Merge-SfM: Merging Partial Reconstructions

Meiling Fang[1]
meiling.fang@igd.fraunhofer.de
Thomas Pollok[2]
thomas.pollok@iosb.fraunhofer.de
Chengchao Qu[2]
chengchao.qu@iosb.fraunhofer.de

[1] Fraunhofer IGD
Fraunhoferstr. 5
64283 Darmstadt
Germany

[2] Fraunhofer IOSB
Fraunhoferstr. 1
76131 Karlsruhe
Germany

Abstract

Recovering a 3D scene from unordered photo collections is a long-studied topic in computer vision. Existing reconstruction pipelines, both incremental and global, have already achieved remarkable results. This paper addresses the problem of fusing multiple existing partial 3D reconstructions, in particular finding the overlapping regions and transformations (7 DOF) between partial reconstructions. Unlike the previous methods which have to take the entire epipolar geometry (EG) graph as the input and reconstruct the scene, we propose an approach that reuses the existing reconstructed 3D models as input and merges them by utilizing all the internal information to avoid repeated work. This approach is divided into two steps. The first is to find overlapping areas between partial reconstructions based on Fisher similarity lists. Then, based on those overlaps, pairwise rotation between partial reconstructions is estimated by solving an ℓ_1 approximation optimization problem. After global rotation estimation, translation and scale between each pair of partial reconstructions are computed simultaneously in a global manner. In order to find the optimal transformation path, the maximal spanning tree (MST) is constructed in the second stage. Our approach is evaluated on diverse challenging public datasets and compared to state-of-the-art Structure from Motion (SfM) methods. Experiments show that our merging approach achieves high computational efficiency while preserving similar reconstruction accuracy and robustness. In addition, our method has superior extensibility which can add partial 3D reconstructions gradually to extend an existing 3D scene.

1 Introduction

Structure from Motion (SfM) [1, 11, 15, 29] is used to estimate the interior and exterior camera orientation, and to recover the 3D scene structure at the same time. Recently, SfM has achieved impressive results from unordered image collections due to the improvement of the feature extraction and matching [20, 42], geometric verification [27, 33, 41], and bundle adjustment (BA) [10, 22, 40]. Traditional SfM pipelines treat the entire image set as the input, where image processing such as feature extraction and matching must be performed at first. Imagine that a large-scale 3D structure is already reconstructed, and a new set of images

*This work was done when Meiling Fang was an intern at Fraunhofer IOSB.

capturing a different area with small overlaps of the existing scene should be reconstructed to expand the current 3D scene. Moreover, it is often desirable for the merging process to be flexible, meaning that new partial 3D models need to be added to the existing reconstruction progressively. The traditional SfM pipeline cannot satisfy this requirement, since methods in the literature have to either restart the process from scratch, or add new images into the existing graph incrementally, and a lot of repetitive work has to be done compared to the direct fusion approach we propose in this paper. To estimate camera poses in a global manner, the epipolar geometry (EG) graph whose nodes represent images and edges link image pairs with consistent matching features is often leveraged. Obviously, the global estimation is sensitive to the structure of the EG graph and correspondence outliers. Others add new images into existing reconstructions incrementally and are prone to the order of the merging process. Besides, the repeated triangulation and local BA are time-consuming, and this decreases the scalability and efficiency of the reconstruction. In addition to traditional 3D reconstruction merging algorithms, which rely on 2D keypoints, some existing work [26, 34, 43] fuses 3D reconstructions by registering point clouds based on the learned 3D features from neural networks. For instance, [43] proposed a Fuseption-ResNet to learn multi-view local 3D descriptors from RGB-D frames. However, the capability of neural networks at the expense of losing camera poses is a trade-off that has to be made in this method. Similar to our approach, [14, 21] used diverse graph partition algorithms to generate clusters for partial reconstructions and then merged them together. While the partition based on the EG graph leads to a better clustering result, the entire EG graph must be built in advance, making the graph hard to expand. [11, 14] drastically ignored image-to-image connectivity and correspondences to build a skeletal geometry of iconic images [16] and registered other images to this skeletal model. Although those methods improve scalability of the system (*i.e.*, more suitable for large-scale image registration), loss of connectivity between most images leads to a decrease of the accuracy and consistency of the merged model. To solve the problems described above, we propose a new merging pipeline which preserves internal information of partial 3D reconstructions and builds reconstruction-to-reconstruction connectivity based on overlapping image detection to avoid repetitive triangulation or generation of the maximal spanning tree (MST). In this work, all partial reconstructions will be aligned into a unified world coordinate system using the pairwise similarity transformation $\mathcal{T}_{ab} \in Sim(3)$ between partial models $\texttt{Model}_a$ and $\texttt{Model}_b$ that contains rotation, translation, and scale. Compared to other hybrid SfM methods [3, 37] which roughly merge clusters of images, our approach reduces the impact of feature matching outliers in the translation stage. It also improves robustness with the help of the filtering process of inconsistent correspondences while estimating pairwise rotation between partial reconstructions. Moreover, a reconstruction graph is built with nodes corresponding to clusters of partial reconstructions and edges linking connected clusters, which is similar to an EG graph for matched image pairs. This means that the entire reconstruction graph can be extended by merging partial 3D models either in a batch mode or in an incremental manner instead of restarting the entire process again from scratch.

