Computer Vision and Image Understanding 157 (2017) 179–189

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Efficient tree-structured SfM by RANSAC generalized Procrustes analysis

Yisong Chen[a,c,*], Antoni B. Chan[b], Zhouchen Lin[a], Kenji Suzuki[c], Guoping Wang[a]

[a] School of Electronics Engineering and Computer Science, Peking University, Beijing, China
[b] Department of Computer Science, City University of Hong Kong, Hong Kong
[c] Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL, United States

ARTICLE INFO

Article history:
Received 29 November 2015
Revised 15 February 2017
Accepted 18 February 2017
Available online 28 February 2017

Keywords:
Structure-from-motion

ABSTRACT

This paper proposes a tree-structured structure-from-motion (SfM) method that recovers 3D scene structures and estimates camera poses from unordered image sets. Starting from atomic structures spanning the scene, we build well-connected structure groups, and propose RANSAC generalized Procrustes analysis (RGPA) to glue structures in the same group. The grouping-aligning operations hierarchically proceed until the full scene is reconstructed. Our work is the first attempt of using GPA for modern 3D reconstruction tasks. RGPA is able to merge multiple structures at a time and automatically identify outliers. The reconstruction tree is much more compact and balanced than previous hierarchical SfM methods and has a very shallow depth. These advantages, along with the resulting removal of intermediate bundle adjustments, lead to significantly improved computational efficiency over state-of-the-art SfM methods. The cameras and 3D scene can be robustly recovered in the presence of moderate noise. We verify the efficacy of our method on a variety of datasets, and demonstrate that our method is able to produce metric reconstructions efficiently and robustly.

1. Introduction

Structure-from-motion (SfM) techniques have become popular for reconstructing cameras and 3D scenes from unstructured and unconstrained image collections. The pioneering work of Brown and Lowe uses SIFT features and bundle adjustment to do metric reconstruction on a small set of weakly calibrated images (Brown and Lowe, 2005; Lowe, 2004; Triggs et al., 2000). The work is then improved to handle large-scale data under weaker constraints with more powerful computational tools (Snavely et al., 2006; Agarwal et al., 2009). Recently proposed SfM systems have enabled significant progress in vision and graphics applications (Pollefeys et al., 2004; Wu et al., 2011).

The dominant approaches for SfM are incremental or sequential algorithms (Snavely et al., 2006; Pollefeys et al., 2004), which start with a small seed reconstruction, then grow by repeatedly including additional cameras and scene points. Incremental methods tend to be computationally intensive, making repeated use of bundle adjustment as well as inconsistent measurements removal. This can be alleviated by employing multi-core optimization (Wu et al., 2011), or by splitting the problem into more tractable components (Snavely et al., 2008; Nister and Stewenius, 2006; Shah et al., 2014). Incremental methods also suffer from drift in scenes with weak visual connections.

One approach to SfM, which is less sensitive to drift, is to use a hierarchical reconstruction (Ni et al., 2012; Corsini et al., 2013). By organizing a hierarchical cluster tree and merging partial reconstructions along the tree, these methods are able to distribute errors evenly throughout the reconstruction, thus making them less sensitive to initialization and drift (Gherardi et al., 2010). This scheme can cut the computational complexity by one order of magnitude provided that the cluster tree is well balanced (Nister, 2000). In hierarchical methods, care must be taken to avoid bad structures caused by outlier matches (Nister, 2000; Havlena et al., 2009).

Another approach to SfM is to use global optimization (Martinec and Padjla, 2007; Sinha et al., 2012; Arie-Nachimson et al., 2012). First, camera rotations are estimated separately using two-view geometries. Then these rotations are fed into further optimization steps that solve for camera translations and structure (Jiang et al., 2013; Moulon et al., 2013). The global pose registration approach lacks built-in robustness to noise, and may fail to provide a good initialization (Hartley et al., 2013; Crandall et al., 2011). Although some attempts are made to handle noise (Chatterjee and Govindu, 2013; Enqvist et al., 2011), the global method tends to be

* Corresponding author.
E-mail address: chenyisong@pku.edu.cn (Y. Chen).

http://dx.doi.org/10.1016/j.cviu.2017.02.005
1077-3142/

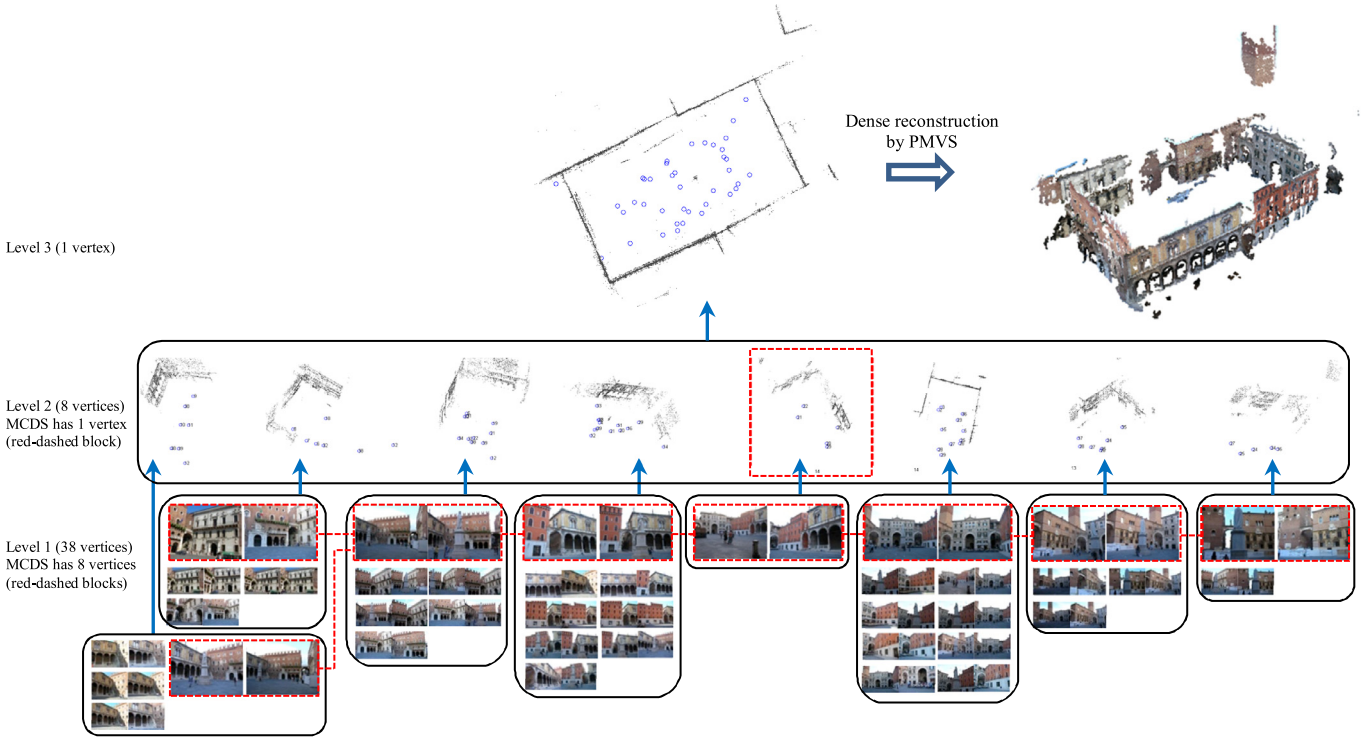


Fig. 1. Tree-structured SfM algorithm. The algorithm starts with a pairwise reconstruction set spanning the scene (represented as image-pairs in the leaves of the reconstruction tree). From these atomic structures a minimum connected dominating set (MCDS) is computed (MCDS vertices are image pairs with red-dashed boxes, and connections are red-dashed lines). Each MCDS vertex induces a group, which contains the vertex itself and all non-MCDS vertices connected to it (black rounded-rectangle), as well as the adjacent MCDS vertices (connected by red-dashed line). All structures in the same group are aligned and merged to form the higher-level structures (blue arrows) by RANSAC generalized Procrustes analysis (RGPA). The MCDS and RGPA operations are repeated hierarchically until the complete scene is built at the top-level of the tree. The very small tree depth is a key advantage of our method. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

unstable and inaccurate when applied to highly unstructured data (Wilson and Snavely, 2014).

