HSfM: Hybrid Structure-from-Motion

Hainan Cui[1], Xiang Gao[1, 2], Shuhan Shen[1, 2], and Zhanyi Hu[1, 2, 3]

[1]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing , China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{*hncui, xiang.gao, shshen, huzy*}*@nlpr.ia.ac.cn*

Abstract

Structure-from-Motion (SfM) methods can be broadly categorized as incremental or global according to their ways to estimate initial camera poses. While incremental system has advanced in robustness and accuracy, the efficiency remains its key challenge. To solve this problem, global reconstruction system simultaneously estimates all camera poses from the epipolar geometry graph, but it is usually sensitive to outliers. In this work, we propose a new hybrid SfM method to tackle the issues of efficiency, accuracy and robustness in a unified framework. More specifically, we propose an adaptive community-based rotation averaging method first to estimate camera rotations in a global manner. Then, based on these estimated camera rotations, camera centers are computed in an incremental way. Extensive experiments show that our hybrid method performs similarly or better than many of the state-of-the-art global SfM approaches, in terms of computational efficiency, while achieves similar reconstruction accuracy and robustness with two other state-of-the-art incremental SfM approaches.

1. Introduction

Structure-from-Motion (SfM) technique is to estimate the 3D scene structure and camera poses from a collection of images [2, 38]. It usually consists of three modules: features extraction and matching, initial camera poses estimation, and bundle adjustment. According to the difference of initial camera poses estimation manner, SfM can be broadly divided into two classes: incremental and global.

For incremental SfM approaches, one way [34, 38] is to start from selecting a few seed images for initial reconstruction, then repeatedly add new images. Another way [21, 41] is to cluster the images into atomic models first, then reconstruct each atomic model and incrementally merge them after. Arguably, incremental manner is the most popular strategy for 3D reconstruction [22, 39]. However, it is sensitive to the initial seed model reconstruction and the manner of model growing. In addition, the reconstruction error is accumulated with the iterations going on. For large-scale scene reconstruction, the reconstructed structure may suffer from scene drift [24]. Furthermore, the time-consuming bundle adjustment (BA) [42] is repeatedly performed, which dramatically decreases the system scalability and efficiency. To tackle these weaknesses, global SfM approaches become popular in the past few years.

For global SfM approaches [31, 13], initial camera poses are estimated simultaneously from the epipolar geometry graph (EG), whose vertices correspond to images and edges link matched image pairs, and the bundle adjustment is performed only once, which brings a better potential in system efficiency and scalability. The generic pipeline for global camera poses estimation consists of two steps: rotation averaging and translation averaging. For the rotation averaging, its accuracy mainly depends on two factors: the structure of EG and the accuracy of pairwise epipolar geometries [43]. Currently many literatures [5, 17] only minimize the residuals on the epipolar edges. As a result, when the cameras are not well distributed, for example the Internet data [44], those methods sometimes perform poorly. For the translation averaging, since epipolar geometry only encodes the direction of pairwise translation, it is difficult to determine camera positions. Moreover, the translation estimation is more sensitive to feature match outliers. In comparison, incremental SfM approaches benefit from RANSAC technique to discard bad epipolar geometries. Thus, it is desirable to take the advantages of both incremental and global manners.

Contribution: (1) we propose a new **hybrid SfM** approach to tackle the issues of efficiency, robustness and accuracy in a unified framework; (2) a community-based rotation aver-

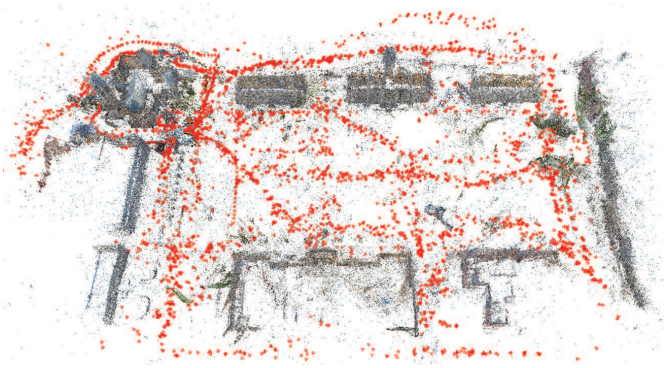

Figure 1. Result of Quad [7] with 5061 images registered out of 5520 images, where the calibration time-cost of our hybrid method is about 55 mins and the median calibration accuracy is 1.03m.

aging method is proposed in a global manner, which considers both the structure of EG and the accuracy of pairwise geometries; (3) based on the estimated camera rotations, camera centers are estimated in an incremental way. For each camera addition, both camera rotations and intrinsic parameters are kept as constant, while the camera centers and scene structure are refined by a modified bundle adjustment.

In our hybrid SfM, global rotation averaging decreases the risk of scene drift, and incremental centers estimation increases robustness to the noisy data. With known camera rotations, camera centers registration only needs two scene points, thus the RANSAC technique makes our method become more robust to outliers and more cameras could be calibrated at each camera adding step. Additionally, since only the scene structure and camera centers are refined in each camera addition, the bundle adjustment in our hybrid work is much faster than the conventional one [38, 39].

In the experiments, we evaluate our hybrid SfM system on both sequential and unordered image data. Our method outperforms many recent state-of-the-art global SfM methods [11, 31, 40, 44], in terms of the number of reconstructed cameras, indicating that our method is more robust to outliers. In terms of reconstruction efficiency, our method performs similarly or better than the global SfM methods, while it is up to 13 times faster than a parallelized version of Bundler [38], and 5 times faster than the parallelized version of Theia [39]. Fig. 1 illustrates a reconstruction result on the public dataset Quad [7], where more than 5K images are calibrated by our hybrid SfM system. With a comparable calibration accuracy, the speed of our hybrid method is 50 times faster than Bundler [38], and 7 times faster than DISCO [7].

