconstruction buildup that gives higher priority to these subsets and merges new information according to this ordering. Another advantage of our buildup strategy is that loop closing is done unbiased by drift in spite of adding correspondence data incrementally. Correspondence information is only locally verified and merged into reconstructions without immediately removing outliers based on globally reconstructed structure. We demonstrate the robustness and scalability of our approach on several large reconstructions from unordered sets of images and indicate the achieved accuracy by preserving the topology of the 3D structure and cameras.*

## 1. Introduction

Current Structure from Motion (SfM) methods may be classified based on the image data they use. Two major types of image sources are still images and video sequences. Recent trends in photo community websites and the ubiquitous availability of consumer grade compact cameras have shifted the research interest towards reconstruction from unordered image data sets [18]. This choice of input data usually has strong implications on the selection of algorithms used to solve the SfM problem. Unordered sets of still images rely on wide baseline image matching techniques to establish correspondence information and the SfM problem may be solved for all input images simultaneously or in an incremental way. Two examples of this type of systems are [10] and [17].

This paper presents an approach to build reconstructions robustly and incrementally from unordered sets of input images. Incremental methods can add new images into existing reconstructions when they become available and make the system more flexible. The same is true for using unordered sets of input images instead of video streams. Compared to feature point tracking in video sequences, wide baseline matching provides additional links between images separated by time. Recent image matching methods make the harder wide baseline matching problem more tractable and scalable.

We focus our attention in this work on making the incremental still image based reconstruction process more robust and flexible. We base our reasoning about feature track compatibility and image connectivity on image triplets because they are well known to be more robust against false feature matches and naturally extend to graph based representations. Given a set of pairwise correspondences for input images, we transform these correspondence into corresponding image triplet reconstructions. This reduces the number of outliers compared to epipolar geometry, verifies track compatibility locally using pairs of triplet reconstructions, detect overlap in this triplet representation efficiently with overlapping views and find starting points for the reconstruction that are most reliable. Another important aspect of our work is that we do not use the 3D points in the evolving reconstructions to generate tracks of reprojection inliers. This common strategy makes the implicit assumption that no drift is present and needs an explicit loop closing strategy. We avoid this step by closing loops implicitly, using only the local correspondence information from triplet to triplet registration.

We motivate the concept of *local correspondences* with the example shown in Figure 1. In a reconstruction from a small set of unordered images, drift builds up due to slight camera calibration errors. This reconstruction was built by registering individual views to the evolving, global structure. Towards the end, correspondences that would close the loop are immediately discarded as the accumulated drift implies a large re-projection error. As a result the reconstruction is distorted and the loop cannot be closed (see right image in Figure 1) because the required correspondences are classified as outliers. In general, incremental SfM pipelines usually discriminate correspondences into in- and outliers by evaluating some robust estimator on the global structure obtained in the build up process. When drift is present this can introduce severe bias and correct classification degrades with the number of added images. We propose to avoid this step and establish global feature correspondences by adding only information from local triplet to triplet correspondences. The in- and outlier classification is done by checking the feature compatibility of triplet pairs. This gives an unbiased classifier for in- and outliers that does not depend on the succession of image insertions.

The main contributions of the proposed method are a strategy to identify the most reliable parts of unordered sets of images, build the reconstructions incrementally, using this information and add only locally verified correspondences. This creates a core structure of the input, i.e. reconstructions using the most reliable information. Because feature point tracks are verified only locally on image triplets, the loop closing problem is not biased by drift.

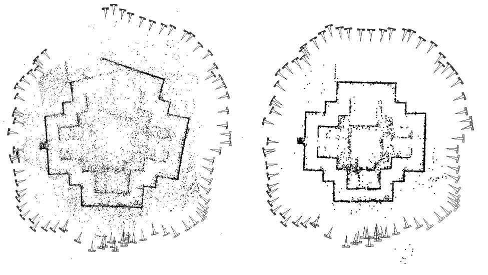


Figure 1. *Incremental reconstruction with and without drift.* This example shows the output of an incremental reconstruction pipeline where new views and feature point tracks are registered using the global 3D structure. Due to slight errors in (a) the camera calibration drift builds up in this reconstruction. If additional views are added at the point where the structure breaks up, the new views are registered only at one end of the loop and do not add additional constrains to the true global topology. Figure (b) shows the correct topology.