The main contributions of our approach, coined *Merge-SfM*, are summarized as follows:

- We introduce a highly extensible pipeline to handle the imbalance of accuracy, robustness, and efficiency in the merging problem. While the improved matching algorithm and the estimation of transformation in a global manner achieve high computation efficiency, the way of picking sufficient edges, and the cycle consistency constraint, retain more reliable connectivity.
- Our improved image matching algorithm, which builds connectivity between partial

reconstructions, accelerates the matching process by exploiting the internal information of the individual models. Then, pairwise rotations are estimated in a global manner by minimizing the global rotation error between pairwise models using all connectivities instead of solely the best-shot edge to enhance robustness to matching outliers.

- We propose to estimate the pairwise translation and scale parameters simultaneously via a linear optimization system composed of sufficient image-to-image connectivity once all partial reconstructions are under a unified world coordinate. This global method, accounting for the scale factors between partial reconstructions and the scales between feature correspondences at the same time, can efficiently solve the scale ambiguity problem.

2 Related Work

Detection of overlapping image pairs To find parts of the same scene among the images, feature descriptors such as [19] are usually used to compute the similarity between images. An effective, but prohibitively expensive approach is the brute force method, which tests all possible image pairs and features. In other words, the computational complexity $\mathcal{O}(N_I^2 N_F^2)$ is determined by the number of images and features. To accelerate the search for potential matches, several state-of-the-art SfM methods [6, 18, 23, 30] used image retrieval techniques assuming image pairs with high visual similarity scores are likely to match. The Vocabulary Tree [23], as an example, hierarchically quantizes descriptors from image keypoints. However, constructing a Vocabulary Tree is time-consuming and can lead to a large memory footprint. Cui *et al*. [6] demonstrated that Fisher vectors are much faster to compute and more indicative of possible matches than Vocabulary Trees. Inspired by [6], we also use Fisher vectors to reduce the dimensionality of image representation and advance the propagation of the *GraphMatch* algorithm. The output of this step is a set of potentially overlapping image pairs, which will be verified in a geometric manner.

Incremental SfM One straightforward way for camera pose estimation is to build a 3D model by initializing the structure from a few seed images, and then adding cameras one by one into a unified coordinate system. However, the incremental methods are sensitive to the initial seed model and the order of adding further images. In addition, the accumulated error may cause scene drift, especially for large-scale image collections. Furthermore, all incremental methods suffer from high time consumption because of frequent updating of local MSTs. To decrease the reconstruction error accumulated with iterations, COLMAP [29] adopted re-triangulation tracking while adding images. Moreover, some methods [13, 37] including COLMAP [29] proposed the *next-best-view* algorithm to handle the drawback of an incremental system, *i.e*., the dependence on the initial image pairs and the manner of model growing.

Global SfM The problem of global SfM can be represented as an EG graph. These approaches [2, 4, 8, 35] estimate all camera poses from the available relative poses at the same time and construct the MST only once. Therefore, compared with incremental and hierarchical SfM methods, the global systems can lower the computational burden. Moreover, global SfM methods can effectively avoid drifting errors and have potential in large-scale image collections. The typical global methods estimate all camera poses altogether in two steps: First, rotation averaging for camera rotations is computed and then all camera positions are determined in the next stage, which is referred to as translation averaging. Computation is also difficult because pairwise relative translation $\mathbf{t}_{ij}$ is only known up to a scale. Moreover,

translation estimation is more vulnerable to feature matching outliers.

Efficient SfM methods Recently, some hybrid SfM approaches employ exclusively clustering methods before the reconstruction pipeline. Cui *et al*. [5] computed camera rotations through a community-based rotation averaging method and estimated camera positions in an incremental way, which is time-consuming. [14] produces clusters of overlapping camera views, which means camera clusters share duplicate images. Both methods need access to the complete EG graph in advance which is not available in our pipeline. The work in [17] is based on a lightweight representation of the complete scene which is similar to ours. The progressive pipeline takes an ordered sequence of images with feature correspondences as its input and updates the resulting clusters and reconstructions with each image added to the scene, while we take partial reconstructions as the input and merge them into an extended model.

Cycle consistency In terms of correspondence outliers in the EG graph, cycle consistency is helpful to identify global rotations robustly by filtering false essential geometry. Two classes of methods based on spanning trees or cycles [12, 24] stand out. First, edges, which can build a cycle, are evaluated based on their relative rotation error $\mathbf{R}_b^\top \mathbf{R}_{ab}\mathbf{R}_a$, which represents the discrepancy between the relative and global rotation. Second, Enqvist *et al*. [9] performed rotation consistency $\mathbf{R}_{ab}\mathbf{R}_{bc}\mathbf{R}_{ac}$ in the MST which is extracted by a weighted EG graph with weights corresponding to the number of inlier correspondences. Inspired by those approaches, we also take into account cycle consistency to construct a simplified MST graph, where nodes represent reconstructions and edges are weighted based on a combination of the number of overlapping images and the rotation consistency to achieve the optimal rotation path. Experiments show that it performs excellently, *e.g*., on the *Gendarmenmarkt* [38] dataset with highly symmetrical architectures.