Robustness to noise and computational efficiency are two major challenges to modern SfM approaches. On one hand, adverse camera poses and mismatches can cause bad atomic reconstructions, and may significantly bias nonlinear optimization even on small datasets. Many studies have been conducted to improve the stability (Zach et al., 2010; Govindu, 2006; Olsson et al., 2011; Heinly et al., 2014). On the other hand, great efforts have been devoted to speed up key SfM modules, like image clustering, feature matching and nonlinear optimization (Wu, 2013, Bhowmick et al., 2014; Frahm et al., 2010).

The attempt on full reconstruction from all available matches is often undermined by highly unstructured dataset, and may take unnecessary computational cost. Instead, a spanning structure has been proved sufficient for many tasks (Snavely et al., 2008; Lou et al., 2012; Havlena et al., 2010). In this paper, we propose a novel SfM method to build a spanning reconstruction from unordered images. Our method addresses both the robustness and efficiency challenges, and is able to recover cameras and scene structures quickly from highly unstructured datasets. We first compute a spanning tree that traverses all images and generate the associated atomic structures. We then group the structures into local clusters with close intra-similarities, and build larger reconstruction by robustly aligning multiple structures in the same group using a RANSAC extension of generalized Procrustes analysis (GPA). This grouping-aligning process iteratively proceeds until a full reconstruction is achieved. To the best of our knowledge, our work is the first attempt of practical usage of GPA for modern reconstruction tasks. Fig. 1 gives an example of our tree-structured SfM method.

Our work has two key contributions:

1. We propose a RANSAC generalized Procrustes analysis method for multiple structure alignment, which is fast and robust to outliers (Section 2.3).
2. We design a shallow reconstruction tree for organizing unordered images and grouping local structures, which enables quick and reliable 3D reconstruction (Sections 2.2 and 2.3.2).

Intrinsically our approach partitions the problem into smaller instances and combines them hierarchically. Therefore, it has the advantages of a hierarchical solution, such as high computational efficiency and insensitivity to initialization and drift. Our method is advantageous over state-of-the-art SfM methods in the following aspects:

1) In comparison to hierarchical methods (Gherardi et al., 2010), our method has much shallower tree depth and thus faster speed, which is due to merging of more than two structures simultaneously;
2) In comparison to global optimization methods (Havlena et al., 2009; Martinec and Padjla, 2007; Sinha et al., 2012; Arie-Nachimson et al., 2012), our method is more tolerant of large camera rotations and avoids the problem of registration failure;
3) In comparison to incremental methods (Snavely et al., 2006; Agarwal et al., 2009; Pollefeys et al., 2004), our divide-and-conquer strategy is much faster, less sensitive to drift, and easier to parallelize.

2. Tree-structured SfM based on RANSAC generalized Procrustes analysis

In this section, we propose a tree-structured SfM algorithm based on RANSAC generalized Procrustes analysis (RGPA), which can quickly and robustly recover the cameras and the 3D scene in the presence of moderate noise. First, we build point tracks and perform pairwise reconstruction. Next, we build a spanning structure set containing two-view reconstructions that spans the maximum connected component of the image set. The spanning structures are then grouped and merged hierarchically in a bottom-up manner. RGPA is used to align and merge a group of structures, while also detecting and removing outlier matches. After all structures are merged, we perform a final bundle adjustment to refine the reconstructed 3D points and cameras.

2.1. Track generation and pairwise reconstruction

In this preprocessing step, tracks are progressively established from SIFT features using an algorithm similar to the one described in Olsson and Enqvist (2011). For all image pairs with sufficient matching features in the track list, we use the 5-point algorithm (Nister, 2004) to estimate camera poses and triangulate points. For the calibration matrix, we use the standard camera model with the principal point (x_0,y_0) being at the center of the image and the focal length f being extracted from the cameras' EXIF data (Snavely et al., 2006).

One important operation at this stage is to detect and remove bad image pairs with incorrect epipolar geometry, which may destroy future structure merging. We carry out several checks to remove suspicious pairs (Pollefeys et al., 2004; Wu et al., 2011; Snavely et al., 2008): 1) we remove pairs with fewer than 15 matches; 2) we remove pairs with too similar scenes (the outlier ratio $r < 0.01$ for a planar homography test); 3) we remove pairs where the distance d_1 between the camera centers is too small compared to the median distance d_2 between the cameras and the reconstructed points ($d_2 > 10d_1$); 4) We check rotation-consistency for all triplets of 2-view reconstructions and remove each pair causing inconsistency over 50% of the triplets it belongs to.

2.2. Spanning structure building

We aim to build a spanning reconstruction from the most reliable two-view reconstructions. To do this we first build a graph with each image being a vertex and each valid pairwise reconstruction being an edge. We propose a method to extract a spanning structure set ψ from the maximum connected component H of this initial graph. ψ contains all images in H and all edges in ψ corresponds to the most reliable two-view reconstructions. The advantage of using a spanning structure is twofold: 1) a spanning structure generally covers the main body of the scene with a sufficient number of features, which is computationally more efficient than a redundant full reconstruction; 2) the atomic reconstructions in the spanning structure are the most reliable ones and thus further reduce the risk of failure caused by bad epipolar geometries.

Previous work uses maximum leaf spanning tree to define the spanning structure (Snavely et al., 2008). This works well for incremental SfM but may result in relations between initial structures that are too weak for hierarchical SfM. Here, we propose a method for building a spanning structure for robust structure merging, which computes the spanning tree of a graph by heuristically collecting 2-view reconstructions with more valid matches.

Assume that $v_{i,j}$ is the number of valid matches for image pair (i, j). After most bad image pairs are removed in the preprocessing step in Section 2.1, it is reasonable to assume that a larger $v_{i,j}$ corresponds to a better reconstruction (Enqvist et al., 2011). To justify this, we separated the 162 2-view reconstructions of the "Dante" dataset into good and bad ones with a threshold t of the average re-projection error $e_{i,j}$, and conducted a logistic regression to estimate the goodness of reconstruction as a function of $v_{i,j}$. The result in Fig. 2 verified that a larger $v_{i,j}$ corresponds to a better reconstruction. Moreover, local structures sharing more matches generally have stronger connections and lead to better merging results. Therefore, we use $v_{i,j}$ to define the edge weights, and look for the spanning tree with the largest total weights (Prim, 1957). Fig. 3 plots an example maximum spanning tree for 6 images.

After the optimal spanning tree ψ is computed, the two-view reconstructions associated with the edges of ψ make the atomic structures for the subsequent merging step, and act as the leaves of the reconstruction tree. The 38 leaf vertices of Fig. 1 give the atomic structure set for the dataset Dante in our experiment.

The above scheme helps us quickly build a spanning tree with strong enough connections among its nodes. Taking into account the distribution of matches or other semantic information may lead to better models at the cost of more advanced techniques. This is an interesting future work.

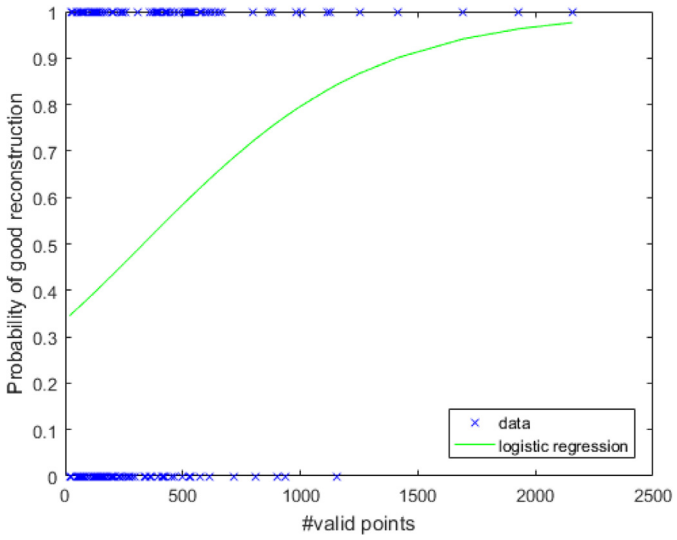


Fig. 2. Logistic regression for the 162 2-view reconstructions of "Dante". The threshold t of the average re-projection error for separating good and bad reconstructions is set to the median value of all $e_{i,j}$. This experiment justifies the assumption that a larger $v_{i,j}$ corresponds to a better reconstruction.