Recent literature in Structure from Motion comprises a number of approaches addressing the problem of reconstructing a scene from unordered collections of images [17, 8, 10]. Snavely et al. describe in [17] a system that is able to reconstruct a scene from a very diverse set of images, like image collections gathered from the web. A limiting factor regarding the scalability of this approach is mainly pair-wise image matching and large scale bundle adjustment. In [18] the latter problem is addressed by computing a small but representative *skeletal* set of a scene. Targeting at efficiency, Ni et al. propose in [7] an out-of-core bundle adjustment capable of tackling larger reconstructions. Closely related is the approach described in [8] where object recognition techniques are utilized to compute a small subset of *iconic images* that represent the important aspects from a scene. Again the algorithm is designed to work on large-scale image collections gathered from the Internet. Common for all the approaches is to rely on some calibration information, often directly derived from EXIF information. In our approach, we also utilize calibrated cameras with known focal length.

Considering view triplets as the basic SfM building block was addressed by many authors for image sequences [4] and unordered image collections [22]. The triple relation based on the trifocal tensor imposes stronger geometric constraints and allows the local detection of mismatches. For instance, Zach et al. apply in [22] a non-monotone Bayesian reasoning based on view triplets to detect incorrect two view geometries.

Our approach builds upon recent advances in image retrieval, utilizing vocabulary tree data structures and inverted files [13] to speedup matching. Schindler et al. have shown [14] that this approach is also suitable for large scale location recognition. In our system a vocabulary tree structure is used for coarse image matching and is therefore related to [6]. Recently, vocabulary based indexing structures were also successfully applied for Simultaneous Localization and Mapping (SLAM) tasks as described in [3].

## 2. Structure and Motion Computation

Our algorithm consists of three major steps. (i) An epipolar graph $G_\mathcal{E}$ is created with images as nodes and correspondences verified by epipolar geometry as edges. The feature matching process is accelerated with a bag of words approach. (ii) This graph is then transformed into a graph $G_\mathcal{T}$ of triplet reconstructions. The nodes in this graph are all trifocal reconstructions created from $G_\mathcal{E}$ and are basically connected by overlapping views. These connections, i.e. edges, of $G_\mathcal{T}$ are created when triplets share at least one view and pass a test for 3D point compatibility. The feature correspondences of triplets are established by using tracks from the overlapping views. (iii) These edges of $G_\mathcal{T}$ are then merged incrementally into reconstructions, while loop closing is handled implicitly. Because the process is incremental, additional sets of unordered images can be easily added to extend created reconstructions after a set of input images has been processed.

### 2.1. Epipolar Graph $G_\mathcal{E}$

Given a set of unordered input images, SIFT features [9] are extracted from every image. We use a vocabulary tree [13] based approach for coarse matching of similar images. Hence we can greatly reduce the computational effort of pair-wise image matching, by only matching the most relevant images as reported by the vocabulary scoring.

In our system the vocabulary tree is trained in an unsupervised manner with a subset of 2.000.000 SIFT feature vectors randomly taken from 2500 images. Thus our vocabulary is generic and allows the generalization to different data sets. The descriptor vectors are hierarchically quantized into clusters using a k-means algorithm. As proposed in [13], we set the branch factor to 10 and allow up to 7 tree levels. The image retrieval performance can be increased by using a higher branching factor [14]. Once the vocabulary tree is trained, searching the visual vocabulary is very efficient and new images can be inserted on-the-fly.

In our current setting we rely on an entropy weighted scoring similar to the *tf-idf* "term frequency inverse document frequency" as described in [16]. Let $\mathcal{D}$ be an image in our database and $t$ be the term in the vocabulary associated to feature $f$ of the current query image $\mathcal{Q}$, then our scoring function $sim(\mathcal{Q}, \mathcal{D})$ is,

$$sim(\mathcal{Q}, \mathcal{D}) = \frac{1}{|\mathcal{Q}| + |\mathcal{D}|} \sum_{t \in \mathcal{Q} \cap \mathcal{D}} \log \left( \frac{N}{n(t)} \right) \qquad (1)$$

where $N$ is the total number of images in the collection, $n(t)$ is the number of images that contain term $t$ and $|\mathcal{Q}|$, $|\mathcal{D}|$ are the number of features from the query and database image, respectively. This weighting allows fairness between database images with different number of features.

