Efficient Large-scale Photometric Reconstruction Using Divide-Recon-Fuse 3D Structure from Motion

Yueming Yang[1], Ming-Ching Chang[1,2], Longyin Wen[1], Peter Tu[2], Honggang Qi[3] and Siwei Lyu[1]

[1]State University of New York, Albany, NY, USA, {yyang7,mchang2,lwen,slyu}@albany.edu
[2]GE Global Research, Niskayuna, NY, USA, {changm,tu}@ge.edu
[3]University of China Acaemy of Science, Beijing, China, {hgqi}@ucas.ac.cn

Abstract

We propose an efficient framework for large-scale 3D reconstruction from a large set of photos following the Structure-from-Motion (SfM) paradigm with divide-conquer and fusion. Our main novelty is to ensure commonality from overlaps between image sets corresponding to their reconstructions, which facilitates effective stitching and fusion. Specifically, such commonality is ensured by selecting a set of duplicated images (which are termed anchor images*) in adjacent image sets prior to the 3D reconstruction. The anchor images can assist accurate fusion of the 3D point clouds. We describe an efficient RANSAC scheme for pairwise stitching. Our method is intuitively scalable to large site reconstruction via subdivision and fusion following a graph construct. We further describe another RANSAC algorithm to improve loop closure in our anchor image approach. Experimental results on reconstructing a large portion of a university campus demonstrate the efficacy of our method.*

1. Introduction

3D model reconstruction of large sites and urban environments from multiple images with Structure-from-Motion (SfM) has been an active topic in the last decade. The SfM process starts with finding correspondences between two images, and generates a point cloud by calculating the 3D locations of the corresponding points with triangulation. The process is continued with more images added sequentially to generate a denser point could. With advanced hardware, SfM has been successfully extended to reconstruct large-scale man-made structures of popular tourist sites in the world [17, 8, 18, 1, 6]. 3D reconstruction models are useful in many research areas, such as virtual and augmented reality in education [4], urban planning, digital heritage, smart city, and surveillance [24]. However, naive SfM reconstruction requires image feature matching between *all pairs* of images, which becomes a bottleneck for running time and storage for such algorithms to scale up to large image sets. The time complexity for pairwise matching of n images is $O(n^2)$, which takes days when there are thousands of images[17, 22].

In this work, we propose a scalable framework for large scale 3D reconstruction that we term **Divide-Recon-Fuse SfM**. The idea is to divide the large site into different scenes, and then cluster the images into sets of individual scenes. 3D reconstruction on each set can be performed independently and in parallel, and resulting 3D models can be stitched and fused together. The overall reconstruction time can be reduced by the order of the subdivision, with the cost to fuse individual reconstructions into a single model. Note that the outcome of SfM is a 3D model in an arbitrary virtual coordinate system. Thus, the fusion involves finding the best scaling and registration parameters to align and stitch various 3D point clouds together. On a related front, reliable and accurate fusion is crucial for both on-line and off-line reconstruction methods, thus a successful fusion method is the key for large-scale reconstruction.

Main contributions of our work are three-folds.

- We describe a divide-and-conquer strategy to accelerate large scale SfM reconstruction. The framework is general that it supports any 3D reconstruction methods that generate 3D models from a set of images, including unorganized image sets, and provides the mapping between 2D image features to 3D correspondences.
- We propose a novel formulation of adding sets of duplicated "anchor images" in adjacent image sets, which provides powerful hints in the stitching and fusion of individual reconstructions.
- We further investigate the well-known loop closure problem in fusion, and propose an efficient RANSAC