3 Merging Partial 3D Reconstructions

3.1 Overview

Our method takes multiple partial 3D reconstructions with internal information such as camera poses and reconstructed 3D points as the input and outputs a single merged SfM reconstruction (see Fig. 1). The core assumption is the existence of overlapping regions between the 3D models. The complete EG graph is gradually built by adding image-to-image connectivity between partial reconstructions. To detect the overlapping images efficiently, Fisher similarity lists [6] and an improved matching algorithm are used. Then the transformation $\mathcal{T}_{ab} \in Sim(3)$ between pairwise local 3D models $\texttt{Model}_a$ and $\texttt{Model}_b$ is estimated. The best pairwise rotation is created by minimizing the global rotation error. Then, a weighted combination of cycle consistency and the number of overlapping images are performed to pick a reference reconstruction and construct a simplified reconstruction MST graph. The scale factors are computed in terms of partial reconstructions instead of feature correspondences. Therefore, the translation and scale factors are obtained simultaneously by solving a linear optimization system composed of the remaining pairwise feature correspondences following known rotations of each camera, *i.e*., the output of the previous step. Generally, this method enables efficient fusion of multiple reconstructions because it takes full advantage of the input internal information and avoids repetitive computation of feature matching, triangulation and the MST.

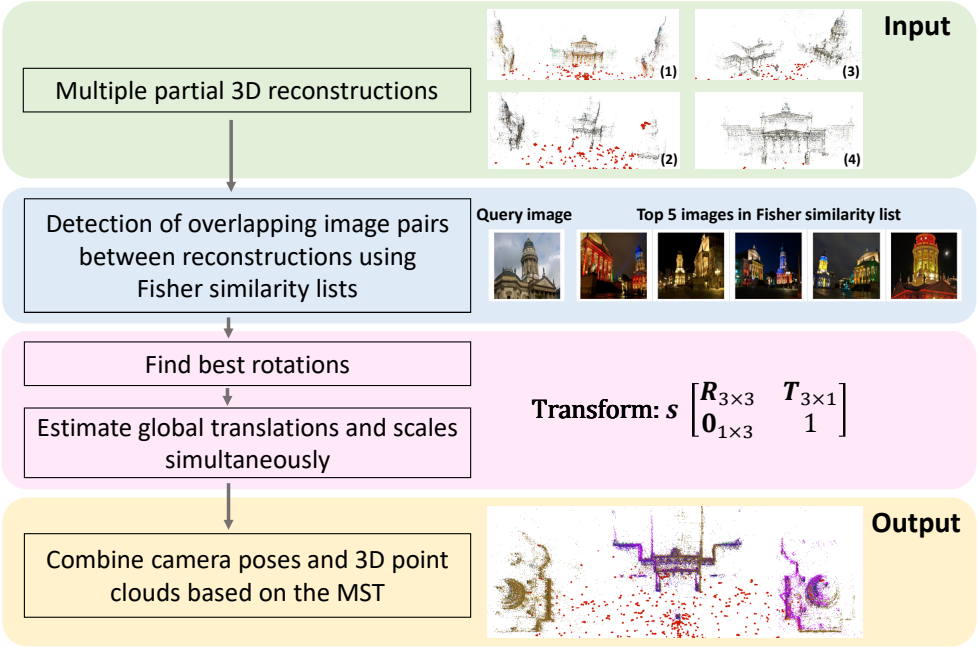

Figure 1: Overview of the proposed merging pipeline.

3.2 Image Matching Based on Fisher Similarity Lists

Overlapping images share similar visual content. The most frequently used approach to filter non-matching image pairs is the brute force method which is effective but computationally expensive testing all possible $\mathcal{O}(N^2)$ pairs of images. Yet, the vast amount of 75% to 95% of all image pairs do not match in most photo collections [39]. Finding high-quality matching image pairs efficiently and mitigating the cost of exhaustive matching approaches is a crucial problem. To tackle this challenge, [6] leverages two priors that can predict which image pairs are likely to match. The first is the Fisher similarity prior based on the distance between the Fisher vectors of any two images. The second prior is based on the graph distance between images in the underlying matching graph. To make use of the existing internal information, we modifiy the second prior, which is in our method ordered neighbor lists of every image according to the number of feature matches. The image matching process combines these two priors into an iterative sample-and-propagate scheme. An ordered neighbor list for every image depends on the number of feature matches between image pairs. In this case, image I_i is a neighbor of image I_j, *i.e.*, I_i and I_j have enough matching correspondences (correspondence threshold is set to 15 experimentally) after geometric verification that can be considered as a matched pair of images. Based on the neighbor lists, new suitable pairs of the image can be identified, and the search process is accelerated.

Once the Fisher similarity lists and neighbor lists are obtained, iterations between the sampling and propagation step start. The sampling stage is the same as [6], using Fisher similarity lists to guide the search for matching image pairs, while the propagation stage delivers new potential matching image pairs based on the neighbors on top of the listings, which is different to [6], and the neighbors of each image are sorted according to the number of feature matches based on the existing partial individual reconstruction in the propaga-

Figure 2: Fisher examples from the dataset *Gendarmenmarkt* [53]. The first column shows images in the query. The last five columns are the top 5 images in the order in Fisher similarity list from another reconstruction. Note that images in the top Fisher similarity list of this partial 3D reconstruction are all taken at night.

tion stage. The top ten percent of the images on the ordered neighbor list will be used for geometric verification.

Afterwards, the relative motions of image pairs are used to compute the transformation between pairwise partial reconstructions.