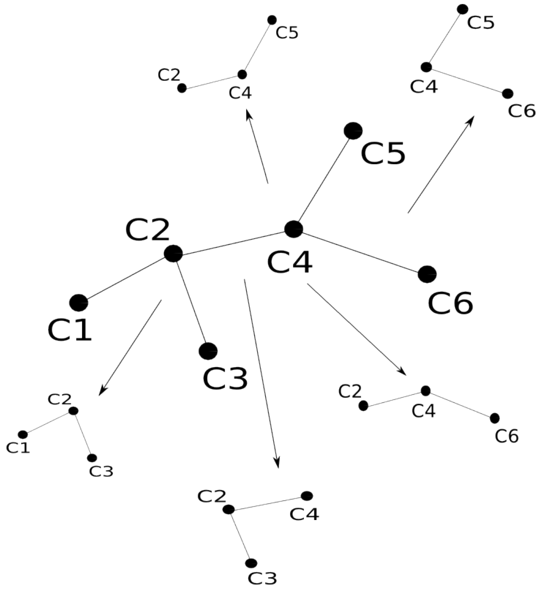
The tentative sparse image correspondences retrieved from the vocabulary tree are then matched using an approximated nearest neighbor technique. The epipolar geometry is computed using a five-point [12] minimal solution inside a RANSAC loop. The correspondence inlier set is used to build the epipolar graph $G_{\mathcal{E}}$ of image connections. The nodes are the images and the edges the inlier set of the pairwise epipolar geometry.

## 2.2. Trifocal Graph $G_{\mathcal{T}}$

By using three images as the basic geometric entity, the number of false correspondences can be reduced for points that are visible in all three views and a more reliable basic graph representation of the image connections can be established. Image triplets are the nodes of this graph and are created from the epipolar geometries. These nodes are basically connected by overlapping views. Degenerate configurations can also be present in the trifocal case, but these configurations are usually geometrically incompatible with other triplets and are at most present at the fringe of the graph.

### 2.2.1 Trifocal Reconstructions

In the next step we create all potential trifocal reconstructions from the edge information in the epipolar graph $G_{\mathcal{E}}$. A Minimum Spanning Tree (MST) of $G_{\mathcal{E}}$ is created in a similar way as in [20] but is only used to enumerate potential image triplets efficiently. The MST is traversed and all possible triplet candidates are generated from this list. Then the edges of the MST are removed from $G_{\mathcal{E}}$ and the next MST is generated. Figure 2 shows how triplets are enumerated from one MST. This process is iterated until all edges of $G_{\mathcal{E}}$ are processed. The advantage of the MST creation over a brute force triplet enumeration is that it can be stopped after a few iterations and uses the best globally connected matching epipolar geometries first.



(a)

Figure 2. *Triplet enumeration.* This example shows how a MST is used to enumerate image triplets. The MST consists of 6 cameras $C1...C6$ and consists of 5 triplet candidates.

*Triplet Reconstruction and Reconstruction Quality:* We reconstruct the trifocal structure if the three images are fully connected by three epipolar edges. Two connections suffice if all three views share at least one point but we require that all three connections are present. For all three pairwise relative orientations obtained with a minimal solver [12] the third view is inserted with a three point calibrated absolute pose solver [5] inside a RANSAC loop. The configuration with the highest inlier count is selected and optimized with bundle adjustment [21].