978-1-5090-3811-4/16/$31.00

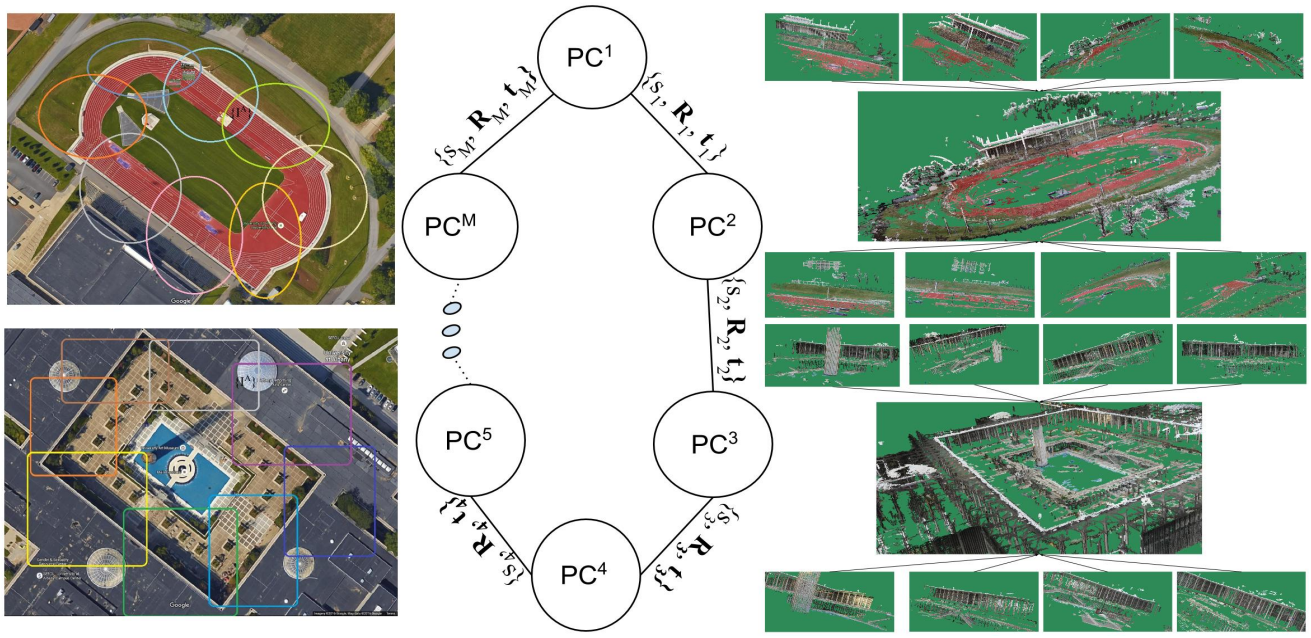


Figure 1. **Divide-and-conquer large-scale 3D reconstruction** *of two campus sites: Track Field (*15.52*M points, top) and Campus Podium (*10.65*M points, bottom) , both using 8 sub-divisions followed by point cloud alignment and fusion. Top down aerial views on the left are courtesy of Google Map. The* Scene Adjacency Graph *(SAG) is illustrated with each sub-model denoted as a* node*, and the fusion between two sub-models denoted as a* link*.*

solution based on the anchor images.

We practically evaluate the proposed Divide-Recon-Fuse SfM pipeline on two large-scale campus datasets in Section 4.

2. Background

We briefly survey literatures on large-scale photometric reconstruction and the loop closure problem in 3D model fusion. Musialsk *et al*. [12] survey recent technologies of urban reconstruction, and point out that current state-of-art do not scale well to the huge amount of data. The processing in terms of 3D reconstruction and the management of 3D models remain a challenge.

With dozens of computer cores, the groundbreaking work of Snavely *et al*. [17, 1] reconstructed sparse point-clouds of a few world-renown tourist sites solely from Internet photos using SfM. It leads to an upswing of large scale 3D reconstruction, Frahm *et al*. [6] and Wu[22] later developed efficient implementations of the same algorithm with GPU acceleration. Furukawa *et al*. [7] developed a patch-based multi-view stereo algorithm to derive enhanced dense point-cloud. Research derives from these lines of works include datasets with geo-registered images annotated with 3D orientations [11], heritage preservation via 3D digital documentation [13] and immersive virtual reality [23].

3D reconstruction from Internet photos is an active topic, since on-line resources represent a very large dataset available for reconstruction. However, photos on community sites such as Flicker are typically "iconic" (i.e., covering only famous landmarks) [15, 2, 9], and usually very redundant. Given its large volume, 3D reconstruction from Internet photos without divide-and-conquer is generally considered unachievable. Preemptive feature matching [22], which matches the first few features for a pair of images to determine whether to continue all matchings or not, can degrade the final matching quality. To overcome the lack of suitable data for 3D reconstruction, Untzelmann *et al*. [20] solicited photos using a crowd-sourcing system, and generate an unified point cloud using GPS information.