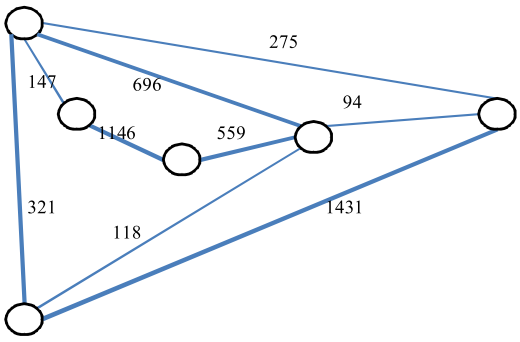


Fig. 3. An example spanning structure from 6 images by solving the maximum spanning tree. Each edge connects two images with sufficient valid matches. The number of valid matches between two images (i,j) $v_{i,j}$ are used to define the edge weights. The bold edges represent the maximum spanning tree, which defines the spanning structure.

2.3. Bottom-up structure merging by RGPA

Each leaf of the reconstruction tree corresponds to a two-view reconstruction and provides a set of reconstructed 3D points. Globally any two of these 3D point sets are related by a 7-degree-of-freedom similarity transform (3 for rotation, 3 for translation and 1 scale). Therefore, these initial structures can be aligned using their common points by a carefully designed 3D registration algorithm (Eggert et al., 1997). We use a modified version of generalized Procrustes analysis (GPA) (Crosilla and Beinat, 2002), a statistical shape analysis tool, for the registration task.

2.3.1. Generalized Procrustes analysis

We first give a brief problem statement for generalized Procrustes analysis (Pizarro and Bartoli, 2011) for shape alignment. The input shapes are represented by n matrices $D_1, \ldots, D_n$. Each shape $D_i \in R^{d \times m}$ is composed of m d-dimensional points,

$$D_i = (D_{i,1}, \ldots, D_{i,m}), D_{i,j} \in R^d. \quad (1)$$

The shape alignment problem is to find the set of n similarity transformations T_i: $R^d \rightarrow R^d$ and the reference shape $F = (F_1, \ldots, F_m) \in R^{d \times m}$ that minimize the cost function

$$\varepsilon(T, F) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_{i,j} \left\| F_j - T_i\left(D_{i,j}\right) \right\|_2^2, \quad (2)$$

where $\mu_{i,j} \in \{0, 1\}$ allows the model to ignore missing data when not all points can be observed in all shapes. The GPA problem can be solved by alternating between the estimation of T and F. In particular, Procrustes analysis (PA) is used to align all shapes one by one with the current reference F to solve T, and all aligned shapes are superimposed to update F. The algorithm terminates when the change of the reference is sufficiently small.

The standard GPA algorithm goes as follows:

1. Choose a reference shape (typically by selecting it among the available instances).
2. Superimpose all instances to current reference shape by PA using their sharing points.
3. Compute the mean shape of the current set of superimposed shapes.
4. Go back to step 2 if the mean shape and the reference is close enough, otherwise, return the reference and terminate the algorithm.

In our context, n is the number of partial reconstructions in a group, and m is the total number of reconstructed 3D points for these n reconstructions. We consider non-reflected similarities in 3D space ($d = 3$). Our GPA model is different from the model in (2) in that the factor variable $\mu_{i,j}$ is not fixed, but dynamically changes based on the quality of alignment of on-the-fly updated matches. In an iterative alignment and reference update module, reliable points are progressively updated until convergence, while spurious points are detected and removed by cross-checking over multiple shapes. This makes our model much less sensitive to noise. We provide details for structure grouping and merging in the remainder of this section.

2.3.2. Structure set grouping

Given a set of partially reconstructed structures S_i, $i = 1,\ldots,s$, composed of the reconstructed 3D points set P_i, and the cameras set C_i, we first partition the structures into g groups G_i, $i = 1,\ldots,g$ (not necessarily mutually exclusive), each containing a cluster of structures similar to each other. We have three main requirements for this partitioning:

1) For the sake of reliable merging of local structures, all structures in a group should have enough matches to a reference shape, so that GPA can be smoothly carried out.
2) For the sake of complete reconstruction, any two groups should be connected via a path with strong enough connections between adjacent groups in the path.
3) For the sake of computational efficiency, the total number of groups should be as few as possible.

We model the structure set as a graph with each structure acting as a vertex and the number of common points defines the weight connecting two vertices. With this graph the above requirements can be well described as a minimum connected dominating set (MCDS) problem (Havlena et al., 2010). The MCDS model has been used before for skeletal set building (Snavely et al., 2008; Havlena et al., 2010). In our context we use MCDS to help do structure grouping.

A minimum connected dominating set of a graph Γ is a set Δ of vertices with three properties:

1) Every vertex in Γ either belongs to Δ or is adjacent to a vertex in Δ.
2) Any vertex in Δ can reach any other vertex in Δ by a path that stays entirely within Δ.
3) Δ has the smallest possible cardinality among all sets of Γ that satisfy 1) and 2).

We can see that the 3 properties of MCDS relate perfectly to the 3 requirements for structure grouping. For grouping structures at the same level of the reconstruction tree, we employ a variant of the greedy algorithm proposed in Guha and Khuller (1998) to compute an approximation of MCDS. The algorithm is illustrated in Fig. 4. In particular, during greedy searching we give higher priority to those vertices sharing more common points with their neighbors. Our algorithm proceeds as follows:

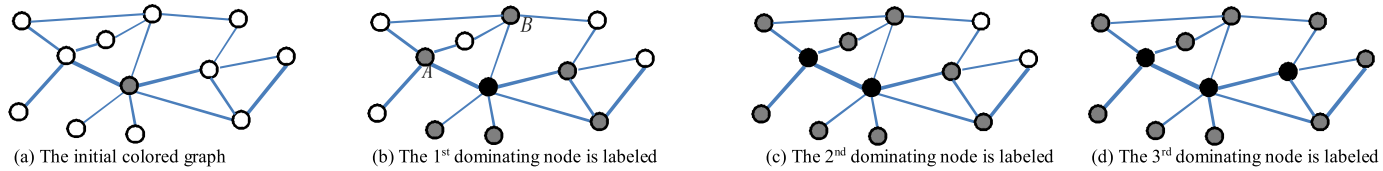
1. Initialization: Select the vertex with the maximum degree and color it gray. Color all other vertex white.
2. Select a gray vertex v with 1) maximum degree, and 2) strongest connection, to white vertices, and color it black.
3. Color every white neighbors of v gray.
4. Go back to step 2, until all vertices are black or gray.

When the algorithm ends, the black vertices make the minimum connected dominating set. Each MCDS vertex acts as the reference of a group, which is composed of this vertex and all vertices connected to it. The dashed boxes connected by dashed lines at the bottom level of Fig. 1 give an example of MCDS.

2.3.3. RANSAC generalized Procrustes analysis

We now present our RANSAC GPA (RGPA) algorithm for merging all structures in a group. We first choose the MCDS vertex P_r as the initial reference shape. Next we iterate between an alignment step and a reference update step. In the alignment step, each structure in the group G is aligned with the reference P_r using valid matches determined by a RANSAC (Fischler and Bolles, 1981) extension of Procrustes analysis (PA). Particularly, two structures are aligned by conducting multiple PA trials with a small sample of points (5 points), and the result with the most inliers is selected. For the RANSAC strategy we choose LMedS (Choi et al., 2009), which is fast and involves no parameter tuning for less than 50% outliers. For each local alignment 60 RANSAC trials are tested, which guarantee a success probability of over 99% under a conservative estimate of 40% outliers.