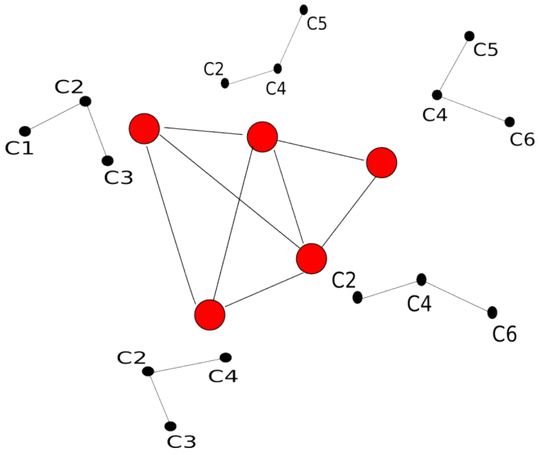
A third view of 3D structure only increases the discriminability of false three view feature point matches if it adds additional information, i.e. reduces the covariance of the triangulated scene point. We use [1] to compute the median roundness of the 3D structure uncertainty that is visible in all three views for all three image pairs and reject a triplet if at least one epipolar combination does not provide a certain baseline. Therefore, we enforce a minimal triangulation angle between all pairs of images, and avoid degenerated triple configurations (e.g. epipolar relations from pure rotational motions).

Each accepted trifocal reconstruction is inserted as node into $G_{\mathcal{T}}$. In the next step, connections between these nodes will be created.

### 2.2.2 Trifocal Graph Edges

Overlap between the set of trifocal reconstructions has to be established. We distinguish between the detection of potential overlap of two trifocal reconstructions (connectivity) and the geometric consistency of two trifocal reconstructions (compatibility).

[4], two image triplets can share zero, one or two images. To simplify matching we only consider triplets that have at least one view in common for a potential edge in $G_{\mathcal{T}}$. Correspondences between the two 3D point sets are established with the common images. The correspondence information from common views does not take all possible structure matches into account, but different combinations of triplets will usually contain this information.



(a)

Figure 3. *Triplet Graph.* Each node in this graph represents a reconstructed image triplet. Common views of the reconstructed image triplets are used to determine the connectivity in this trifocal graph $G_{\mathcal{T}}$.

*Compatibility:* Two trifocal reconstructions that are potentially connected are registered into one coordinate system by computing the similarity transform of the two point sets in a RANSAC loop and evaluating the reprojection errors of the transformed points. The two aligned triplets are then used to create all possible local tracks and contradictory measurements are removed. For two triplets with two overlapping views this means local tracks of length two, three and four are created.

This local correspondence information is inserted into $G_{\mathcal{T}}$ as edge. Figure 3 shows how overlapping views induce the graph structure. The edge information will be used in the next step to create a global representation of the scene.

## 2.3. Reconstruction

The main idea of our merging process of image triplets into global reconstructions is that we start from the most reliable parts simultaneously and merge only the "local" edge information of $G_{\mathcal{T}}$ into reconstructions in an incremental way and handle loop closing implicitly. In contrast to methods that compute all camera positions and 3D points in one step, new images can be easily added to an existing reconstruction and the scalability is only bound by bundle adjustment and therefore numerically stable. Various methods have been proposed to speed up the bundle adjustment optimization, for example, either by reducing the number of iterations and varying views [11, 2] or by repartitioning the problem [15, 19, 7].

### 2.3.1 Identifying Most Reliable Image Triplets

We begin with a full graph $G_{\mathcal{T}}$ of triplet reconstructions and corresponding geometrically verified edges. With an unordered set of images no meta information about topology like in [4] is given. One obvious reconstruction strategy would be to use a MST or skeletal graph (the skeletal graph also reduces the data set, a topic we are not dealing with in this work) of triplets or views connected by triplets, similarly to the epipolar equivalents of [20] or [18] .

We propose to start with the most reliable edges between nodes and start the reconstruction from these points. We select the node $N_C$ that has the highest cardinality of connected triplets, i.e. the vertex with the highest degree and search for a second node $N_{CC}$ that is connected to $N_C$. This edge will be inserted into the reconstructions. The vertex $N_{CC}$ is the one with the highest degree of the set of adjacent nodes of $N_C$. This gives us an edge that represents the vertex $N_C$ with the highest cardinality of connected triplets in the graph and the best (highest cardinality) adjacent node $N_{CC}$. An edge is selected with this strategy, merged into a reconstruction and removed from $G_{\mathcal{T}}$. This is repeated until $G_{\mathcal{T}}$ is empty.