The work of [2] is most relevant to us. 3D reconstruction is performed on Internet photos by constructing an image matching graph using a vocabulary tree. Camera rotation and translation of the connecting images are available from the result of SfM, which is used to estimate the relative transformations between a pair of reconstructions. One strong assumption is that transformation estimation depends completely on the accuracy of the SfM-extracted camera projection matrix, which is known to be noisy. This results in merging gaps between models and degraded the final model, as we compare our results in Figure 3. To address this issue, we perform fusion based on aligning sub-models directly, instead of relying on individual reconstructed camera parameters. Furukawa et al. [16] address the problem of generating a visually plausible model from the fusion of 3D models in divide-and-conquer 3D reconstruction. Under the assumption of Manhattan scene and that partially reconstructed sub-models must form a closed loop, the stitching

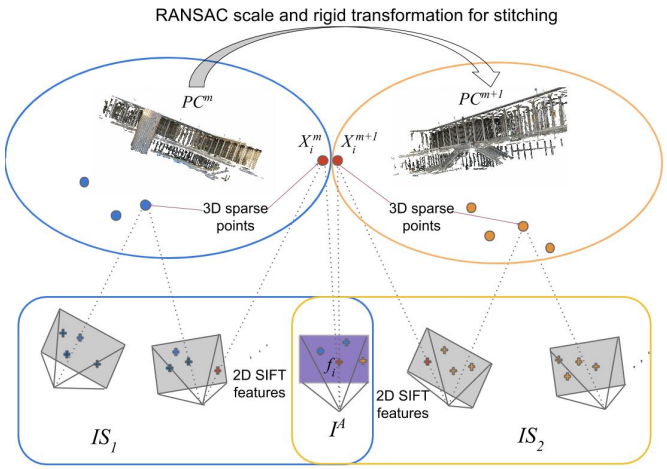


Figure 2. **Pairwise SfM point cloud fusion** by leveraging known correspondence between 2D image features and their reconstructed 3D points. Anchor image I_A (as an overlap of the two SfM reconstructions) serves as the bridge between the two point clouds (PC^m and PC^{m+1}) for fusion (see text). In practice, we have multiple anchor images, and each contains many feature points for the fusion calculation. An efficient RANSAC scheme is developed in Table 1.

is performed under the assistance of semantic labels.

Loop Closure. A main challenge in the sub-models fusion step is how to close the gap when there exists a loop among the consecutive fusion steps. Loop closure is a well known problem in the area of SLAM for years, where mapping registration is typically involved [21].

3. Method

Our 3D reconstruction pipeline starts with dividing the large site into several overlapping sets, as shown in Figure 1. Each *scene section* shares a sufficient overlapping portion with adjacent scene sections for the fusion purpose. Images in each scene section can be used to reconstruct a 3D point cloud representation using the standard Structure-from-Motion (SfM) algorithms [1, 22], followed by a dense reconstruction [7, 10]. The resulting 3D point cloud from SfM is in individual *virtual* coordinate system, which is determined in the SfM process depending on the algorithm's iterative process and camera focal length. We will briefly overview 3D SfM in Section 3.1.

The fusion step is to merge all point clouds of different virtual coordinate systems into one common coordinate system. As such, we need to estimate their relative conformal transforms (i.e., rotation, global scaling and translation). The fusion parameters to be determined includes 7 degrees of freedom — up to a scaling s, rotation $\mathbf{R}$, and translation $\mathbf{t}$. Different from the approach of [2], where the fusion relies on the images of overlapping scenes to be successfully

Table 1. **Procedure** *EstPairTransRANSAC*($\mathbf{X}^m, \mathbf{X}^{m+1}, K, \tau$).

for $k = 1$ to K **do**
  idx = randomly select 6 numbers from $index$
  $\mathbf{X}_i^m \leftarrow \mathbf{X}^m(idx), \mathbf{X}_i^{m+1} \leftarrow \mathbf{X}^{m+1}(idx)$
  $\mathbf{R}_i, s_i, \mathbf{t}_i \leftarrow$ **EstTransformSVD** $(\mathbf{X}_i^m, \mathbf{X}_i^{m+1})$
  $e_i \leftarrow |s_i\mathbf{X}^{m+1} - \mathbf{R}_i\mathbf{X}^m - \mathbf{t}_i|$
  $n_i \leftarrow$ number of items in e_i which are $\leq \tau$
  if $n_i \geq n\_inliers$ **then**
    $n\_inliers \leftarrow n_i$
    $\mathbf{R}, s, \mathbf{t}, e \leftarrow \mathbf{R}_i, s_i, \mathbf{t}_i, e_i$
  end if
end for
return $\mathbf{R}, s, \mathbf{t}, e, n\_inliers$