3.3 Estimation of Rotation

Each independent partial 3D reconstruction is in a different world coordinate system, so an alignment, which is a rotation matrix in $SO(3)$, should be performed to transform them into a unified coordinate system. Cui *et al*. [5] proposed a voting scheme to find the best rotation from many edges for each pair of the communities, *i.e*., to find the best only edge. They considered each edge between two communities as one possible rotation candidate. However, rotation alignment based on one edge is unreliable and error-prone. So we propose a method that takes full advantage of all overlapping images between pairwise partial reconstructions. This approach builds a linear equation system based on the global rotation error, which is solved by an ℓ_1 solver [33]:

$$\mathbf{R}_{ij} = \mathbf{R}_j^\top \mathbf{R}_i \tag{1}$$

$$\rho_e = \{\bar{\mathbf{R}}_{ij}\hat{\mathbf{R}}_{ij}^\top \mid I_i \in \texttt{Model}_a, I_j \in \texttt{Model}_b\} \tag{2}$$

Eq. (1) describes the discrepancy between the relative rotation and the global rotation and $\mathbf{R}_i$ is the global rotation of image I_i. Eq. (2) is an error measure between the rotations of the currently aligned reconstruction $\bar{\mathbf{R}}$ and the global rotations $\hat{\mathbf{R}}$ which is obtained from the feature-based essential matrix decomposition.

Suppose that I_j from partial 3D model $\texttt{Model}_b$ can be rotated by a pairwise rotation matrix $\mathbf{R}_{ab}$ to lie with I_i from $\texttt{Model}_a$ in a unified coordinate system (see Fig. 3). According to the above two equations, $\mathbf{R}_{ab}$, denoting the rotation from $\texttt{Model}_b$ to $\texttt{Model}_a$, is the only unknown parameter to be solved:

$$\mathbf{R}_j \mathbf{R}_{ab} \mathbf{R}_i^\top \hat{\mathbf{R}}_{ij}^\top = \mathbf{I}, \tag{3}$$

where $\mathbf{I}$ is the identity matrix. Eq. (3) can be rewritten as $\mathbf{R}_j\mathbf{R}_{ab} = \mathbf{R}_{ij}\mathbf{R}_i$. Then, all such linear equations from all m pairs of images between the pairwise partial reconstructions are collected into the following single linear equation system:

$$\mathbf{A}_{m\times 3}\mathbf{R}_{ab} = \mathbf{b}_{m\times 3} \tag{4}$$

$$\arg\min_{\mathbf{R}_{ab}} \|\mathbf{A}\mathbf{R}_{ab} - \mathbf{b}\|_1, \tag{5}$$

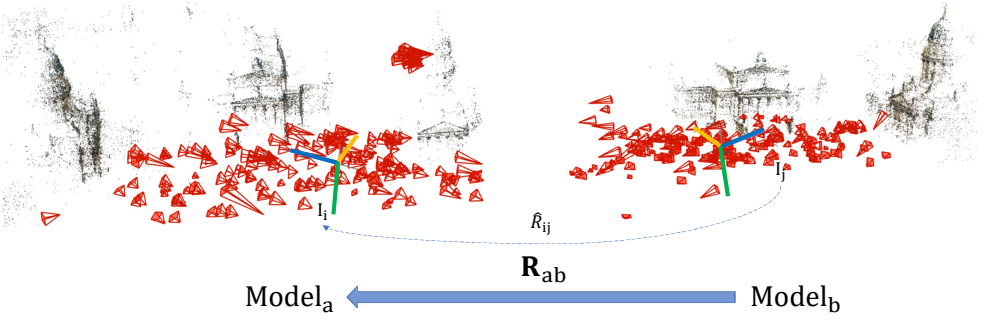

Figure 3: Since the input partial reconstructions $\texttt{Model}_a$ and $\texttt{Model}_b$ are in different coordinate systems, a rotation $\mathbf{R}_{ab}$ is needed to put them into a unified world coordinate, *i.e.*, all images from $\texttt{Model}_b$ will be rotated through $\mathbf{R}_{ab}$ with fixed $\texttt{Model}_a$ by minimizing the global rotation error. After that, partial reconstructions can be merged by translating and scaling.

where $\mathbf{A}_{m\times 3}$ and $\mathbf{b}_{m\times 3}$ are matrices stacking all $\mathbf{A}_n$ and $\mathbf{b}_n$ respectively with $\mathbf{A}_n = \mathbf{R}_j$ and $\mathbf{b}_n = \mathbf{R}_{ij}\mathbf{R}_i$ for the n-th pair of images I_i and I_j. Note that all the above rotation matrices are represented in quaternions for the optimization with orthogonal constraints.

Once all rotation transformations between the connected partial reconstructions are obtained, the EG graph is simplified as a weighted reconstruction graph with nodes representing partial reconstructions and edges linking reconstructions with sufficiently overlapping images, *i.e.*, the global rotation error is less than 10°. A larger s, which is the number of sufficiently overlapping images, stands for larger overlapping regions between partial reconstructions. In addition, cycle consistency is considered. The smaller the value c representing the rotation consistency (see Sec. 2), the higher the consistency of those relative rotations is. The weight w on each edge is a combination of the number of remaining overlapping images and the value of cycle consistency $w = ks + (1-k)(1-c)$ where both s and c are normalized. We construct the MST of the reconstruction graph and set the reference coordinate system as the node with the largest degree in case of multiple nodes with the same degree, then pick the node with the largest sum of weights. Finally, rotations of other partial reconstructions are aligned to the reference reconstruction based on the MST.