2. Related Work

Incremental SfM methods One way to reconstruct the scene starts from two or three "seed" views, then incrementally add new views into system to initialize the final BA [2, 16, 28, 34, 35, 38, 45, 48]. Such approaches are sensitive to the seed selection criteria, and accumulated error may cause scene drift. To decrease the accumulated errors, both VSFM [45] and COLMAP [34] proposed to re-triangulate tracks in the image adding process. Another way [20, 21, 41] is to create atomic 3D models first, and then merge different models. Such hierarchical methods are sensitive to the atomic model selection and model growing scheme. For large image collections, all the incremental methods suffer from scene drift and heavy computation load due to the repeated activation of bundle adjustment.

Global SfM methods Global SfM approaches [3, 7, 10, 11, 23, 29, 31, 33, 40, 44] simultaneously estimate all the camera poses and perform bundle adjustment only once. The camera poses estimation process mainly contains two parts: rotation averaging and translation averaging.

Rotation Averaging Rotation averaging estimates all the camera rotations from pairwise relative rotations simultaneously, which has been well studied in many literatures [5, 14, 17, 27, 43]. Martinec *et al*. [27] proposed to solve this problem under Frobenius norm, and Govindu [14] proposed to transform the rotation averaging problem into a Lie-algebraic averaging. Based on [14], better result is achieved by combining with robust L1 optimization [5]. Recently, Wilson *et al*. [43] found two factors impacting the rotation estimation accuracy, one is the EG structure and the other is the epipolar geometry accuracy, and recommended to cluster cameras first when they distributed unevenly. However, they did not describe how to group and merge images effectively based on this theory. Inspired by this theoretical analysis, we propose a community-based rotation averaging method to automatically determine when and how to cluster, and followed by a greedy merging step.

Translation Averaging Many linear methods [3, 23, 33] proposed to solve the camera positions by matrix decomposition. While efficient, such approaches are sensitive to epipolar geometry outliers. Hence, many global SfM approaches [15, 46, 47] carefully filtered the erroneous edges first. Zach *et al*. [47] proposed to filter edges by loop consistency check, and Wilson *et al*. [44] presented a hashing-like method, called 1DSfM. However, this method [44] requires abundant pairwise associations ($O(n^2)$). Instead of filtering, some methods refined the epipolar relations by local bundle adjustment [11] or multi-view track consistency [40]. Other methods [7, 11, 37, 25] solved the scene points and camera centers together. In this way, not only the collinear motion problem is solved, but all the cameras are fused into a connected parallel-rigid graph [31]. Besides, methods [4, 8, 9, 32, 37] fused auxiliary imaging information to obtain the camera centers. While efficient and scalable, they are heavily relying on the auxiliary information.

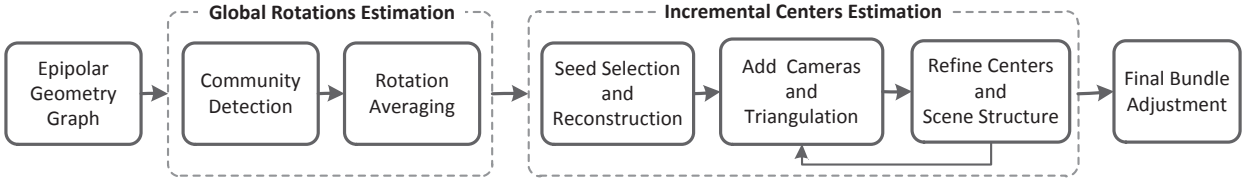

Figure 2. Pipeline of our hybrid SfM system.

3. Overview of Hybrid SfM

Considering that incremental approaches are usually more robust and accurate due to its repeated optimization by bundle adjustment, but its computational load is prohibitively large if the image dataset is large, while global approaches are adept for the estimation of all the rotations, but error-prone to the outliers in the camera centers estimation, here we propose a hybid SfM by taking the advantages of both the incremental and global schemes.

As shown in Fig. 2, the input of our system is the epipolar geometry graph (EG), which includes the pairwise matches on each epipolar edge, and corresponding pairwise geometry estimated from essential matrix decomposition. For example, the essential matrix on edge (i, j) encodes the relative rotation $\mathbf{R}_{ij}$ and the relative translation direction $\mathbf{t}_{ij}$, which is constrained by the following equations:

$$\begin{aligned} \mathbf{R}_{ij} &= \mathbf{R}_j \mathbf{R}_i^T, \\ \lambda_{ij}\mathbf{t}_{ij} &= \mathbf{R}_j(\mathbf{C}_i - \mathbf{C}_j), \end{aligned} \tag{1}$$

where $\mathbf{C}_i$ and $\mathbf{R}_i$ correspond to the camera center and rotation of image i. The equation between the global camera rotations and relative rotations could be transformed into Lie-algebraic space first, then solved using L1 optimization [5]. However, as demonstrated in [43], the accuracy of global rotation averaging is sensitive to both the structure of EG and the accuracy of pairwise geometries. Thus, in the first module of our hybrid SfM in Fig. 2, we propose a community-based rotation averaging method to take account of these two factors.

For camera centers estimation, since the scale factor λ_{ij} is unknown, it is difficult to estimate the camera centers directly, and [31] proved that the essential matrices only determine camera positions in a parallel rigid graph. In addition, the translation estimation is sensitive to erroneous feature matches. Thus, in the second module of our hybrid SfM in Fig. 2, we use an incremental manner, which benefits from RANSAC method to exclude erroneous feature matches, to estimate the camera centers.

When no more cameras could be added, a final bundle adjustment is performed to refine all camera intrinsic parameters, camera poses and the scene structure.

4. Global Rotation Estimation

For sequential images, the connections are commonly evenly distributed. However for unordered images, for example images from Internet [44], the cameras distribution are usually uneven, *e.g.* the place of interest usually gets more attention. As a result, if there are many interested buildings in the scene, the overall connections between buildings become sparse, while denser for each one. To tackle the uneven camera distribution problem, we propose an automatical grouping method inspired by techniques in complex networks analysis. Then, rotation averaging is performed for each community, and an alignment step is followed to fuse them into a united coordinate system.