For each PA operation, the factor variable $\mu_{i,j}$ is activated only for inliers detected by RANSAC. This means that the 3D points used in every PA are not the same, but dynamically change depending on the inlier matches detected by RANSAC. In other words, the registration and optimization of (2) is done in terms of all inliers, which undergo on-the-fly updating during iteration. This allows our model to effectively detect and remove outliers under noisy



(a) The initial colored graph (b) The 1st dominating node is labeled (c) The 2nd dominating node is labeled (d) The 3rd dominating node is labeled

Fig. 4. A step by step illustration of our MCDS computing algorithm. The initial graph in (a) contains 12 structures. Bolder edges correspond to more common points between two structures. In (b), for labeling the 2nd black vertex, vertex *A* has higher priority than vertex *B* because it has stronger edges to adjacent white vertices, although the number of its white neighbors is equal to B (both 3). Therefore, vertex *A* is colored black in (c). The 3 black vertices in (d) make the final MCDS.

(P_1) (P_2) (P_3)

Fig. 5. An example of reference updating in a 3-structure group. The 3 structures P_1, P_2 and P_3 are represented by 3 images from "Dante". P_2 is the initial reference P_r. The common features C_{ij} between P_i and P_j are plotted as red (C_{12}), green (C_{23}) and blue (C_{13}) markers, respectively. When aligning P_1 and P_2, the common features C_{12} (mainly in the red ellipse) act as tie points, and likewise for P_3 and P_2 (C_{23}, green dashed ellipse). At the end of each GPA iteration, the reference shape P_r is updated as the set of features with at least two occurrences (i.e., all features in C_{12}, C_{23} and C_{13}). The features occurring only in one structure are never selected as tie points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Algorithm 1. RANSAC generalized Procrustes analysis for one group *G*.

Input: Group $G=\{S_i(P_i,C_i),S_{i+1}(P_{i+1},C_{i+1}),\ldots,S_j(P_j,C_j)\}$,
Reference index $r \in \{i,\ldots,j\}$.
Initialization: Set reference shape $P_r^0 = P_r$
Set iteration times $numIter = 3$
for $l = 1 \ldots numIter$
for $k = i \ldots j$
Align P_k to P_r^{l-1} using RANSAC Procrustes analysis to obtain the similarity transform T_k^l and the aligned points set P_k^l:
$[P_k^l\ T_k^l] = \text{RansacProcrustesAnalysis}(P_k, P_r^{l-1})$; %only inliers detected by RANSAC are used for alignment.
end
Superimpose all aligned shapes to build an intermediate set of points:
$P = \text{merge}(P_i^l, \ldots, P_j^l)$;
Update the reference shape P_r by collecting valid points in P:
$P_r^l = \text{collectValidpoints}(P)$; %only points in at least two structures are used for reference updating.
end
Align all structures in the group into a new structure:
$S = \text{align}(P_i^l, \ldots, P_j^l, T_i^l, \ldots, T_j^l)$;
Output: merged structure *S*.

environments. Note that multiple shapes can be aligned at a time with the reference shape in one alignment step, which leads to very small tree depth and significant time saving. This is an important advantage of RGPA.

In the reference update step, a new reference shape is computed by superimposing all aligned shapes, and all points occurring in at least two structures are averaged and accepted as new reference points. This step can add some points that are not originally in P_r, but in at least two other structures, into the reference. Note that only points in at least two structures are used for shape alignment and reference updating to ensure an unbiased estimate (see Fig. 5 for an illustration). The iteration proceeds for 3 times, which is long enough for the reference shape to converge under varied noise levels (see experiments). We denote our algorithm for aligning and merging all structures in a group as RANSAC GPA (RGPA). The RGPA module is summarized in Algorithm 1. Under noisy environment RGPA is more efficient than the early robust GPA attempt (Crosilla and Beinat, 2006), which has to iteratively process increasingly bigger data. In comparison to other SfM works that conduct robust estimation of the 3D similarity transform to merge two 3D shape (Gherardi et al., 2010; Li et al., 2008), RGPA is able to merge multiple shapes at the same time, and brings the benefit of a shallow reconstruction tree.

One distinctive property of our RGPA module is that it is resistant to outliers randomly distributed in the original shapes, thus it is able to work robustly under a relatively noisy environment. This benefit is mainly from the dynamic reference update operation between different iteration rounds. During the reference update, inliers tend to be attracted towards a fixed pivot point, which improves the reliability of the reference for future alignment. On the contrary, outliers generally have inconsistent positions in 3D and fail to converge to a steady point, and thus have little chance of surviving repetitive outlier checking. The dynamic change of the reference shape is another important advantage of RGPA, which equips the conventional GPA framework with a desirable property of counteracting noise. Particularly, in the context of our SfM al-

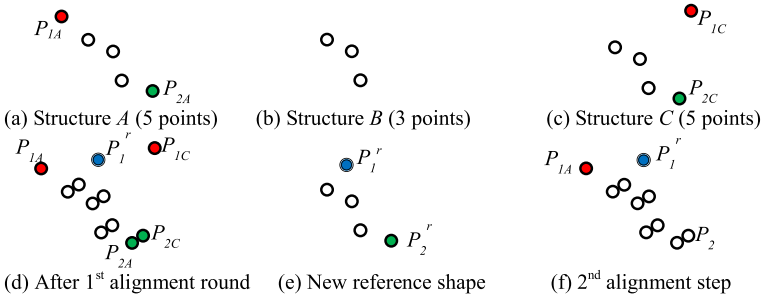


Fig. 6. Illustration of reference updating and outlier detection by RGPA. Three structures *A, B* and *C* are merged by RGPA, with *B* acting as the reference. *A* and *C* have two common track points P_1 (red points) and P_2 (green points) invisible to *B*. P_1 is an outlier whereas P_2 is an inlier. After the 1st alignment round both P_1 and P_2 are added to the new reference shape because they occur in at least two structures (*A* & *C*). However, we see from (d) that P_{1A} and P_{1C} are not well aligned as P_{2A} and P_{2C}, and the corresponding new averaged reference point $P_1{}^r$ (blue point) in (e) is distant from both P_{1A} and P_{1C}. As a result, in the 2nd alignment round for structure *A*, P_1 has big alignment error and is labeled as an outlier, while P_2 successfully survives as a new reference point (f). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 7. Some erroneous track matches detected and removed by RGPA due to their large alignment errors. These mismatches may cause future alignment or BA failure if not detected and removed at an early stage.

gorithm, the above advantage makes RGPA capable of effectively detecting and removing erroneous tracks by cross-checking among multiple structures in one group. We use the median alignment error to estimate the noise variance σ^2 as in Raguram et al. (2011), and employ the 3σ criterion to identify outliers. Tracks containing outliers are removed from the merged structure. The reference updating and outlier labeling mechanism is illustrated in Fig. 6. The outlier in the figure is difficult to detect by popular error checking methods based on 3D-2D re-projection, but can be easily detected by our 3D-3D alignment method in iterative RGPA. With this scheme mismatches as shown in Fig. 7 can be effectively identified and removed.

Algorithm 1 is applied to each group $G_g{}^l$ at level-l and results in a new aligned structure S at a higher level $l+1$. Note that there might be some duplicate cameras between structures in a group (e.g., structure 1 contains cameras 1,2 and structure 2 contains cameras 2,3, then camera 2 is duplicate). In this case, after alignment we retain the camera associated with more 3D points and discard the others, to ensure the uniqueness of the cameras in one structure.