The advantage of this strategy is twofold. Firstly, all edges in $G_{\mathcal{T}}$ are to some degree compatible and the node with the highest cardinality is therefore a highly reliable starting point and is located in a neighborhood with high information redundancy. The same observation is valid for $N_{CC}$. This is a simple measure of reliability that incorporates more information/views than searching simply for the best edge weight based on feature correspondences between two geometric primitives. We focus on the "easy" parts of the problem first. Errors introduced due to mismatches are usually introduced at a late stage in reconstruction and do not affect the robust cost functions during bundle adjustment. Secondly, this strategy reduces the number of reconstruction merging operations because the probability of selecting an edge with a node already present in a reconstruction is maximised. This is of course only valid for edge selection strategies that do not create an insertion ordering by spanning the graph.

The only exception to the cardinality strategy is made when an overlap of cameras in different reconstructions is detected. If cameras are present in multiple reconstructions, edges in $G_{\mathcal{T}}$ connecting these reconstructions are preferred. This reduces the number of cameras and 3D points that are reconstructed multiple times.

### 2.3.2 Integrating Local Correspondence Information

The edges selected by the cardinality strategy have to be merged into potentially existing reconstructions. Four different update strategies exist: (i) If both nodes are not present in a reconstruction, a new one is created. (ii)If one triplet of the edge $e_{ij}$, connecting triplet $i$ and $j$, is present in a reconstruction new correspondences and potentially new cameras have to be added. Adding new correspondences of existing cameras is straightforward. The correspondences from $e_{ij}$ are already local inliers and are simply added to existing tracks or triangulated using the cameras from the reconstruction if the track is new. If new cameras have to be added, the reconstructions can be transformed into the global system by using the similarity transform $T_r$ from the local edge to edge registration and a similarity transform $T_g$ transforming the triplet that is already in the reconstruction into the global one. No outlier rejection is done using the structure or cameras of the global reconstructions. (iii) If both triplets of the edge $e_{ij}$ are already present in the same reconstruction, only new correspondences have to be added. (iv) If both triplets of the edge $e_{ij}$ are already present in different reconstructions, $R_k$ and $R_l$, these reconstructions should be merged.

*Merging Reconstructions:* The reconstructions $R_k$ and $R_l$ should be merged if an edge $e_{ij}$ connecting them is inserted. A similarity transform is computed using RANSAC and the reprojection errors of overlapping cameras and the involved 3D points. The reconstruction with fewer cameras is transformed into the coordinate system of the larger reconstruction. New cameras and tracks of the smaller reconstruction are added in a similar way as in the triplet merging process.

*Adding New Images:* Because the reconstruction is built incrementally, it is possible to add new sets of images. The new images can be matched with the old and new ones and the epipolar graph is extended. The new triplet edges are added to $G_{\mathcal{T}}$ and merged into the existing reconstructions.

| Data Set | Images | Triplets | LCCC |
|---|---|---|---|
| SoL | 248 | 1724 | 139 |
| Opera | 347 | 1951 | 304 |
| Cathedral | 1920 | 7395 | 564 |

Table 1. *Overview of data sets.* This Table shows the number of input images, reconstructed triplets and the number of views in the largest connected component (LCCC).

*Implementation Aspects:* Bundle adjustment is used to integrate the correspondences and cameras. Given the reconstruction problem $\mathbf{x}_j^i = P^i \mathbf{X}_j$ where the 2D point measurements $\mathbf{x}_j^i$ are the observations of unknown 3D points $\mathbf{X}_j$ observed in the unknown cameras $P^i$, bundle adjustment is defined as the (in practice local) minimum of the cost function $\mathcal{C}(P^i, \mathbf{X}_j) = \sum\limits_i \sum\limits_j v_{ij} d(P^i \mathbf{X}_j, \mathbf{x}_j^i)^2$, where $v_{ij}$ is a binary variable that is 1 if the point $\mathbf{X}_j$ is visible in image $P^i$ and 0 otherwise. We replace $d(P^i \mathbf{X}_j, \mathbf{x}_j^i)^2$ with two robust cost functions. These two types of bundle adjustment iterations are: Huber cost function $\gamma_h(e)$ iterations and saturated error $\gamma_s(e)$ (also called Blake-Zisserman cost function) iterations. The two cost functions with inlier threshold $b$ are:

$$\begin{aligned} \gamma_h(e) &= \begin{cases} e^2 & \text{if } e < b \quad \text{inlier} \\ 2b|e| - b^2 & \text{if } e \geq b \quad \text{outlier} \end{cases}, \\ \gamma_s(e) &= \begin{cases} e^2 & \text{if } e < b \quad \text{inlier} \\ b & \text{if } e \geq b \quad \text{outlier} \end{cases} \end{aligned} \tag{2}$$