4.1. Community Detection

Community detection [12, 6] has been widely used in the complex networks analysis, which aims to divide a graph into groups with denser connections inside and sparser connections outside. Let A_{ij} be an element of the adjacent matrix of our epipolar geometry graph (EG). $A_{ij} = 1$ if an edge exists between camera i and camera j, otherwise $A_{ij} = 0$. The degree of node i in the EG is the number of cameras that connect to it, denoted as $d_i = \Sigma_j A_{ij}$. Let $m = \frac{1}{2}\sum_{ij} A_{ij}$ be the total number of edges in EG. If EG is randomized without a community structure, the probability of an edge existing between camera i and camera j is $\frac{d_i d_j}{2m}$ [6]. To measure the difference of the fraction of intra-community connections between EG and the random graph, we use the modularity indicator Q proposed in [6]. Suppose that camera i belongs to a community S_p and camera j belongs to a community S_q, then Q is defined as:

$$Q = \frac{1}{2m}\sum_{ij}(A_{ij} - \frac{d_i d_j}{2m})\delta(S_p, S_q), \tag{2}$$

where $\delta(S_p, S_q) = 1$ if S_p = S_q and 0 otherwise. To enhance the impact of good edges with more matches inlier, we use a weighting adjacent matrix and the edge weight A_{ij} is set to $\sqrt{N_{ij}}$, where N_{ij} is the number of feature match inliers between camera i and camera j. To partition the EG, we assume each node belongs to a sole community first, then iteratively joins separate communities whose amalgamations result in the largest increase in Q [36]. As pro-

posed in [6], the modularity has a single peak Q_{max} over the generation of the dendrogram which indicates the most significant community structure. In practice, we found that $Q_{max} > 0.4$ indicates that EG has a significant community structure. Thus, we take the partition result when the peaking value is larger than 0.4. Otherwise when $Q_{max} < 0.4$, all the cameras are just considered as one community.

4.2. Rotation Averaging

For each community, the global rotation averaging method proposed in [5] is used for rotation averaging. As a result, the estimated rotations for each community is under a different coordinate system. When we have two or more communities, an alignment should be performed to put them into a unified coordinate system. The transformation between any pair of communities is a rotation matrix in $SO(3)$, but there are usually many edges between two communities in the original epipolar geometry graph. Thus, we propose a voting scheme to find the best transformation for each pair of communities. For each edge between two communities, we get one possible rotation transformation candidate. Then, based on this candidate rotation, the residual of other edges between this pair of communities are calculated. The best transformation is the one which has the most inliers, where the inlier is defined as the edge whose residual angle is less than 15 degrees.

After all the transformations between connected communities are obtained, the original epipolar graph is simplified as a weighted community-graph with nodes corresponding to communities and edges linking communities with connections, the weight on each edge is defined as the ratio of inliers of the best transformation corresponding to. We set the reference coordinate system as the node with the largest degrees, and construct the maximal spanning tree (MST) of this community-graph for the alignment. Based on the MST, the rotations of other communities are aligned to the reference community. Fig. 3 shows our community-based rotation averaging result on the Gendarmenmarkt [44] dataset, where each curve denotes the cumulative distribution function (CDF) of global camera rotation errors. From the result of L1RA [5], which corresponds to the case that all cameras are considered as a sole community, we can see that its rotation estimation is erroneous. However, after it is divided into four communities, the estimated rotation for each community becomes more accurate. Our final result is shown in red, which significantly improves the camera rotation accuracy.

5. Incremental Centers Estimation

Once the rotation of each camera is achieved, for the robustness concern, we estimate the camera centers in an incremental way. With the estimation process going on, the scene structure is also reconstructed. In the next, we in-

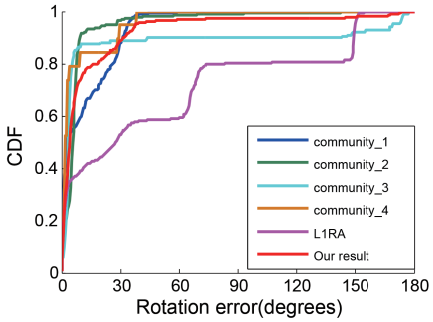

Figure 3. The cumulative distribution function (CDF) of global camera rotation errors for the Gendarmenmarkt [44] data, whose epipolar edges contain significant rotation errors.

troduce three constraints to select a pair of good cameras for initial reconstruction first. Then, based on the correspondence between the estimated scene structure and tracks across images, camera centers are iteratively estimated and refined by the modified bundle adjustment.

5.1. Initial Camera Selection and Reconstruction

To obtain a good initial reconstruction, a pair of cameras should satisfy three constraints: more feature matches, wide baseline and accurate camera poses. All edges in the EG could be considered as the candidates for our initial camera pair selection. Thus, we augment EG by tagging each edge with these three constraints.

First, the inlier number of feature matchings at each edge in EG, which is verified by the 5-point algorithm [30], is recorded. Then, with known rotations, the angle of a pairwise normalized feature match $(\mathbf{p}_i, \mathbf{p}_j)$ between camera i and camera j is computed by: $acos(\mathbf{R}_i^T \mathbf{p}_i, \mathbf{R}_j^T \mathbf{p}_j)$. For each edge in EG, we compute all the angles corresponding to its feature matches and record the median value to indicate the length of baseline.

Since the ground-truth camera poses are not known, we cannot find an image pair with the real best camera poses. However, the rotation averaging in each community [5] could be regarded as finding the best camera rotation for each camera that minimizes the median residual of its connected edges, which means that the more edges the camera connect, the more accurate the estimated rotation possibly would be. Let n_i be the number of neighbors of the camera i in the EG, the camera poses accuracy of an edge between camera i and camera j is indicated by $\sqrt{(n_i^2 + n_j^2)/2}$.

Based on these three indicators on each edge, we propose a cascaded scheme to select an initial camera pair. We consider the camera poses accuracy first to get an accurate initial reconstruction. However, even with accurate camera poses, if the baseline is small, the reconstruction still suffer. Thus, we choose the baseline factor with the second priority for the scene structure concern. At last, we consider the

number of match inliers. In practice, the pose accuracy indicators are sorted in a descending order first, and we only select the first α_1 edges (in our work, α_1 is set to 60%). Then, the edges with a median angle less than 10 degree is discarded in the initial camera selection to avoid the pure rotation problem. Finally, for all the remaining edges, we choose an edge with the largest number of image matches.