2.3.4. Higher level grouping and merging

An optional operation is to do sparse bundle adjustment (BA) (Wu et al., 2011; Lourakis and Argyros, 2009) for each newly generated structure before proceeding to grouping and merging at the next higher level of the reconstruction tree. Fortunately, due to the built-in outlier detecting mechanism of RGPA, the structures generally have good accuracy and this optional BA operation becomes unnecessary. Therefore, the structure grouping and merging operations are performed bottom-up without any intermediate BAs. After the top level is reached, we execute a final BA to refine the structure. The removal of intermediate BAs leads to big saving of running time.

Fig. 1 plots the tree structure of our SfM algorithm for the "Dante" dataset. The reconstruction is accomplished in as few as 3 levels. The very small tree depth leads to fast reconstruction. It also helps reduce the risk of error propagation. It is worth noting from Fig. 1 and Algorithm 1 that all atomic RGPA operations in the same level of the reconstruction tree, as well as all atomic Procrustes analyses in the same iteration round of a group, are independent of each other. This means that RGPA can be parallelized to further reduce the running time.

3. Experiments

We implement our RGPA-based SfM algorithm in Matlab, with some core components (SIFT feature extraction, 5-point algorithm, BA) obtained as pre-compiled C code (Lowe, 2004; Wu et al., 2011; Nister, 2004). Experiments except preprocessing were run on a desktop PC (3.4 GHz dual core, 16 GB RAM).

We first test the performance of RANSAC in RGPA under noisy environment and experimentally show the robustness of RGPA. Considering the fact that the choice of the threshold parameter for the standard RANSAC is not trivial in large-scale SfM tasks (Moulon et al., 2012), we take the LMedS RANSAC model (Choi et al., 2009), which needs no parameter tuning at a moderate noise rate. We test on EPFL Fountain-P11 dataset (Strecha et al., 2008), which contains 11 images that can be grouped with one RGPA merging operation. We synthetically add various ratios (e) of outlier noise. Fig. 8 plots the estimated inlier ratios as RGPA iterates under different amounts of outlier noise. The inlier ratio quickly increases because outliers are effectively detected and discarded. Under all noise levels the estimated inlier ratios converge and become very close to 1.0 after the 3rd iteration. This implies that 3 iterations are sufficient for RGPA.

In Table 1 we compare the performances of LMedS, and using no RANSAC in RGPA. It is clear that without RANSAC reliable recon-

Table 1
Comparison of RANSAC strategies before the final BA.

Outlier ratio (e)		Origin	0.1	0.2	0.3	0.4
Time (s)	LMedS	1.27	1.24	1.21	1.20	1.19
	noRANSAC	0.42	0.42	0.42	0.42	0.42
Number of recovered points	LMedS	11,114	10,585	10,166	9698	9715
	noRANSAC	11,116	11,116	11,054	11,120	11,120
Average reproj. error (pixel)	LMedS	0.74	0.77	0.81	0.82	0.85
	noRANSAC	1.14	11.51	11.92	10.03	11.26

Table 2
Performance evaluation on 3 small datasets with ground-truth cameras (R_{err} in degree and T_{err} in mm. Bold indicates the best score).

Dataset\Method	Fountain-P11		Herz-Jesu-P25		Castle-P30	
	R_{err}	T_{err}	R_{err}	T_{err}	R_{err}	T_{err}
Global1 (Arie-Nachimson et al., 2012)	0.421	23.0	0.313	48.0	–	–
Global2 (Jiang et al., 2013)	0.195	14.0	0.188	64.0	0.480	235.0
VSFM (Wu et al., 2011)	0.041	7.6	**0.068**	25.0	0.156	175.8
HIER (Gherardi et al., 2010)	**0.035**	**5.4**	0.128	**15.6**	0.158	**126.7**
RGPA	**0.035**	**5.4**	0.127	**15.6**	**0.139**	143.8

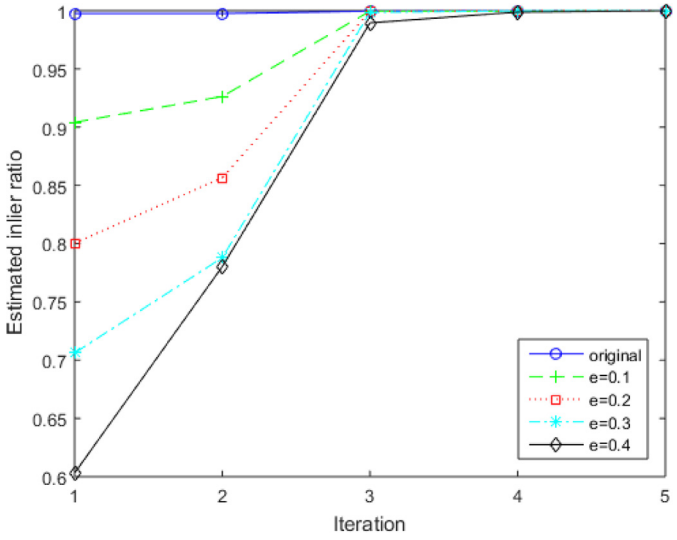


Fig. 8. Inlier ratio vs. RGPA iteration under different noise levels.

struction cannot be achieved and the re-projection error is large even under low noise ratio.

We then compare RGPA with state-of-the-art hierarchical (Gherardi et al., 2010) (HIER), global (Arie-Nachimson et al., 2012; Jiang et al., 2013) and the incremental (Wu, 2013) (VSFM) SfM methods using 3 small datasets with ground-truth cameras (Strecha et al., 2008). The initial focal lengths are extracted from EXIF. Table 2 reports mean errors for recovered camera rotations (R_{err}, in degree) and translations (T_{err}, in mm) with respect to ground truths. Although all methods yield correct structures from which good densely reconstructed points can be computed (Furukawa et al., 2010) (see supplementary material), the camera pose accuracy of the global methods (Arie-Nachimson et al., 2012; Jiang et al., 2013) is inferior to other approaches. This weakness is amplified on more unstructured datasets, and prohibits the usage of global methods on some challenging datasets. The two hierarchical methods (HIER and RGPA) perform the best in this test and have comparable results.

Finally we test bigger datasets with more unstructured organizations (Hane et al., 2013; Samantha site; BigSFM site). We report the results of the 5 image sets in Table 3. For scenes with big camera rotations the camera registration of the global methods (Arie-Nachimson et al., 2012; Jiang et al., 2013) becomes unstable and fails to provide reliable initialization (see Supplementary Material). Therefore, we only compare RGPA with Gherardi et al. (2010) and Wu. (2013) in Table 4. It is worth noting that due to the outlier detection scheme in Section 2.3.3 the intermediate BAs in Gherardi et al. (2010) can also be removed to achieve a much faster reconstruction. The running time in Table 4 excludes the preprocessing step of track generation and pairwise reconstruction, which is common for the three algorithms. For preprocessing we used a 192-core cluster to speedup feature matching, and the running times are given in the last column of Table 3. The RANSAC estimated inlier ratios are also reported in Table 3.

"Building" is the easiest one among the 5 datasets. All three methods recover all 128 cameras. RGPA and HIER rebuild fewer points because they only perform a spanning reconstruction that ignores many redundant points. This spanning reconstruction covers the main structure of the scene and the densely rebuilt scenes exhibit comparable qualities. Albeit written in non-optimized Matlab code, RGPA and HIER are both much faster than VSFM. The reasons are twofold: 1) in RGPA and HIER the intermediate BAs are removed; 2) the tree structure of hierarchical SfM is more efficient than the line structure of incremental SfM. Even though RGPA uses an iterative alignment procedure for 3D registration, it still runs faster than HIER because RGPA has a much shallower reconstruction tree.

"Dante" is a very challenging dataset, which attempts to cover a moderately complicated, closed scene with only 39 unordered cameras. It has the highest outlier ratio among the 5 datasets, and the maximum local outlier ratio reaches 30% (see Table 3). Big camera rotations and sparse associations between cameras cause problems for global methods (see supplementary materials). Incremental and hierarchical methods are not as vulnerable as global methods. VSFM recovers 38 cameras but fails to recover the last one due to too few inlier projections, whereas RGPA and HIER successfully recover all cameras. The reconstructed points are satisfactory for RGPA, but noisy for HIER due to drift (see Fig. 9).