$\gamma_h(e)$ bundle adjustment iterations are used after a new camera is added. This cost function is robust to outliers and still establishes the global topology when loops are closed. $\gamma_s(e)$ iterations are only used after at least 10 cameras have been added and strong outliers (we set $b = 5$ for $\gamma_h(e)$ and $b = 25$ for $\gamma_s(e)$) are then removed. This removes mainly contradictory feature tracks (usually two concatenated tracks that have length two in each triplet) and prevents error build up that can negatively affect the Huber cost function.

## 3. Results

*Data Sets:* We have tested our method with several large image collections. Table 1 summarizes basic properties of three of those collections. The data sets Opera and SoL contain mostly images from the one scene, the data set Cathedral contains a large number of images from different in- and outdoor locations. Table 2 summarizes timing results and the number of triangulated points for some data sets.

*Reconstructions:* Table 3 presents some results obtained with our method and shows a comparison with the publicly available bundler software [1]. The topological properties of the data sets are reflected in the visualizations of the trifocal

[1] bundler 0.3 http://phototour.cs.washington.edu/bundler/

| Data Set | $G_{\mathcal{E}}$ | $G_{\mathcal{T}}$ | LCCP2 | Merge |
|---|---|---|---|---|
| SoL | 187 min | 13 min | 23210 | 5 hrs |
| Opera | 8 hrs | 30 min | 116961 | 7 hrs |
| Cathedral | 50 hrs | 2 hrs | 321000 | 40 hrs |

Table 2. *Performance overview.* This Table shows timing results for the epipolar and trifocal graph creation, the number of triangulated points visible in at least two views of the largest connected component (LCCP2) and the timing results of the complete edge merging/reconstruction steps (Merge). The timing results were obtained on an Intel Pentium D CPU with 3.00 GHz.

and epipolar graphs for the data sets. Nearly all images of the Opera data set are connected and the topological structure is present in the graph representations. The diverse nature of the images of the Cathedral data set are also visible in the corresponding trifocal graph $G_{\mathcal{T}}$. The last column, City Block, shows an experiment around a closed loop in a city. The image data set is weakly linked at some points. Our method obtains two reconstructed blocks for the data set. The first graph of the City Block column shows the epipolar graph $G_{\mathcal{E}}$ of the complete data set. The topological structure of the loop is correct but not all epipolar geometries are suitable for reconstruction. The second graph of the City Block column shows the trifocal graph $G_{\mathcal{T}}$. The data is split into two parts that can be reconstructed reliably.

The comparison with the bundler software shows that [17] has difficulties with data sets that have low redundancy. These kind of data sets are prone to build up of drift. In the Opera data set drift builds up and bundler is not able to close the loop. The Cathedral data set works also well with the bundler approach because the images of the cathedral are very well textured. In the City Block experiment, bundler follows the epipolar links around the graph and reconstruction fails at the geometrically weak points at both ends of the reconstruction.

Unordered community photo collections of famous locations are usually densely sampled and consist of many redundant views. The image data sets we use here were obtained at least to some degree with reconstruction in mind and are more of a sequential nature because we want to sample larger parts of a city for image based localization and therefore want to take as few images as possible. This makes the reconstruction process more challenging and prone to drift.

## 4. Conclusion

We have introduced a novel method for building the core structure of image based reconstructions from unordered sets of input images incrementally and robustly. The main contributions of the proposed method are (i) the implicit loop closing while building the reconstructions incrementally by using only correspondence information from locally matched image triplets and (ii) a strategy to identify and start with reliable subsets of images and their corresponding geometric primitives that have the highest mutual compatibility consensus. Experimental results show that the algorithm can build large reconstructions incrementally without depending on global structure to separate in- and outliers. Furthermore, experimental results indicate that this strategy provides superior results when repetitive structure and a high amount of feature matching outliers (even in image triplets) are present.