For the selected initial pair camera i and camera j, we recalculate its relative rotations $\mathbf{R}_{ij} = \mathbf{R}_j\mathbf{R}_i^T$ where $\mathbf{R}_i$ and $\mathbf{R}_j$ are the results from rotation averaging. Then, by fixing this new relative rotation, the corresponding relative translation $\mathbf{t}_{ij}$ are refined by solving the linear system $\mathbf{p}_j^T[\mathbf{t}_{ij}]_\times\mathbf{R}_{ij}\mathbf{p}_i = 0$, which only needs two points to solve. Based on the RANSAC technique, $\mathbf{t}_{ij}$ is refined and the feature matches between this image pair are re-evaluated by the distance from image point to its new epipolar line. After the verification, the feature matches inlier are triangulated and refined by a modified bundle adjustment which only refines camera centers and currently reconstructed scene structure, as shown in Sec. 5.4.

5.2. Camera Registration

Based on the known camera rotations, camera centers could be estimated by only two scene points. The reason is that for a camera, with the known rotation and its two scene points, two projection rays from the scene points both go through the camera center. For example, for the camera i with a visible scene point $\mathbf{X}_j = \{\mathbf{X}_{jx}, \mathbf{X}_{jy}, \mathbf{X}_{jz}\}$ and its corresponding image coordinates $\mathbf{x}_{ij}$, the projection equation could be transformed into:

$$\mathbf{X}_j - \mathbf{C}_i = \lambda\mathbf{R}_i^T\mathbf{K}_i^{-1}\mathbf{x}_{ij}. \tag{3}$$

Let $\mathbf{h}_i = \{\mathbf{h}_{ix}, \mathbf{h}_{iy}, \mathbf{h}_{iz}\} = \mathbf{R}_i^T\mathbf{K}_i^{-1}\mathbf{x}_{ij}$ and the camera center $\mathbf{C}_i = \{\mathbf{C}_{ix}, \mathbf{C}_{iy}, \mathbf{C}_{iz}\}$, then we get the following two independent equations

$$\begin{aligned}(\mathbf{X}_{jx} - \mathbf{C}_{ix})\mathbf{h}_{iz} &= (\mathbf{X}_{jz} - \mathbf{C}_{iz})\mathbf{h}_{ix},\\ (\mathbf{X}_{jy} - \mathbf{C}_{iy})\mathbf{h}_{iz} &= (\mathbf{X}_{jz} - \mathbf{C}_{iz})\mathbf{h}_{iy}.\end{aligned} \tag{4}$$

Since the DOF (degrees of freedom) of camera center is 3, we need two points at least. For each camera observing more than two scene points, we use RANSAC technique to find the best camera center which has the largest number of visible scene points inliers, whose projection errors are less than γ_1 pixels. Then, two constraints are further checked: the number of inliers should be larger than β_1, and the corresponding inlier ratio should be larger than β_2. If both the constraints are satisfied, we consider the camera center is estimated successfully (in our work, β_1 = 16, β_2 = 60%).

Sometimes though the number of visible points is large (*e.g.* larger than 30), the estimated camera center is still wrong because some estimated camera rotations may be not accurate enough. In this case, we use a $P3P$ [26] method to find a possible camera pose. Similarly, if both the inliers number and ratio satisfy the above two constraints, we update the corresponding camera rotation and center. Note that since the camera rotation is estimated by the scene points, it is still in the original rotation coordinate system.

5.3. Triangulation

After some new cameras are added into the SfM system, all the tracks with equal or more than 2 calibrated cameras are triangulated. Here we use a RANSAC-based triangulation method. For each iteration, we randomly choose two visible views, and then check the angle between two projection rays. If the angle is larger than 3 degrees, we consider it is currently well-conditioned and use the DLT [18] method to triangulate. Then, after we get a scene point triangulation, both the number of its consistent measurements and the corresponding view cheirality are checked. Note that all the cheirality [19] of calibrated cameras in the track should be positive, and the measurement in a track is considered as consistent to the current scene point estimation if the corresponding re-projection error is less than γ_1 pixels. For each track, we find the best scene point which has the largest number of consistent measurements.

5.4. Refine Camera Centers and Scene Structure

To mitigate the impact of accumulated errors, we perform bundle adjustment (BA) after each camera adding and triangulation process. To account for the potential outliers, *Huber* function is employed as the loss function in our BA.

Formulation To guarantee the camera rotation coordinate system fixed in the camera adding process, we only refine the camera centers and reconstructed scene structure by keeping intrinsic parameters and camera rotations unchanged. Thus, the modified bundle adjustment is formulated as:

$$\underset{\mathbf{C}_i,\mathbf{X}_j}{\textbf{minimize}} \quad \sum_{i=1}^{N}\sum_{j=1}^{M}\delta_{ij}\|\mathbf{x}_{ij} - \gamma(\mathbf{K}_i, \mathbf{R}_i, \mathbf{C}_i, \mathbf{X}_j)\|_{huber}, \tag{5}$$

where $\delta_{ij} = 1$ if camera i observes scene point j, otherwise $\delta_{ij} = 0$. $\mathbf{K}_i, \mathbf{R}_i, \mathbf{C}_i$ corresponds to the intrinsic camera matrix, rotation and center of the camera i, respectively. $\gamma(\mathbf{K}_i, \mathbf{R}_i, \mathbf{C}_i, \mathbf{X}_j)$ is the projection function, and $\mathbf{x}_{ij}$ denotes the measured 2D image point positions.

Re-Triangulation Similar to VSFM [45] and COLMAP [34], we perform a re-triangulation step for tracks to decrease the accumulated error because the camera centers become more accurate after bundle adjustment.

Further BA With a new scene structure computed by re-triangulation, we perform another BA to obtain more accurate camera poses. After this BA, re-triangulation step is performed again, and the tracks with large reprojection errors are filtered.