In Table 5 we report results with and without intermediate BAs for "Building" and "Dante" to show that intermediate BAs are not necessary in RGPA. The removal of intermediate BAs causes larger re-projection error before the final BA. However, after the final BA the differences become insignificant, and the recovered structures are visually close.

Table 3
A description of the five datasets tested in our experiment.

Dataset	No. of images	Average inlier ratio	Min local inlier ratio	Mean error before/after BA	Preprocessing time (s)
Building	128	98%	90%	3.46 / 0.89	33
Dante	39	91%	70%	3.99 / 0.46	6
Piazzabra	380	94%	79%	3.37 / 0.61	289
Trevi	751	99%	93%	8.78 / 2.71	887
Colosseum	822	97%	76%	8.92 / 2.13	948

Table 4
Results comparison on the five image sets in Table 3 (* indicates that two partial reconstructions were obtained).

Dataset	Algorithm	No. of points recovered	No. of views recovered	Time (s)	Tree depth
Building	VSFM (Wu et al., 2011)	82,275	128	129	–
	HIER (Gherardi et al., 2010)	54,611	128	37	26
	RGPA	55,281	128	31	4
Dante	VSFM	26,786	38	15	–
	HIER	19,090	39	11	11
	RGPA	19,127	39	11	3
Piazzabra	VSFM	182,370/23,792*	348/42*	781	–
	HIER	148,795	379	134	37
	RGPA	144,973	379	118	5
Trevi	VSFM	197,051	750	1282	–
	HIER	107,368	735	679	236
	RGPA	115,617	735	257	3
Colosseum	VSFM	244,431	821	1088	–
	HIER	158,888	809	597	178
	RGPA	164,138	809	317	4

Table 5
A comparison of RGPA with and without intermediate BAs.

Dataset	BA strategy	# BAs	Mean re-projection error (before/after final BA)
Building	With inter. BAs	20	1.34 / 0.87
	No inter. BAs	1	3.46 / 0.89
Dante	With inter. BAs	9	0.73 / 0.42
	No inter. BAs	1	3.99 / 0.46

Table 6
Time spent on RGPA operations for the original and the simulated parallel implementations.

Dataset	Original time t_o (s)	Parallel time t_p (s)	t_o/t_p
Building	19.69	3.82	5.15
Dante	4.76	1.00	4.76
Piazzabra	66.19	10.49	6.31
Trevi	136.72	31.87	4.29
Colosseum	122.18	19.50	6.27

"Piazzabra" contains many narrow baseline cameras and duplicate structures, which lead to a lot of incorrect epipolar geometries and a large proportion of mismatches. There are several local structures with relatively high outlier ratios of around 20%. VSFM is negatively affected and recovers two separate parts, but fails to join them together. RGPA and HIER successfully recover the full scene. The result of HIER is not as good as RGPA due to some visible stitching cracks (see Fig. 10).

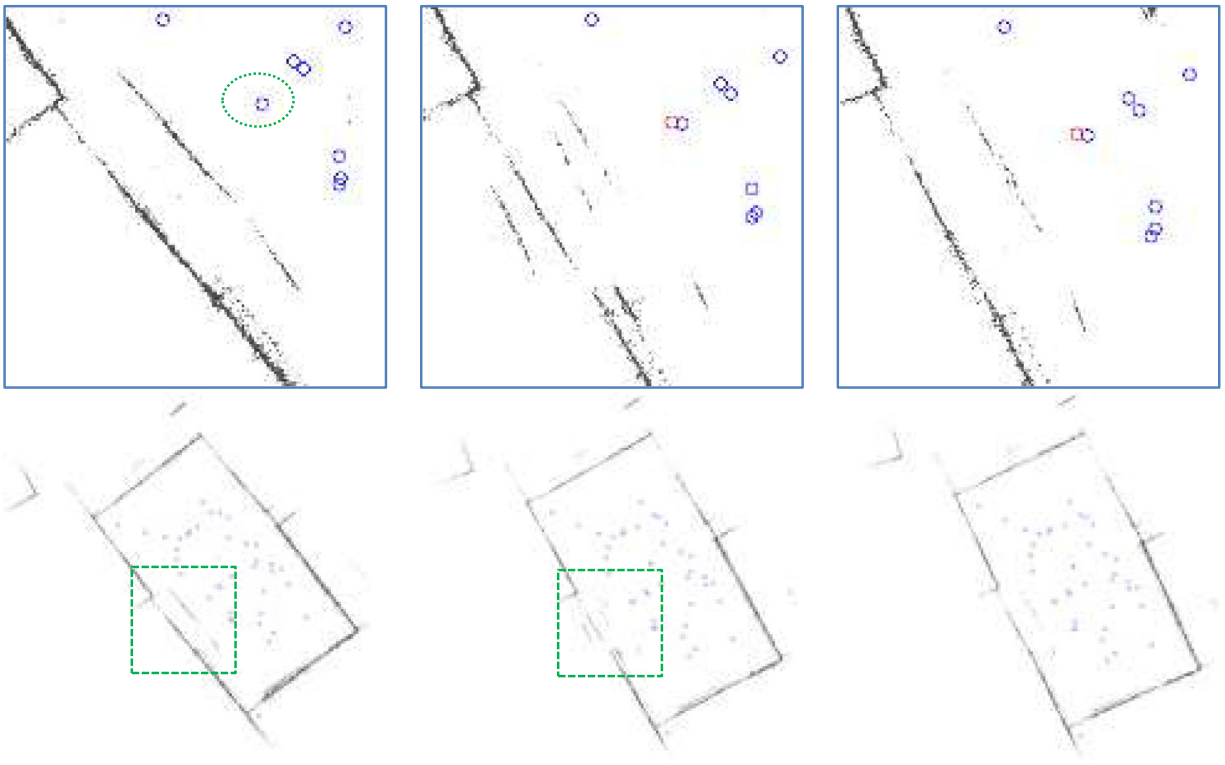
"Trevi" and "Colosseum" are two biggest subsets in the Rome16k dataset (BigSFM site). All photos with valid EXIF focal information are used in our test. Both RGPA and VSFM give satisfactory results. HIER yields acceptable results, but with a few visible errors (see Fig. 11). This is because the small tree depth and the cross-checking among multiple structures make RGPA less sensitive to drift than HIER. RGPA and HIER recover fewer cameras than VSFM because some "singleton" images are automatically removed to ensure a good status of the maximum connected component of the scene. The advantage of RGPA over HIER in speed is evident on these two datasets. The MCDS grouping in RGPA is superior to the one-by-one strategy in HIER, and is able to merge as many as 80 structures at a time. As a result, RGPA takes only 3 or 4 levels to complete the reconstruction, whereas HIER needs 100+ levels. The reason why "Trevi" has 3 levels whereas "Colosseum" has 4 is that the images in "Trevi" are better aggregated and thus MCDS is able to group more structures at a time.

For all experiments, the mean relative deviations of the focal lengths computed by the final BA from the initial EXIF values are below 2%. The results of RGPA are plotted in Fig. 12.

We also give a coarse estimate of the running speed of parallel RGPA. Table 6 compares the total time for the original and the simulated parallel implementation of RGPA. Using parallelization, the running speed is 4–7 times faster. This is because RGPA can be parallelized at the same tree level. Note that HIER cannot be easily parallelized, because in HIER every merging operation is determined after the previous merging has finished.