## 5. Acknowledgments

The use of Bundler by N. Snavely is gratefully acknowledged. In part funded by FWF contract W1209 and Christian Doppler Laboratory for Handheld Augmented Reality.

## References

[1] C. Beder and R. Steffen. Determining an initial image pair for fixing the scale of a 3d reconstruction from an image sequence. In *DAGM*, pages 657–666, 2006.

[2] H. S. C. Engels and D. Nister. Bundle adjustment rules. In *Photogrammetric Computer Vision*, 2006.

[3] E. D. Eade and T. W. Drummond. Unified loop closing and recovery for real time monocular SLAM. In *BMVC*, 2008.

[4] A. W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open omage sequences. In *Proc. ECCV*.

[5] B. Horn, H. Hilden, and S. Negahdaripour. Closed-form solution of absolute orientation using orthonormal matrices. *J. Opt. Soc. Am. A*, 5:1127–1135, 1988.

[6] A. Irschara, C. Zach, and H. Bischof. Towards wiki-based dense city modeling. In *Workshop on Virtual Representations and Modeling of Large-scale environments (VRML)*, 2007.

[7] F. Kai Ni Steedly, D. Dellaert. Out-of-core bundle adjustment for large-scale 3d reconstruction. In *Proc. ICCV*, pages 1–8, 2007.

[8] X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *Proc. ECCV*, 2008.

[9] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[10] D. Martinec and T. Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *Proc. CVPR*, 2007.

[11] E. Mouragnon, F. Dekeyser, P. Sayd, M. Lhuillier, and M. Dhome. Real time localization and 3d reconstruction. In *Proc. CVPR*, pages 363–370, 2006.

[12] D. Nistér. An efficient solution to the five-point relative pose problem. *TPAMI*, 26(6):756–770, 2004.

[13] D. Nistér and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. CVPR*, pages 2161–2168, 2006.

[14] G. Schindler, M. Brown, and R. Szelisk. City-scale location recognition. In *Proc. CVPR*, 2007.

[15] H.-Y. Shum, Z. Zhang, and Q. Ke. Efficient bundle adjustment with virtual key frames: A hierarchical approach to multi-frame structure from motion. *Proc. CVPR*, 2:2538, 1999.