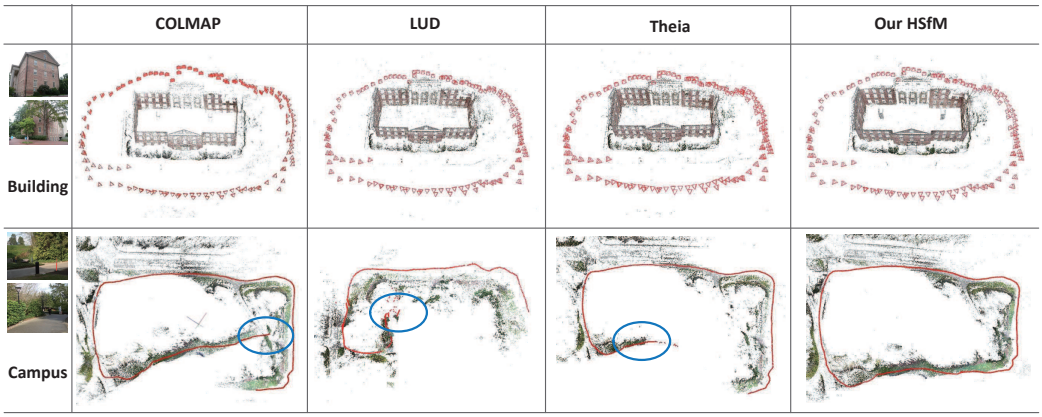

Figure 4. Reconstruction results on sequential image data: Building [47] and Campus [11].

6. Experiment

All of our experiments are performed on a PC with an Intel Xeon E5-2603 2.50GHz CPU(4 cores) and 32G RAM. We use Ceres Solver [1] for bundle adjustment, and the error threshold γ_1 for inlier judgement is set to 16 pixels.

6.1. Evaluation on Sequential Image Data

Many global SfM methods simultaneously estimate the initial camera poses using triplet constraints, thus they are robust to drift error but usually some images are left uncalibrated. Incremental SfM methods are robust to outliers and do not depend on the image triplet, but the accumulated error cannot be avoided. We demonstrate our hybrid SfM system on two public sequential image datasets: the ***Building*** dataset with 128 images from [47], the ***Campus*** dataset with 1040 images from [11], and compare our method with two state-of-the-art incremental SfM methods: COLMAP [34] and Theia [39], and a state-of-the-art global SfM method LUD [31]. Fig. 4 shows the reconstruction results on these two image datasets.

From the results of ***Building*** which has more clean epipolar geometries, we can see that all the methods in comparison could successfully reconstruct the scene. However, for the dataset ***Campus***, there are many trees and the camera trajectory is a loop. The pairwise geometries estimation for this dataset are contaminated and many track outliers are easily appeared due to the matched trees. From the results in the second row of Fig. 4, we can see that both COLMAP [34][1] and Theia [39] suffer from significant drifting error, which cannot get a loop closure. In addition, since the camera trajectory is approximately a linear motion and the pairwise geometries are sparse, the translation based method LUD [31] fails on this dataset. In comparison, our result achieves the loop closure and is effective on the collinear camera motion. Besides, we also evaluate our hybrid SfM method on a large-scale streetview dataset with 2048 images from [9], and the corresponding reconstruction result is shown in Fig 6(a).

[1]The feature matching method is set to preemptive matching.

6.2. Evaluation on Unordered Image Data

To demonstrate our hybrid SfM, we evaluate it on the Internet datasets published in [44], which contain twelve groups of medium-scale data, two large-scale data: *Piccadilly* and *Trafalgar*, and a challenging dataset *Gendarmenmarkt* with symmetric architectures. We also test on a dataset *Temple*, which has many symmetric structures and trees in scene.

We compare our method with four state-of-the-art global SfM approaches, including [44, 31, 11, 40], and two state-of-the-art incremental SfM approach Theia [38, 39]. Since these methods with the same input which is published in [44], the reconstruction accuracy and time-cost comparison are fair. However, as the result published in the literature [34] is obtained by using its own epipolar geometry graph and tracks, which is different from the data published in [44], thus for the fairness concern, we don't compare with COLMAP [34] quantitatively and just show some qualitative comparison results in Fig. 5. We use the calibration results of the state-of-the-art incremental SfM system Bundler [38] as the reference ground-truth, and the corresponding mean and median camera position errors for each method in comparison are computed, as well as the number of calibrated cameras.

The quantitative comparison results are shown in Table 1. From Table 1, we can see that our hybrid SfM method reconstructs the most number of cameras in most cases, indicating that our method is more robust to outliers. For the calibration accuracy, our method achieves similar or better accuracy than these state-of-the-art methods. Table 2 shows the corresponding reconstruction time-cost comparison, from which we can see that our method is much faster than the state-of-the-art incremental method, and performs

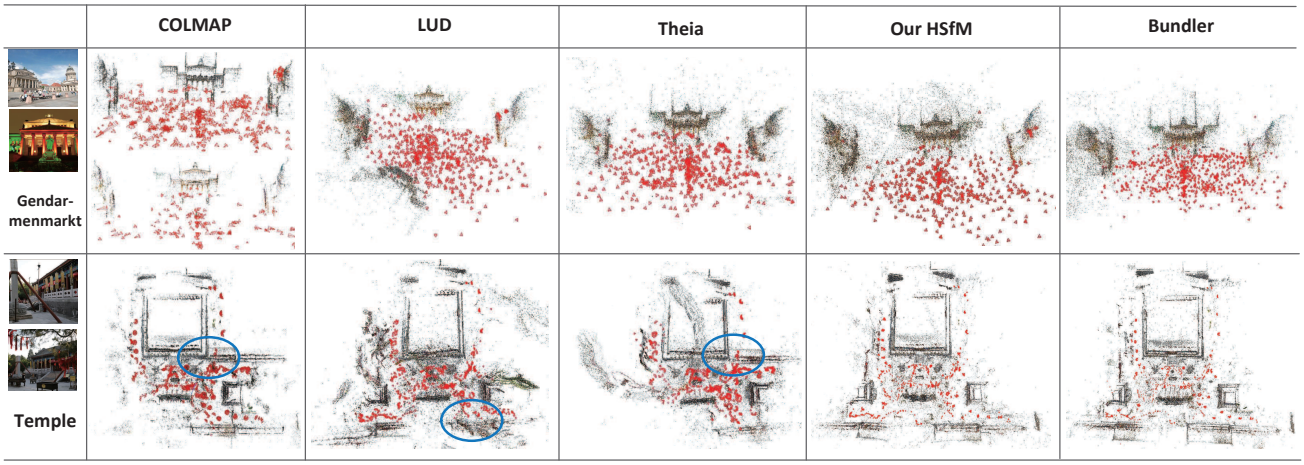

Figure 5. Reconstruction results on unordered image data: Gendarmenmarkt [44] and Temple.