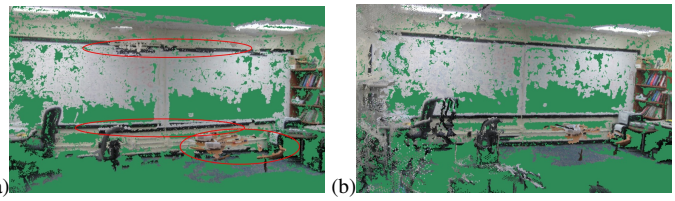
matched, we select **anchor images** that are shared across different sets, and use them to provide reference points to fuse multiple point clouds.

The fusion process can be formulated as a graph construct, motivated from the Regions Adjacency Graph (RAG) [3] in image segmentation [19]. In our case, each sub-model is denoted as a *node*; and the fusion between two sub-models is denoted as a *link*, thus the fusion forms an undirected graph, which we term the *Scene Adjacency Graph* (SAG). Specifically, each link in the SAG involves the use of anchor images from the clusters of the two ending nodes to calculate an optimal set of transformations $\{s, \mathbf{R}, \mathbf{t}\}$ for fusion alignment. More details about anchor images is given in Section 3.2.

One major issue in the fusion of divide-and-conquer 3D reconstruction is the problem of *loop closure* [21], which has been extensively studied in Visual SLAM. In general the loop closure problem is best compensated by amortizing (or spreading) registration errors properly across multiple fusion steps, and thus avoid error accumulation. In the SAG representation, any closed loop in the SAG graph represents a potential inconsistency (or glitch) in the fusion progress. In Section 3.3, we will describe our solution to the loop closure problem, in which the matching of 3D points from anchor image features between consecutive image sets are explicitly enforced and examined to ensure proper loop closure, following an efficient global RANSAC algorithm.

3.1. Divide and Conquer SfM Reconstruction

The Structure-from-Motion (SfM) 3D photometric reconstruction [17, 1, 22] is summarized as follows. SIFT-like feature keypoints are detected for image pairwise matching. A set of sparse 3D correspondences are calculated using Bundle Adjustment. The output of SfM includes: (1) a colored 3D point cloud, (2) 3D correspondences of 2D image feature points in the 3D reconstruction, (3) estimated camera parameters for the 2D images, including the projection matrix $\mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}]$, where $\mathbf{K}$ is the camera's intrinsic parameters, and $[\mathbf{R}|\mathbf{t}]$ are the extrinsic parameters. The work in [2] uses the R and T of the overlapping images derived from SfM to align and merge point clouds, thus it can



(a) (b)

Figure 3. **Pairwise fusion** of two point clouds of a lab room. (a) Fusion following [2] from direct estimation of scale, translation, and rotation of the two clouds using known SfM camera parameters (center and orientation) from the "connection images" leads to inconsistency in fusion (red). (b) Our RANSAC scheme leveraging the 2D-3D-2D feature matching described in Figure 2 and Table 1 aligns the two clouds well.

lead to fusion gaps due to the inevitable SfM optimization error. In comparison, we perform robust estimation using a RANSAC scheme in finding the corresponding 3D point sets, by tracing 2D image features in the anchor images. Our alignment is shown to be more accurate as in Figure 3. Lastly, the crucial step in producing a refined 3D model is dense correspondence by stereo matching[7], which is using information from confirmed, overlapping image pixels in an epipolar line search.

The scaling issue is the primary bottleneck for the SfM computation, since the computational time of pairwise image feature matching dominates the whole pipeline. To see this, given a set of N images divided into K image sets with each set contains about $\frac{N}{K}$ images. The Computational efficiency gain is approximately K times faster, which is calculated as follows. The brute-force time complexity is C_2^N. In the divide-and-conquer scheme, the complexity is $K \times C_2^{\frac{N}{K}}$. The speed up ratio is then $\frac{C_2^N}{K \times C_2^{\frac{N}{K}}} = K\frac{N-1}{N-K} \simeq K$. As N is much larger than K, speed gain is approximately K times faster. Experimental results in Table 3 confirms such observation. Note that each image set should contain sufficient number of images for success SfM reconstruction, so K cannot be arbitrarily large.