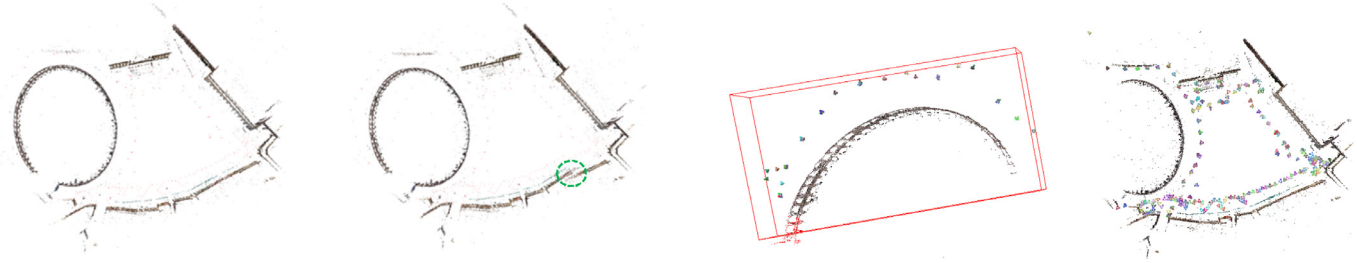
In summary, RGPA runs faster than HIER and VSFM because its tree-structured organization can merge multiple structures at a time, and it needs no intermediate BAs. The built-in outlier detecting mechanism makes RGPA insensitive to errors that may hinder other SfM approaches, thus achieving more robust reconstruction.

The limitation of RGPA is that it relies moderately on the tightness of connections among structures to achieve a full reconstruction. Roughly speaking, for loosely connected structures the scene will split into several connected components at weak edges without sufficient common points, and only partial reconstructions can be accomplished for each connected component. This generally occurs for randomly collected image sets of a large-scale scene. Fig. 13 provides the partial reconstructions of the maximum connected components of 3 less-connected datasets (Wilson and Snavely, 2014). The recovered cameras are significantly fewer than the original dataset, but the scene structures are correctly reconstructed. Treating large datasets with loose connectivity more reliably is an interesting future work.



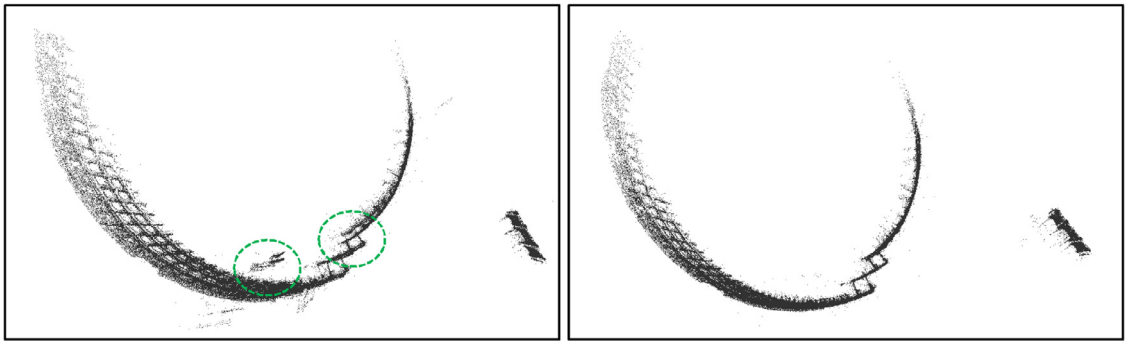
(a) VSFM (38 cameras) (b) HIER (39 cameras) (c) RGPA (39 cameras)

Fig. 9. The results of 3 algorithms for "Dante". The bottom row shows the global top views. All cameras except camera 20 are plotted as blue circles. Camera 20 (plotted as red circles for HIER and RGPA) is not recovered by VSFM (within the dotted ellipse of (a)). The top row shows an enlargement of the region marked with a dashed rectangle. The HIER result has noise in this region. Camera 20 is also in this region. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) RGPA (b) HIER (c) VSFM (two parts)

Fig. 10. The results of 3 algorithms for "Piazzabra". RGPA yields a satisfactory reconstruction. HIER yields an acceptable result but with visible stitching cracks (encircled by green dashed ellipse). VSFM yields two separate parts of the scene (The 2nd part in red solid cube). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) HIER (b) RGPA

Fig. 11. A comparison of reconstructed 3D points by HIER and RGPA for "Colosseum". RGPA yields satisfactory reconstruction. HIER yields an acceptable but less clean result, with visible errors caused by drift (encircled by green dashed ellipses). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

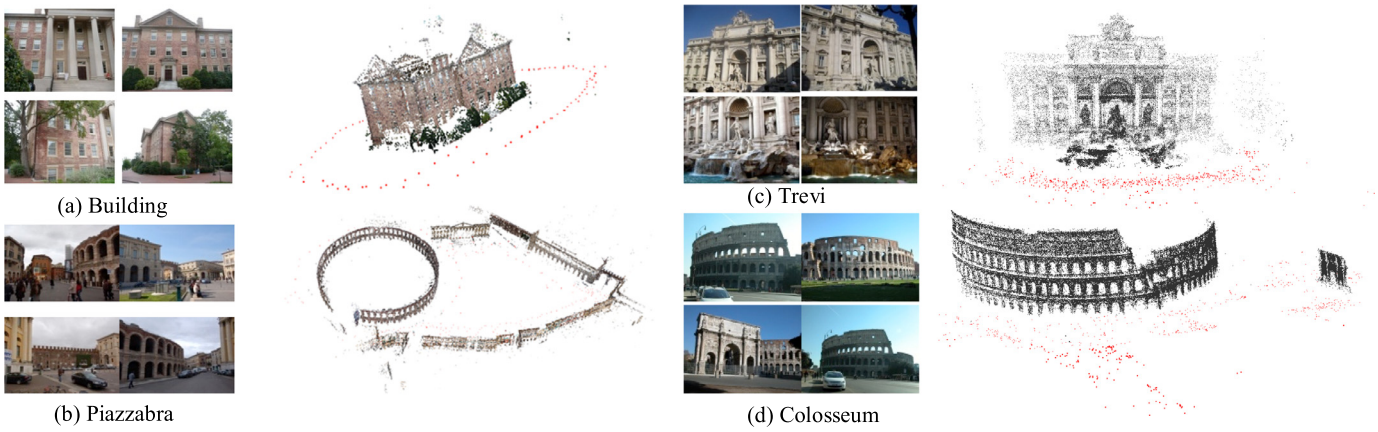


Fig. 12. RGPA results of 4 image sets (The result of "Dante" is in Fig. 1). Left: 4 images of the image set; Right: the reconstructed points and cameras. Cameras are plotted as red points. For (a) and (b), 3D scenes are rendered using colored points because the photos are taken in consistent lighting; For (c) and (d), 3D scenes are rendered using uniform gray-colored points because lighting varies among photos.

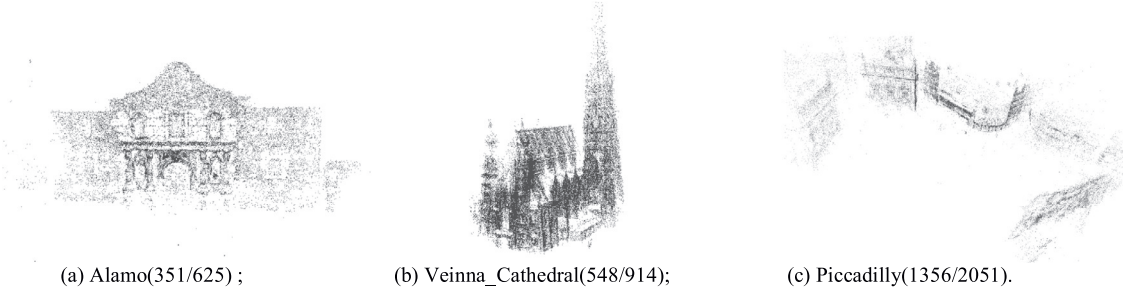


Fig. 13. Reconstruction of 3 less-connected datasets. The numbers in the parentheses denote the number of images in the maximum connected component and the original dataset respectively.

4. Conclusion

In this paper, we have proposed a tree-structured SfM algorithm that can be executed quickly, while also reliably handling outliers. Our method has several advantages. 1) By employing RGPA we can merge multiple structures simultaneously. This substantially increases the efficiency. 2) Using built-in cross-checking between multiple structures we can detect and remove outliers effectively. This makes our algorithm resistant to noise. 3) Using a spanning structure and minimum connected dominating sets we organize unordered images and group structures efficiently. This leads to a quick and reliable bottom-up reconstruction approach. The experiments confirm that our method outperforms the state-of-the-art in both efficiency and robustness for highly unstructured scenes.