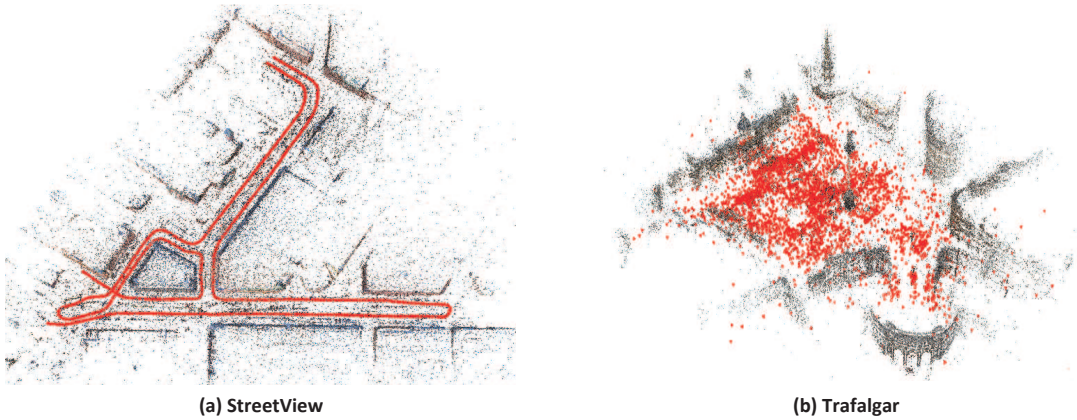

Figure 6. (a) Reconstruction result on a streetview dataset from [9]; (b) reconstruction result on the dataset Trafalgar from [44].

similarly or better than the global methods. For the dataset *Piccadilly*, our method is 13 times faster than the incremental system Bundler [38]. Considering both accuracy and time-cost, we can conclude that while our hybrid SfM inherits the robustness of the incremental manner, it also possesses the speed advantage from the global manner.

Fig. 5 shows the reconstruction results on two unordered image datasets. For the dataset *Gendarmenmarkt*, which is reported as a failure case in [44] due to its repetitive scene structures. The correct model is shown in the last column produced by Bundler [38], and with an incremental manner, Theia [39] also achieves a similar scene structure. From the other reconstruction results, we can see that LUD [31] could not get a reasonable scene structure. Though COLMAP [34] achieves a similar structure, it calibrates two different models. In comparison, our method could achieve a similar scene structure with Bundler [38], which mainly benefits from the community-based rotation averaging method. Most of the global SfM methods [44, 31, 10] use the rotation result produced by L1RA [5] directly, which assumes all the cameras just in a community. However, the median and mean rotation error for the *Gendarmenmarkt* is 40.6 degrees and 54.9 degrees, respectively, which is too erroneous for the center estimation. If we use this rotation result, our incremental center estimation model would also fail. Since the modularity value Q_{max}, which is shown in the Table 2, is larger than 0.4, it means that this dataset actually has a great community character. Based on our community-based rotation averaging method, the corresponding median and mean rotation errors decrease to 9.8 degrees and 15.9 degrees respectively, which is accurate enough for our camera center estimation module.

For the dataset *Temple*, we can see that the scene structures reconstructed by both COLMAP [34] and Theia [39] are wrong. The reason is that some feature match outliers are considered as inliers in their SfM processes, then many cameras use these points for the pose estimation. With the iteration going on, the accumulated pose errors transfer to

Table 1. Accuracy Comparison on Internet image data. $\tilde{x}$ and $\bar{x}$ respectively denote the median and mean position errors in meters for different methods by taking the result of [38] as a reference. N_i is the number of cameras in the largest connected component of input EG graph which is published in [44], and N_c is the number of reconstructed cameras. The bold font highlights the best result in each row.