Scene Clustering. Given a large set of photos of a site which we wish to reconstruct, there are two popular ways to cluster the images. One can cluster images based on appearance features from the image pixels in the "scene space". In comparison, one can leverage available GPS position of the camera and form clusters in the "camera space". Note that scene-space clustering is not always correct due to limitations of the appearance matching; and camera-space clustering can mix clusters, since photos shooting the same scene can be taken from far away with different lens and field of view.

We follow a scene-space clustering scheme (see Figure 1), where images are first automatically clustered and later manually examined and organized into several scene clusters with the false clustered and blur images removed. Images are collected from different positions and viewpoints,

Table 2. **Procedure** *EstLoopTransRANSAC*($\mathbf{X}$, K, L, τ, ϕ).

R $\leftarrow$ []
t $\leftarrow$ []
$s \leftarrow$ []
for $l = 1$ to L **do**
  $n\_inliers\_global \leftarrow 0$
  $\mathbf{R}_l \leftarrow$ []
  $\mathbf{t}_l \leftarrow$ []
  $s_l \leftarrow$ []
  for $m = 1$ to M **do**
    n = the length of $\mathbf{X}$
    $\mathbf{R}_t, s_t, \mathbf{t}_t \leftarrow \mathbf{EstPairTransRANSAC}(\mathbf{X^m}, \mathbf{X^{m+1}}, K, \tau)$.
    $\mathbf{R}_l$ **append** $\leftarrow \mathbf{R}_t$
    $\mathbf{t}_l$ **append** $\leftarrow \mathbf{t}_t$
    s_l **append** $\leftarrow s_t$
  end for
  $\mathbf{X_t^1} \leftarrow \mathbf{X^1}$
  for $m = 1$ **to** $M - 1$ **do**
    $\mathbf{X_t^1} = \frac{1}{s_l[m]}(\mathbf{R}_l[m]\mathbf{X_t^1} + \mathbf{t}_l[m])$
  end for
  $e_j \leftarrow |s_l[M]\mathbf{X^1} - \mathbf{R}_l[M]\mathbf{X_t^1} - \mathbf{t}_l[M]|$
  $n_j \leftarrow$ **number of items in** e_j **which are** $\leq \phi$
  if $n_j \geq n\_inliers\_global$ **then**
    $n\_inliers\_global \leftarrow n_j$
    $\mathbf{R}, s, \mathbf{t}, e \leftarrow \mathbf{R}_l, s_l, \mathbf{t}_l, e_j$
  end if
end for
return $\mathbf{R}, s, \mathbf{t}, e, n\_inliers\_global$

where the camera views are limited within the range of the scene cluster as much as possible. Our method works on unorganized image sets, thus scene images can be collected at different time and weather conditions. The SfM algorithm allows incremental 3D reconstruction, at the cost of pairwise matching each new image with all images in the scene cluster.

3.2. Pairwise SfM Fusion Using Anchor Images

We show that two adjacent 3D point clouds produced from SfM can be precisely transformed, aligned, and fused by matching the feature points on the anchor images. Figure 2 explains this process. Since we select anchor images to be from the union of the two image sets, image features on the anchor images are guaranteed to present in the SfM sparse reconstruction points, thus they can be matched for fusion. The fusion process involves the matching of 2D image features (on one anchor image in one image sets) to the corresponding 3D point (where such correspondence is known from the SfM step), and the matching of the two versions of the very 3D point in the two sub-models. We propose an efficient RANSAC process detailed in Table 1.