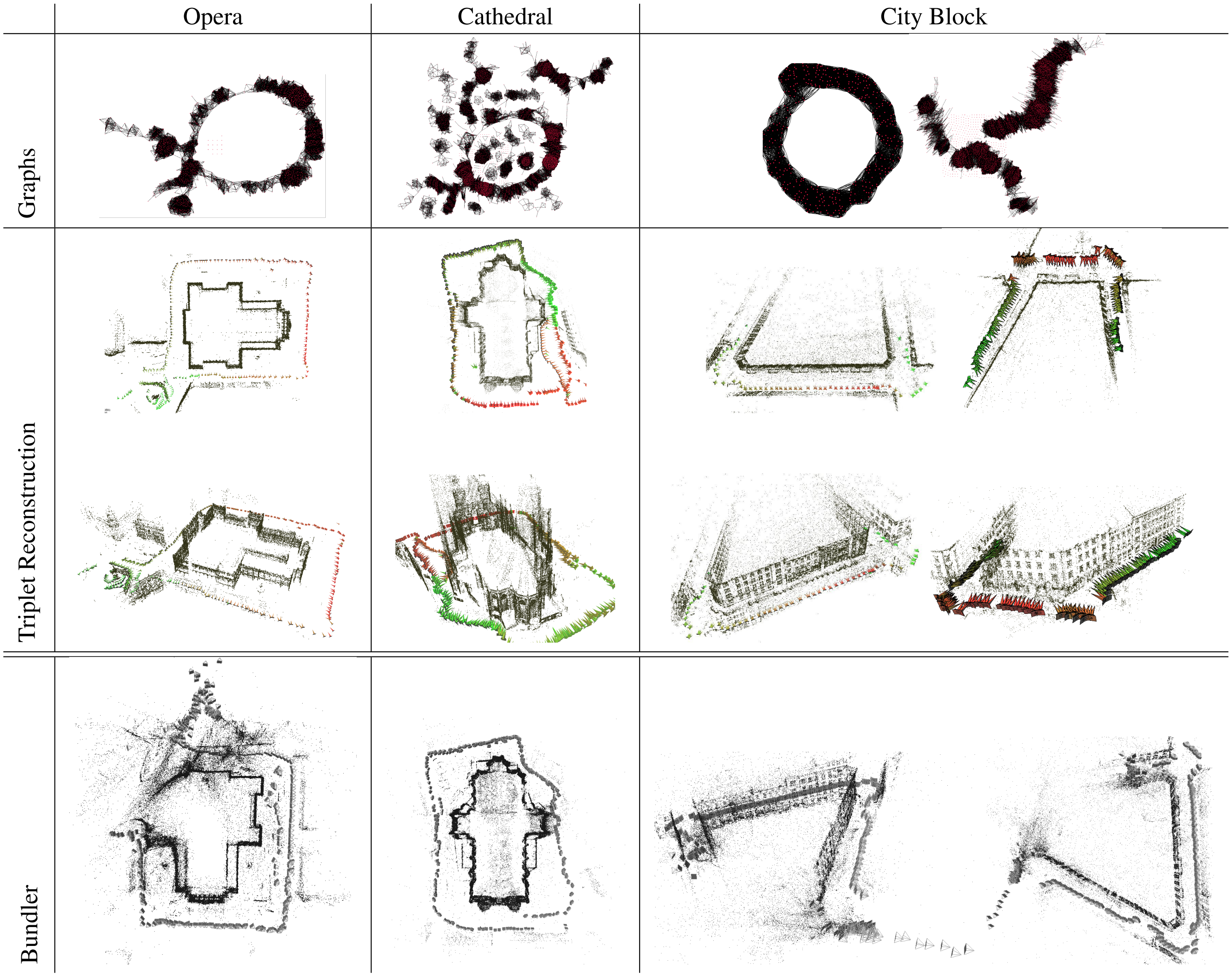


Table 3. *Results and comparison with bundler software.* The first and second column present results for the Opera and Cathedral data sets. The third column City Block shows two results of the same data set and a comparison of the epipolar and trifocal graph topology. *Opera:* The trifocal graph shows that nearly all images of the Opera data set are connected and the topological structure is present in the graph representation. Bundler fails to close the loop on this data set. *Cathedral:* The diverse nature of the images of the Cathedral data set are also visible in the corresponding trifocal graph. Bundler succeeds on this data set. *City Block:* The first graph of City Block shows the epipolar graph $G_{\mathcal{E}}$ and the second graph the corresponding trifocal graph $G_{\mathcal{T}}$. Note that the epipolar graphs shows the correct topological structure of the data set, but some images are weakly linked for reconstruction. This can be seen in the topology of the trifocal graph. The loop is split into two parts and our reconstruction method obtains two major blocks of the data set. We show two views of the result obtained with bundler, the reconstruction diverged at both ends of the structure.

[16] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, pages 1470–1477, 2003.

[17] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3D. In *Proceedings of SIGGRAPH 2006*, pages 835–846, 2006.

[18] N. Snavely, S. M. Seitz, and R. S. Szeliski. Skeletal graphs for efficient structure from motion. In *CVPR*, pages 1–8, 2008.

[19] D. Steedly, I. Essa, and F. Delleart. Spectral partitioning for structure from motion. In *Proc. ICCV*, page 996, Washington, DC, USA, 2003. IEEE Computer Society.

[20] K. L. Steele and P. K. Egbert. Minimum spanning tree pose estimation. In *Proc. 3DPVT*, pages 440–447, 2006.

[21] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – A modern synthesis. In *Vision Algorithms: Theory and Practice*, pages 298–375. 2000.

[22] C. Zach, A. Irschara, and H. Bischof. What can missing correspondences tell us about 3D structure and motion? In *Proc. CVPR*, pages 1–8, 2008.