Dataset		1DSfM [44]			LUD [31]			Cui [11]			Sweeney [40]		Theia [39]			Our HSfM		
Name	N_i	N_c	$\tilde{x}$	$\bar{x}$	N_c	$\tilde{x}$	$\bar{x}$	N_c	$\tilde{x}$	$\bar{x}$	N_c	$\tilde{x}$	N_c	$\tilde{x}$	$\bar{x}$	N_c	$\tilde{x}$	$\bar{x}$
Alamo	627	529	**0.3**	2e7	547	**0.3**	2.0	**574**	0.5	3.1	533	0.4	520	0.4	1.8	566	**0.3**	**1.5**
Ellis Island	247	214	**0.3**	3.0	–	–	–	223	0.7	4.2	203	0.5	210	1.7	**2.8**	**233**	2.0	4.8
Metropolis	394	291	0.5	7e1	288	1.5	4.0	317	3.1	16.6	272	**0.4**	301	1.0	**2.1**	**344**	1.0	3.4
Montreal N.D.	474	427	0.4	1.0	435	0.4	1.0	452	**0.3**	1.1	416	**0.3**	422	0.4	**0.6**	**461**	**0.3**	**0.6**
Notre Dame	553	507	1.9	7.0	536	**0.2**	0.7	549	**0.2**	1.0	501	1.2	540	**0.2**	**0.5**	**550**	**0.2**	0.7
NYC Library	376	295	0.4	**1.0**	320	1.4	7.0	338	**0.3**	1.6	294	0.4	291	0.4	**1.0**	**344**	**0.3**	1.5
Piazza del Popolo	354	308	2.2	2e2	305	1.0	4.0	340	1.6	2.5	302	1.8	290	**0.8**	**1.5**	**344**	**0.8**	2.9
Piccadilly	2508	1956	0.7	7e2	–	–	–	2276	**0.4**	2.2	1928	1.0	1824	0.6	**1.1**	**2279**	0.7	2.0
Roman Forum	1134	989	**0.2**	3.0	–	–	–	1077	2.5	10.1	966	0.7	942	0.6	**2.6**	**1087**	0.9	8.4
Tower of London	508	414	1.0	4e1	425	3.3	10.0	465	1.0	12.5	409	**0.9**	439	1.0	**1.9**	**481**	**0.9**	6.4
Union Square	930	710	3.4	9e1	–	–	–	570	3.2	11.7	701	2.1	626	**1.9**	3.7	**827**	2.8	**3.4**
Vienna Cathedral	918	770	**0.4**	2e4	750	4.4	10.0	842	1.7	4.9	771	0.6	738	1.8	3.6	**849**	1.4	**3.3**
Yorkminster	458	401	**0.1**	5e2	404	1.3	4.0	417	0.6	14.2	409	0.3	370	1.2	1.8	**421**	1.2	**1.7**
Trafalgar	5433	4957	–	–	–	–	–	4945	3.6	8.6	–	–	3873	**2.6**	**4.0**	**4966**	**2.6**	7.2
Gendarmenmarkt	742	–	–	–	–	–	–	609	4.2	27.3	–	–	597	2.9	28.0	**611**	**2.8**	**26.3**

Table 2. Running times comparison. Q_{max} denotes the modularity, and T_D, T_R, T_C, T_{BA} denote the time-cost of community detection, rotation estimation, centers estimation, final bundle adjustment, respectively. T_Σ is the total time-cost of corresponding SfM method.

Dataset		Our HSfM					1DSfM [44]	LUD [31]	Cui [11]	Sweeney [40]	Theia [39]	Bundler [38]
Name	Q_{max}	T_D	T_R	T_C	T_{BA}	T_Σ	T_Σ	T_Σ	T_Σ	T_Σ	T_Σ	T_Σ
Alamo	0.12	1	27	332	20	380	910	750	578	198	1271	1654
Ellis Island	0.08	1	6	120	10	137	171	–	208	33	213	1191
Metropolis	0.31	1	12	108	13	134	244	142	60	161	294	1315
Montreal N.D.	0.10	1	11	472	25	509	1249	553	684	266	1110	2710
Notre Dame	0.08	1	25	298	93	417	1599	1047	552	247	2726	6154
NYC Library	0.19	1	6	173	13	193	468	200	213	154	453	3807
Piazza del Popolo	0.08	1	8	73	17	99	249	162	194	101	292	1287
Piccadilly	0.27	23	277	2405	588	3293	3483	–	1480	1246	3698	44369
Roman Forum	0.59	5	4	501	72	582	1457	–	491	1234	2004	4533
Tower of London	0.41	1	2	312	51	366	648	228	563	391	975	1900
Union Square	0.47	2	3	201	27	233	452	–	92	243	698	1244
Vienna Cathedral	0.12	2	110	270	40	422	3139	1467	582	607	3183	10276
Yorkminster	0.32	1	13	242	38	294	899	297	663	102	858	3225
Trafalgar	0.53	49	318	3850	631	4848	12240	–	2901	–	10210	29160
Gendarmenmarkt	0.41	2	3	161	30	196	–	–	214	–	799	–

scene points. For our method, based on the RANSAC technique, it is easier to find two real inliers for the camera center estimation. As a result, our method is more robust and the reconstruction result is more reasonable, which is similar to that of Bundler [38]. For the global SfM method LUD [31], the estimated scene structure is errornous, indicating that it is more sensitive to outliers. The reconstruction result of *Trafalgar* produced by our hybrid SfM is shown in Fig 6(b). We also perform our hybrid SfM method on the Quad [7] which has 348 ground-truth camera positions measured by the differential GPS (with an accuracy about 10cm), and we achieve similar median position accuracy with: DISCO [7] 1.16m, Bundler [38] 1.01m and ours 1.03m. Our reconstruction result on this dataset is shown in Fig. 1.

Based on the results of unordered image datasets, we can conclude that our method inherits the advantages of both incremental and global manners. More reconstruction results are shown in the supplementary material[2].

[2]http://vision.ia.ac.cn/Faculty/hncui/index.htm

7. Conclusion

In this paper, a new hybrid SfM method is proposed to increase the reconstruction efficiency, accuracy and robustness in a united framework. We propose a community-based rotation averaging method in a global manner first. Then, based on the estimated camera rotations, camera centers are estimated in an incremental manner. Our hybrid SfM method possesses both robustness advantage inheriting from incremental manner and efficiency advantage inheriting from global manner. Extensive experiments show that our method produces superior results in both reconstruction accuracy and computation efficiency compared to many of the state-of-the-art SfM methods.

Acknowledgement

This work was supported by NSFC under Grant 61333015, National Key Technology R&D Program under Grant 2016YFB0502002, and NSFC under Grants 61421004, 61632003. We thank Zhaopeng Cui for sharing the Campus dataset.

References

[1] Ceres solver. http://ceres-solver.org/.

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

[3] M. Arie-Nachimson, S. Z. Kovalsky, I. Kemelmacher-Shlizerman, A. Singer, and R. Basri. Global motion estimation from point matches. In *3DIMPVT*, pages 81–88. IEEE, 2012.

[4] R. Carceroni, A. Kumar, and K. Daniilidis. Structure from motion with known camera positions. In *CVPR*, volume 1, pages 477–484. IEEE, 2006.

[5] A. Chatterjee and V. M. Govindu. Efficient and robust large-scale rotation averaging. In *ICCV*, pages 521–528. IEEE, 2013.

[6] A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6 Pt 2):264–277, 2005.

[7] D. J. Crandall, A. Owens, N. Snavely, and D. P. Huttenlocher. SfM with MRFs: Discrete-continuous optimization for large-scale structure from motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(12):2841–2853, 2013.