3.4 Simultaneous Estimation of Translation and Scale

Computing the translation and scale simultaneously is challenging. The first reason is that an essential matrix can only encode the direction of a relative translation without an absolute scale. The second reason is that the feature correspondences of image pairs between different partial reconstructions contain outliers. Even if geometric verification is performed with the random sample consensus (RANSAC) method, which is beneficial to exclude several erroneous feature matches, the translation between two SfM models can still be problematic because of incorrect epiplolar geometry. Because we filter some overlapping images with higher global rotation error in the previous stage, the second problem has been improved to a great extent. Thus, the only problem we encountered is the scale ambiguity.

Zhu *et al*. [44] presented a method that the scale factors regarding clusters and unknown camera positions of each camera are estimated by considering the translation averaging as a convex ℓ_1 problem. But this method has two drawbacks w.r.t. model fusion. First, the premise of this method is that to achieve higher robustness there must be duplicate images in

different clusters instead of depending on feature correspondence between images. Second, camera poses and the 3D point cloud cannot directly be transformed and merged, which still needs triangulation and BA later on. Different to [44], [45] focuses more on the camera-to-camera constraints, *i.e.*, the scale between cameras is considered. Instead, we propose an alternative approach to combine the scale between partial reconstructions and the scale between each pair of cameras.

With the global rotations $\mathbf{R}_j$ computed from the previous step fixed and the camera positions $\mathbf{c}_i$ from partial models, we estimate the pairwise translation $\mathbf{t}_{ab}$ and the scale factor α_{ab} between reconstructions $\texttt{Model}_a$ and $\texttt{Model}_b$ via a linear equation system:

$$\lambda_{ij}\mathbf{t}_{ij} = \mathbf{R}_j(\mathbf{c}_i - (\alpha_{ab}\mathbf{c}_j + \mathbf{t}_{ab})), \tag{6}$$

where $\mathbf{t}_{ij}$ is a relative translation between I_i from $\texttt{Model}_a$ and I_j from $\texttt{Model}_b$, $\mathbf{c}_i$ is the projected camera center of I_i, which can be computed as $-\mathbf{R}_i\mathbf{t}_i^\top$ (similar for $\mathbf{c}_j$), and λ_{ij} is the scale between the feature correspondences.

The parameters to be solved are $\mathbf{U} = \{\lambda_{ij}, \mathbf{t}_{ab}, \alpha_{ab}\}$. Eq. (6) can be rewritten as $\alpha_{ab}\mathbf{c}_j + \mathbf{t}_{ab} + \lambda_{ij}\mathbf{R}_j^\top\mathbf{t}_{ij} = \mathbf{c}_i$. Then the pairwise scale, translation, and all local feature scales between each pair of images can be represented altogether as a vector $\mathbf{x} = [\alpha_{ab}, t_x, t_y, t_z, \lambda_1, \ldots, \lambda_s]^\top$ where s corresponding to the number of remaining image pairs. At the same time, images in the pair from the reference reconstruction are represented as a vector $\mathbf{y}_c = [\mathbf{c}_1, \ldots, \mathbf{c}_s]^\top$. Therefore, the problem is formulated as a convex ℓ_1 problem:

$$\underbrace{[\mathbf{c}_j\mathbf{I}_{3\times 3}\ldots\mathbf{p}\ldots]}_{\mathbf{A}_{ij}}\mathbf{x} = \mathbf{y}_c, \tag{7}$$

where $\mathbf{A}_{ij}$ is a $3\times(4+n)$ matrix with that c_j is placed in the first and the identity matrix $\mathbf{I}_{3\times 3}$ in the second entry. Starting from the fourth column, the appropriate location $(4+k)$ is replaced by $\mathbf{p} = \mathbf{R}_j^\top\mathbf{t}_{ij}$ which corresponds to λ_k. Otherwise it is $\mathbf{0}_{3\times 1}$. Finally, the pairwise translation and scale are obtained with rapid convergence by stacking s image pairs into:

$$\arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}_c\|_1. \tag{8}$$

Once all pairwise rotations, translations and scales are obtained, transformations $\mathcal{T}_{3\times 4}$ with scale are performed based on the optimal MST path (see Sec. 3.3).

4 Experiments and Results

The experiments were conducted on a desktop PC with an Intel i7-6700K CPU with 32 GB RAM. Image collections are divided into subsets for partial reconstructions by an off-the-shelf community detection algorithm [32].

4.1 Evaluation

We demonstrate our pipeline on Internet datasets published in [38], consisting of twelve medium-scale datasets, one large-scale dataset *Piccadilly* and a challenging dataset *Gendarmenmarkt* with symmetric architectures. Image collections are divided into subsets for partial reconstructions by an off-the-shelf community detection algorithm [32].

Table 1: Performance comparison. N_m denotes the number of the geometrical verified image pairs. T_f denotes the time-cost of computing Fisher similarity lists and T_Σ represents the time-cost of complete matching in minutes.