Acknowledgments

This research is supported by National Natural Science Foundation (NSF) of China under grant nos. 61232014, 61421062 and 61472010, National Basic Research Program of China (973 Program) under grant no. 2015CB351806, and National Key Technology R&D Program of China under grant no. 2015BAK01B06. Zhouchen Lin is supported by 973 Program under grant no. 2015CB352502, NSF of China under grant nos. 61625301 and 61231002, and Qualcomm.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cviu.2017.02.005.

References

Agarwal, S., Snavely, N., Simon, I., Seitz, S.M, Szeliski, R., 2009. Building Rome in a day. In: ICCV, pp. 72–79.
Arie-Nachimson, M., Kovalsky, S.Z., Kemelmacher-Shlizerman, I., Singer, A., Basri, R., 2012. Global motion estimation from point matches. In: Proc. 3DPVT.
Bhowmick, B., Patra, S., Chatterjee, A., Govindu, V., Banerjee, S., 2014. Divide and conquer: efficient large-scale structure from motion using graph partitioning. In: ACCV14.
BigSFM site, http://www.cs.cornell.edu/projects/bigsfm/#data.
Brown, M., Lowe, D., 2005. Unsupervised 3d object recognition and reconstruction in unordered datasets. In: 3DIM, pp. 56–63.
Chatterjee, A., Govindu, V.M., 2013. Efficient and robust large-scale rotation averaging. In: ICCV.
Choi, S., Kim, T., Yu, W., 2009. Performance evaluation of RANSAC family. In: BMVC.
Corsini, M., Dellepiane, M., Ganovelli, F., Gherardi, R, Fusiello, A., Scopigno, R., 2013. Fully automatic registration of image sets on approximate geometry. IJCV 102, 91–111.
Crandall, D., Owens, A., Snavely, N., Huttenlocher, D., 2011. Discrete-continuous optimization for large-scale structure from motion. In: CVPR, pp. 3001–3008.
Crosilla, F., Beinat, A., 2002. Use of generalised Procrustes analysis for the photogrammetric block adjustment by independent models. ISPRS J. Photogramm. Remote Sens.
Crosilla, F., Beinat, A., 2006. A forward search method for robust generalised Procrustes analysis. Data Anal. Classif. Knowl. Organ.
Eggert, D., Lorusso, A., Fisher, R., 1997. Estimating 3d rigid body transformations: a comparison of four major algorithms. Mach. Vision Appl. 9 (5), 272–290.
Enqvist, O., Kahl, F., Olsson, C., 2011. Nonsequential structure from motion. In: OMNIVIS.
Fischler, M., Bolles, R., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24 (6), 381–395.
Frahm, J., et al., 2010. Building Rome on a cloudless day. In: ECCV.
Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R., 2010. Towards internet-scale multi-view stereo. In: Proc. CVPR.
Gherardi, R., Farenzena, M., Fusiello, A., 2010. Improving the efficiency of hierarchical structure-and-motion. In: CVPR.
Govindu, V., 2006. Robustness in motion averaging. In: ACCV, pp. 457–466.
Guha, S., Khuller, S., 1998. Approximation algorithms for connected dominating sets. Algorithmica 20 (4), 374–387.

Hane, C., Zach, C., Cohen, A., Angst, R., Pollefeys, M., 2013. Joint 3D scene reconstruction and class segmentation. In: CVPR, pp. 97–104.
Hartley, R., Trumpf, J., Dai, Y., Li, H., 2013. Rotation averaging. IJCV 103 (3).
Havlena, M., Torii, A., Knopp, J., Pajdla, T., Randomized structure from motion based on atomic 3D models from camera triplets, CVPR 2009.
Havlena, M., Torii, A., Pajdla, T., 2010. Efficient structure from motion by graph optimization. In: Proc. ECCV.
Heinly, J., Dunn, E., Frahm, J.M., 2014. Correcting for duplicate scene structure in sparse 3D reconstruction. In: ECCV.
Jiang, N., Cui, Z., Tan, P., 2013. A global linear method for camera pose registration. In: ICCV.
Li, X., Wu, C., Zach, C., Lazebnik, S., Frahm, J., 2008. Modeling and recognition of landmark image collections using iconic scene graphs. In: ECCV.
Lou, Y., Snavely, N., Gehrke, J., 2012. Matchminer: efficiently mining spanning structures in large image collections. In: ECCV.
Lourakis, M., Argyros, A., 2009. SBA: a software package for generic sparse bundle adjustment. ACM Trans. Math. Softw. 36 (1).
Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. IJCV 60 (2), 91–110.
Martinec, D., Padjla, T., Robust rotation and translation estimation in multiview reconstruction. CVPR 2007, pp. 1–8.
Moulon, P., Monasse, P., Marlet, R., 2012. Adaptive structure from motion with a *contrario* model estimation. In: ACCV.
Moulon, P., Monasse, P., Marlet, R., 2013. Global fusion or relative motions for robust, accurate and scalable structure from motion. In: ICCV.
Ni, K., Dallaert, F., HyperSfM, 2012. International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission.
Nister, D., Reconstruction from uncalibrated sequences with a hierarchy of trifocal tensors. ECCV 2000, 649–663.
Nister, D., 2004. An efficient solution to the five-point relative pose problem. IEEE Trans. PAMI 26 (6), 756–777.
Nister, D., Stewenius, H., Scalable recognition with a vocabulary tree. CVPR 2006, 2161–2168.
Olsson, C., Enqvist, O., 2011. Stable structure from motion for unordered image collections. SCIA, 6688, pp. 524–535.
Olsson, C., Erisson, A., Hartley, H., 2011. Outlier removal using duality. In: CVPR.
Pizarro, D., Bartoli, A., 2011. Global optimization for optimal generalized procrustes analysis. In: CVPR, pp. 2409–2415.
Pollefeys, M., Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., Koch, R., 2004. Visual modeling with a hand-held camera. IJCV 59 (3), 207–232.
Prim, R.C., 1957. Shortest connection networks and some generalizations. Bell Syst. Tech. J. 36, 1389–1401.
Raguram, R., Frahm, J.-M., RECON, 2011. Scale-adaptive robust estimation via residual consensus. In: ICCV, pp. 1299–1306.
Samantha site, http://www.diegm.uniud.it/fusiello/demo/samantha/.
Shah, R., Deshpande, A., Narayanan, P., Multistage SFM: revisiting incremental structure from motion, 2014 3DV14.
Sinha, S., Steedly, D., Szeliski, R., 2012. A multi-stage linear approach to structure from motion. Trends and Topics in Computer Vision. Springer, pp. 267–281.
Snavely, N., Seitz, S., Szeliski, R., 2006. Photo tourism: exploring photo collections in 3d. ACM TOG 25, 835–846.
Snavely, N., Seitz, S., Szeliski, R., 2008. Skeletal graphs for efficient structure from motion. In: CVPR.
Strecha, C., von Hansen, W., Gool, L.V., Fua, P., Thoennessen, U., 2008. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: CVPR, pp. 2838–2845.
Triggs, B., McLauchlan, P., Hartley, R., Fitzgibbon, A., 2000. Bundle adjustment—a modern synthesis. Vision Algorithms: Theory and Practice 298–372.
Wilson, K., Snavely, N., 2014. Robust global translations with 1DSfM. In: ECCV.
Wu, C., Agarwal, S., Curless, B., Seitz, S., 2011. Multicore bundle adjustment. In: Proc. CVPR, pp. 3057–3064.
Wu, C., 2013. Towards linear-time incremental structure from motion. In: 3DV.
Zach, C., Klopschitz, M., Pollefeys, M., 2010. Disambiguating visual relations using loop constraints. In: CVPR.