[8] H. Cui, S. Shen, and Z. Hu. Global fusion of generalized camera model for efficient large-scale structure from motion. *Science China Information Sciences*, 60(3):038101, 2017.

[9] H. Cui, S. Shen, Z. Hu, et al. Efficient large-scale structure from motion by fusing auxiliary imaging information. *IEEE Transactions on Image Processing (TIP)*, 22:3561–3573, 2015.

[10] Z. Cui, N. Jiang, C. Tang, and P. Tan. Linear global translation estimation with feature tracks. In *BMVC*, 2015.

[11] Z. Cui and P. Tan. Global structure-from-motion by similarity averaging. In *ICCV*, pages 864–872. IEEE, 2015.

[12] Fortunato and Santo. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[13] T. Goldstein, P. Hand, C. Lee, V. Voroninski, and S. Soatto. Shapefit and shapekick for robust, scalable structure from motion. In *ECCV*, pages 289–304. Springer, 2016.

[14] V. M. Govindu. Lie-algebraic averaging for globally consistent motion estimation. In *CVPR*, pages 1–8. IEEE, 2004.

[15] V. M. Govindu. Robustness in motion averaging. In *ACCV*, pages 457–466. Springer, 2006.

[16] S. Haner and A. Heyden. Covariance propagation and next best view planning for 3d reconstruction. In *ECCV*, pages 545–556. Springer, 2012.

[17] R. Hartley, J. Trumpf, Y. Dai, and H. Li. Rotation averaging. *International Journal of Computer Vision (IJCV)*, 103:267–305, 2013.

[18] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[19] R. I. Hartley. Chirality. *International Journal of Computer Vision (IJCV)*, 26(1):41–61, 1998.

[20] M. Havlena, A. Torii, J. Knopp, and T. Pajdla. Randomized structure from motion based on atomic 3d models from camera triplets. In *CVPR*, pages 2874–2881. IEEE, 2009.

[21] M. Havlena, A. Torii, and T. Pajdla. Efficient structure from motion by graph optimization. In *ECCV*, pages 100–113. Springer, 2010.

[22] J. Heinly, J. L. Schonberger, E. Dunn, and J.-M. Frahm. Reconstructing the world* in six days*(as captured by the yahoo 100 million image dataset). In *CVPR*, pages 3287–3295. IEEE, 2015.

[23] N. Jiang, Z. Cui, and P. Tan. A global linear method for camera pose registration. In *ICCV*, pages 481–488. IEEE, 2013.

[24] F. V. K. Cornelis and L. V. Gool. Drift detection and removal for sequential structure from motion algorithms. In *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1249–1259. IEEE, 2004.

[25] F. Kahl and R. Hartley. Multiple-view geometry under the l-norm. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(9):1603–1617, 2008.

[26] L. Kneip, D. Scaramuzza, and R. Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *CVPR*, pages 2969–2976. IEEE, 2011.

[27] D. Martinec and T. Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *CVPR*, pages 1–8. IEEE, 2007.

[28] P. Moulon, P. Monasse, and R. Marlet. Adaptive structure from motion with a contrario model estimation. In *ACCV*, pages 257–270. Springer, 2013.

[29] P. Moulon, P. Monasse, and R. Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *ICCV*, pages 3248–3255. IEEE, 2013.

[30] D. Nistér. An efficient solution to the five-point relative pose problem. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(6):756–770, 2004.

[31] O. Ozyesil and A. Singer. Robust camera location estimation by convex programming. In *CVPR*, pages 2674–2683. IEEE, 2015.

[32] M. Pollefeys, D. Nister, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. J. Kim, P. Merrell, C. Salmi, S. N. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewenius, R. Yang, G. Welch, , and H. Towles. Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision (IJCV)*, 72(2):143–167, 2008.

[33] C. Rother. Linear multiview reconstruction of points, lines, planes and cameras using a reference plane. In *ICCV*, pages 1210–1217. IEEE, 2003.

[34] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113. IEEE, 2016.

[35] R. Shah, A. Deshpande, and P. Narayanan. Multistage sfm: Revisiting incremental structure from motion. In *3DV*, volume 1, pages 417–424. IEEE, 2014.

[36] T. Shen, S. Zhu, T. Fang, R. Zhang, and L. Quan. Graph-based consistent matching for structure-from-motion. In *ECCV*, pages 139–155. Springer, 2016.

[37] S. N. Sinha, D. Steedly, and R. Szeliski. A multi-stage linear approach to structure from motion. In *Trends and Topics in Computer Vision*, pages 267–281. Springer, 2012.

[38] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *Internation Journal of Computer Vision (IJCV)*, 80(2):189–210, 2008.

[39] C. Sweeney. Theia multiview geometry library: Tutorial & reference. `http://theia-sfm.org`.

[40] C. Sweeney, T. Sattler, T. Hollerer, M. Turk, and M. Pollefeys. Optimizing the viewing graph for structure-from-motion. In *ICCV*, pages 801–809. IEEE, 2015.

[41] R. Toldo, R. Gherardi, M. Farenzena, and A. Fusiello. Hierarchical structure-and-motion recovery from uncalibrated images. *Computer Vision and Image Understanding (CVIU)*, 140:127–143, 2015.

[42] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustment – a modern synthesis. In *Vision algorithms: theory and practice*, pages 298–372. Springer, 2000.

[43] K. Wilson, D. Bindel, and N. Snavely. When is rotations averaging hard ? In *ECCV*, pages 255–270. Springer, 2016.

[44] K. Wilson and N. Snavely. Robust global translations with 1dsfm. In *European Conference on Computer Vision (ECCV)*, pages 61–75. Springer, 2014.

[45] C. Wu. Towards linear-time incremental structure from motion. In *3DV*, pages 127–134. IEEE, 2013.

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

[47] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. In *CVPR*, pages 1426–1433. IEEE, 2010.

[48] E. Zheng and C. Wu. Structure from motion using structure-less resection. In *ICCV*, pages 2075–2083. IEEE, 2015.