	# verified matches			Time (min)			
Method	baseline	Voc. Tree	ours	baseline	Voc. Tree	ours	
Dataset	N_m	N_m	N_m	T_Σ	T_Σ	T_f	T_Σ
Alamo	38,400	10,374	34,892	5.31	10.02	0.41	4.99
Ellis Island	10,821	6,716	8,868	1.63	4.48	0.15	1.36
Metropolis	20,29	753	1,576	1.77	6.05	0.20	1.00
Montreal N.D.	8,009	5,130	7,220	4.04	10.58	0.27	3.90
NYC Library	6,756	4,365	5,272	2.07	6.38	0.17	1.63
Piazza del Popolo	9,228	5,539	7,711	1.82	5.06	0.18	1.58
Roman Forum	9,780	3,243	8,132	21.46	20.80	1.62	20.67
Tower of London	5,782	2,803	4,640	2.99	7.11	0.28	2.45
Union Square	9,288	4,988	7,779	13.89	16.28	0.61	11.37
Vienna Cathedral	37,154	13,661	33,497	13.15	15.79	0.87	6.17
Yorkminster	3,881	1,856	3.303	3.55	9.03	0.25	2.66
Gendarmenmarkt	17,363	6,819	15,770	50.08	31.20	0.69	39.18
Piccadilly	112,311	11,710	104,246	86.30	24.61	2.90	83.85

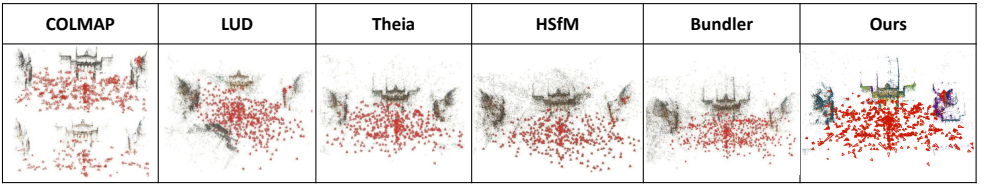

Figure 4: Qualitative reconstruction results on *Gendarmenmarkt*. Each color in our method represents a different partial 3D reconstruction in our result.

Table 1 shows results of the exhaustive method as the baseline, Vocabulary Tree and our matching method based on Fisher similarity lists. We measure the number of matched pairs of images and the runtime for matching between pairwise reconstructions. A publicly available pre-built Vocabulary Tree [28] is used. For fairness, all matching processes are based on COLMAP [29]. As shown in Table 1, our algorithm consistently finds more matched image pairs, which is comparable to the results of the baseline, and also more efficient than the Vocabulary Tree (approx. 2.5 times more edges and 2.5 times faster in most cases).

Furthermore, we compare our method with four state-of-the-art global SfM methods, two incremental and two hybrid SfM methods. Note that our input, *i.e.*, partial reconstructions from COLMAP [29], is different from these methods with the same input (EG graph and tracks) published in [38]. Qualitative comparison on *Gendarmenmarkt* is illustrated in Fig. 4. It is reported by [38] as a failure case because of its symmetric structures. As we can see, LUD (global) [25] and COLMAP (incremental) [29] generate erroneous scene structure. Although Theia (global) [35], HSfM (hybrid) [8] and ours achieve similar reconstruction results, similar to the model from Bundler (incremental) [31], our method has the largest number of registered images. In our result, the symmetrical structure is more complete even than Bundler [31]. Table 2 shows quantitative results. Our method preserves all the recovered images from partial reconstructions. Although we have different inputs and EG graphs as others, we achieve similar accuracy as state-of-the-art SfM methods, with significantly lower time-cost as reported in Table 3, which is up to 2 to 3 times faster than the state-of-the-art global and incremental methods in most cases. On *Metropolis*, our method performs 5.6 times faster than Theia (global). Here we have to point out that due to the different input in comparison with all other methods, those comparisons either in accuracy or in runtime are merely a reference to justify the reliability, efficiency and robustness of our novel merging pipeline. Finally yet importantly, our implementation is based on COLMAP [29], which can be replaced by a state-of-the-art SfM system for better performance.

Table 2: Accuracy Comparison on Internet image data. $\tilde{x}$ and $\bar{x}$ denote the median and mean position errors in meters respectively by taking the result of [31] as a reference. N_i is the number of cameras in the largest connected component of the input EG graph which is published in [33], and N_c is the number of recovered cameras. The bold font highlights the best result in each row.

Dataset		1DSfM [36]			LUD [25]			Cui [9]			Swe [24]			HSfM [9]			Theia [31]			Ours		
Name	N_i	N_c	$\tilde{x}$	$\bar{x}$	N_c	$\tilde{x}$	$\bar{x}$	N_c	$\tilde{x}$	$\bar{x}$	N_c	$\tilde{x}$	$\bar{x}$	N_c	$\tilde{x}$	$\bar{x}$	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	-	566	**0.3**	**1.5**	520	0.4	1.8	**582**	0.5	2.6
Ellis Island	247	214	**0.3**	3.0	–	–	–	233	0.7	4.2	203	0.5	-	**233**	2.0	4.8	210	1.7	**2.8**	232	0.8	4.4
Metropolis	394	291	0.5	7e1	288	1.5	4.0	317	3.1	16.6	272	**0.4**	-	**344**	1.0	3.4	301	1.0	**2.1**	328	1.6	5.3
Montreal N.D.	474	427	0.4	1.0	435	0.4	1.0	452	**0.3**	1.1	416	0.3	-	**461**	**0.3**	**0.6**	422	0.4	**0.6**	374	**0.3**	0.8
NYC Library	376	295	0.4	**1.0**	320	1.4	7.0	338	**0.3**	1.6	294	0.4	-	**344**	0.3	1.5	291	0.4	**1.0**	336	**0.3**	1.3
Piazza del Popolo	354	308	2.2	2e2	305	1.0	4.0	340	1.6	2.5	302	1.8	-	**344**	0.8	2.9	290	0.8	1.5	**344**	**0.5**	**1.1**
Roman Forum	1134	989	**0.2**	3.0	–	–	–	1077	2.5	10.1	966	0.7	-	1087	0.9	8.4	942	0.6	**2.6**	**1109**	0.8	6.4
Tower of London	508	414	1.0	4e1	425	3.3	10.0	465	1.0	12.5	409	0.9	-	**481**	**0.7**	6.4	439	1.0	**1.9**	469	**0.7**	5.4
Union Square	930	710	3.4	9e1	–	–	–	570	3.2	11.7	701	2.1	-	**827**	2.8	**3.4**	626	**1.9**	3.7	724	2.3	5.5
Vienna Cathedral	918	770	**0.4**	2e4	750	4.4	10.0	842	1.7	4.9	771	0.6	-	**849**	1.4	**3.3**	738	1.8	3.6	823	0.7	3.5
Yorkminster	458	401	**0.1**	5e2	404	1.3	4.0	417	0.6	14.2	409	0.3	-	421	1.2	**1.7**	370	1.2	1.8	**431**	1.2	4.2
Gendarmenmarkt	742	–	–	–	–	–	–	609	4.2	27.3	–	–	-	611	2.8	**26.3**	597	2.9	28.0	**704**	**2.4**	38.0
Piccadilly	2508	1956	0.7	7e2	–	–	–	2276	**0.4**	2.2	1928	1.0	-	**2279**	0.7	2.0	1824	0.6	**1.1**	2266	0.7	9.0

Table 3: Runtime comparison in seconds. N_c denotes the number of partial 3D reconstructions. T_m and T_t denote the time-cost of computing relative motions and estimating transformations including pairwise rotations, translations and scales, respectively. T_Σ is the total time-cost of corresponding SfM methods. Note that the input of [31] is a collection of images, which means the total time includes feature extraction and feature matching. The input of other methods is the same EG graph with feature correspondences and relative motions from [33].

Dataset		1DSfM [36]	LUD [25]	Cui [9]	Swe [24]	HSfM [9]	Theia [31]	Parallel SfM [40]	Bundler [30]	Ours		
Name	N_c	T_m	T_t	T_Σ	T_Σ	T_Σ	T_Σ	T_Σ	T_Σ	T_Σ	T_Σ	T_Σ
Alamo	4	910	750	578	198	380	497	264	1654	202	1.5	204
Ellis Island	3	171	-	208	33	137	28	45	1191	61	0.4	61
Metropolis	2	244	142	60	161	134	47	125	1315	5	0.1	5
Montreal N.D.	2	1249	553	684	266	509	163	261	2710	47	0.3	47
NYC Library	3	468	200	213	154	193	61	144	3807	29	0.4	29
Piazza del Popolo	3	249	162	194	101	99	61	93	1287	62	0.4	62
Roman Forum	4	1457	-	491	1234	582	244	902	4533	41	1.2	42
Tower of London	4	648	228	563	391	366	155	410	1900	22	0.6	23
Union Square	4	452	-	92	243	233	48	207	1244	34	0.9	35
Vienna Cathedral	4	3139	1467	582	607	422	244	905	10276	238	2.5	241
Yorkminster	2	899	297	663	102	294	92	281	3225	23	0.2	23
Gendarmenmarkt	4	-	-	214	-	196	72	-	-	82	1.4	83
Piccadilly	6	3483	-	1480	1246	3293	330	1614	44369	362	6.8	369

5 Conclusions

In this paper, we present a novel algorithm coined Merge-SfM that is able to dynamically expand an existing 3D reconstruction by adding new partial 3D reconstructions. We first identify overlapping regions between partial reconstructions based on the Fisher similarity lists, which is highly efficient. Then all models are aligned onto a unified coordinate system based on global rotation averaging. Afterwards, translations and scales are estimated simultaneously by solving an ℓ_1 optimization problem. Finally, partial reconstructions and camera poses are merged by global transformations $\mathcal{T}_{ab} \in Sim(3)$ based on an optimal MST. Our framework does not require any prior global knowledge about the image-to-image connectivity. Furthermore, our approach allows to incrementally extend a 3D reconstruction with no predefined end point. The proposed pipeline achieves superior computational efficiency while being on par with many state-of-the-art SfM methods in terms of reconstruction accuracy.

Acknowledgment

This work has received funding from the European Union's Horizon 2020 research and innovation program in the context of the VICTORIA project under grant agreement No. 740754.

References

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

[2] M. Arie-Nachimson, S. Z. Kovalsky, I. Kemelmacher-Shlizerman, A. Singer, and R. Basri. Global motion estimation from point matches. In *3DV*, 2012.

[3] B. Bhowmick, S. Patra, and A. Chatterjee. Divide and conquer: Efficient large-scale structure from motion using graph partitioning. In *ACCV*, 2014.

[4] L. Carlone, K. Daniilidis, R. Tron, and F. Dellaert. Initialization techniques for 3D SLAM: a survey on rotation estimation and its use in pose graph optimization. In *ICRA*, 2015.

[5] H. Cui, X. Gao, S. Shen, and Z. Hu. HSfM: Hybrid structure-from-motion. In *CVPR*, 2017.

[6] Q. Cui, V. Fragoso, C. Sweeney, and P. Sen. GraphMatch: Effcient large-scale graph construction for structure from motion. In *3DV*, 2017.

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

[8] E. Dunn and J.-M. Frahm. Next best view planning for active model improvement. In *BMVC*, 2009.

[9] O. Enqvist, F. Kahl, and C. Olsson. Non-sequential structure from motion. In *ICCV*, 2011.

[10] A. Eriksson, J. Bastian, T.-J. Chin, , and M. Isaksson. A consensus-based framework for distributed bundle adjustment. In *CVPR*, 2016.

[11] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building Rome on a cloudless day. In *ECCV*, 2010.

[12] V. M. Govindu. Robustness in motion averaging. In *ACCV*, 2006.

[13] S. Haner and A. Heyden. Covariance propagation and next best view planning for 3D reconstruction. In *ECCV*, 2012.

[14] M. Havlena, A. Torii, and T. Pajdla. Efficient structure from motion by graph optimization. In *ECCV*, 2010.

[15] J. Heinly, J. L. Schönberger, E. Dunn, and J. M. Frahm. Reconstructing the world* in six days. In *CVPR*, 2015.

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

[17] A. Locher, Havlena. M., and L. Van Gool. Progressive structure from motion. In *ECCV*, 2018.

[18] Y. Lou, N. Snavely, and J. Gehrke. MatchMiner: Effcient spanning structure mining in large image collection. In *ECCV*, 2012.

[19] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

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

[21] P. Moulon, P. Monasse, and R. Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *ICCV*, 2013.

[22] K. Ni, D. Steedly, and F. Dellaert. Out-of-core bundle adjustment for large-scale 3D reconstruction. In *ICCV*, 2007.

[23] D. Nister and H. Stewénius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006.

[24] C. Olsson and O. Enqvist. Stable structure from motion for unordered image collections. In *SCIA*, 2011.

[25] O. Özyeşil and A. Singer. Robust camera location estimation by convex programming. In *CVPR*, 2015.

[26] C. R. Qi, H. Su, M. Niessner, M. Dai, A. Yan, and M. Guibas. Volumetric and multi-view CNNs for object classification on 3D data. In *CVPR*, 2016.

[27] R. Raguram, J. Tighe, and J.-M. Frahm. Improved geometric verification for large scale landmark image collections. In *BMVC*, pages 77.1–77.11, 2012.

[28] J. L. Schönberger. COLMAP – SfM and MVS. https://demuc.de/colmap/, 2018.

[29] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *CVPR*, 2016.

[30] J. L. Schönberger, A. C. Bergeand, and J.-M. Frahm. PAIGE: Pairwise image geometry encoding for imporved effciency in structure from motion. In *CVPR*, 2015.

[31] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3D. *ACM TOG*, 2006.

[32] S. Sobolevsky, R. Campari, A. Belyi, and C. Ratti. General optimization technique for high-quality community detection in complex networks. *Phys. Rev. E*, 90:012811, 2014.

[33] H. Stewénius, S. H. Gunderson, and J. Pilet. Size matters: Exhaustive geometric verification for image retrieval. In *ECCV*, pages 674–687, 2012.

[34] H. Su, S. Maji, S. Kalogerakis, and E. Learned-Miller. Multi-view convolutional neural networks for 3D shape recognition. In *ICCV*, 2015.

[35] C. Sweeney. Theia multiview geometry library: Tutorial & reference. http://theia-sfm.org, 2016.

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

[37] C. Sweeney, V. Fragoso, T. Höllerer, and M. Turk. Large scale SFM with the distributed camera mode. In *3DV*, 2016.

[38] K. Wilson and N. Snavely. Robust global translation with 1DSfM. In *ECCV*, 2014.

[39] C. Wu. Towards linear-time incremental structure from motion. In *3DV*, 2013.

[40] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. Multicore bundle adjustment. In *CVPR*, 2011.

[41] Q. Xu, J. Li, W. Tao, and D. Ming. Efficient large-scale geometric verification for structure from motion. *Pattern Recogn. Lett.*, 125:166–173, 2019.

[42] L. Zhou, S. Zhu, T. Shen, J. Wang, T. Fang, and L. Quan. Progressive large scale-invariant image matching in scale space. In *ICCV*, pages 2381–2390, 2017.

[43] L. Zhou, S. Zhu, Z. Luo, T. Shen, R. Zhang, M. Zhen, T. Fang, and L. Quan. Learning and matching multi-view descriptors for registration of point clouds. In *ECCV*, 2018.

[44] S. Zhu, T. Shen, L. Zhou, R. Zhang, J. Wang, T. Fang, and L. Quan. Parallel structure from motion from local increment to global averaging. In *arXiv:1702.08601*, 2017.

[45] S. Zhu, R. Zhang, L. Zhou, T. Shen, T. Fang, P. Tan, and L. Quan. Very large-scale global SfM by distributed motion averaging. In *CVPR*, 2018.