Specifically, consider two adjacent sub-models PC^m and PC^{m+1}. We denote the 2D SIFT feature point set $\{f_i\}$ (from an anchor image) corresponds to 3D point set $\{X^m\}$ in the SfM reconstruction of PC^m. Similarly, 2D feature points $\{f_j\}$ correspond to their 3D points $\{X^{m+1}\}$ in PC^{m+1}. In the fusion of PC^m and PC^{m+1}, the intersection $\{f_i\} \cap \{f_j\}$ and their corresponding 3D points $\{X^m\} \cap \{X^{m+1}\}$ are the set that can be matched to solve for the alignment transformation $\{s, \mathbf{R}, \mathbf{t}\}$. We adopt an efficient RANSAC for such calculation, as shown in Table 1, where **EstTransformSVD** [5] is a standard solver for solv-

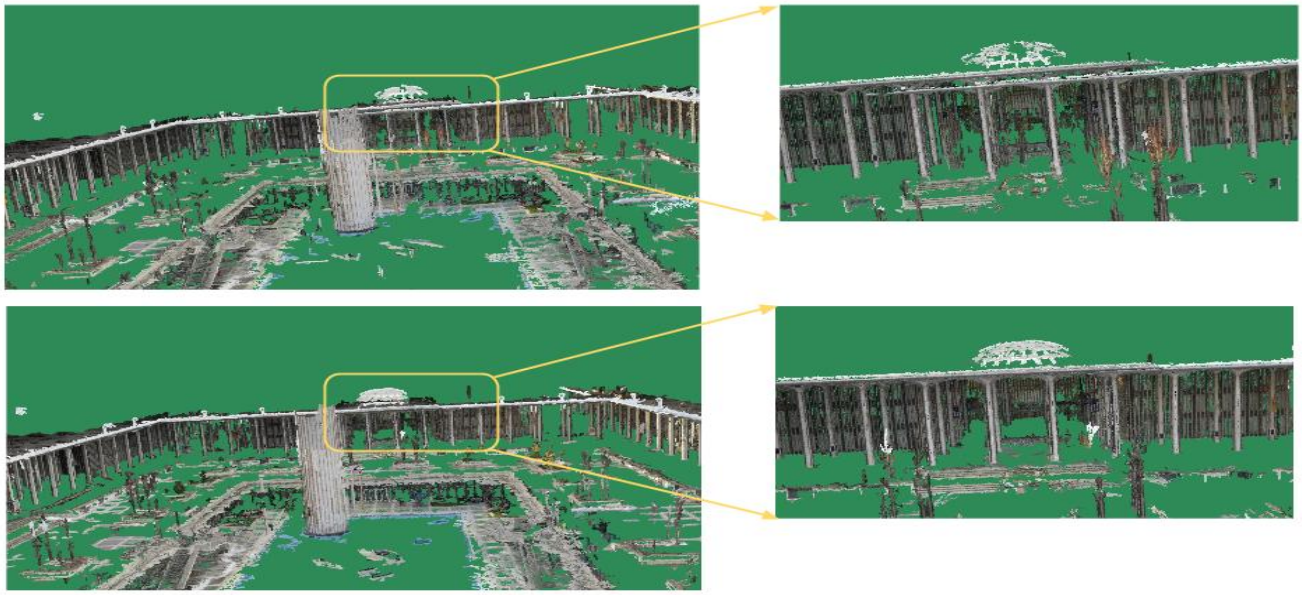


Figure 4. **Loop closure** in the fusion of the Podium dataset following a ring of sub-division sets. (a) Repeated fusion between consecutive pairs from a single iteration of Algorithm 2 might lead to glitch in the loop. (b) Proposed RANSAC Algorithm in Table 2 performs multiple iterations in (a) and the one with smallest loop closure error is chosen in the final fusion.

ing rigid transformation up to a scaling change [24]. K is the number of RANSAC iterations, τ is the threshold separating inliers and outliers. Figure 3 shows an example fusion result.

3.3. Global Loop Closure Using Efficient RANSAC

With rigid transformation estimation of the correspondence from joint feature matching, the alignment between a pair of sub-models are aligned well. However, numerical error will accumulate and become visible at the closing part, a problem known as loop closure (see Figure 4 for an example). To address this problem, we use RANSAC again but globally distribute the errors among all the sequential transformation pair given that each coordinate frames registration error subject to the local RANSAC threshold τ. Our solution successfully closed the loop as shown in Figure 4.

The input of our algorithm is a set of corresponding feature points among all the sub-models in the loop. Suppose we have a single closed loop consisting of M sub-models. The successive two sub-models PC^m and PC^{m+1} are related by their overlap. The rotation matrix, translation vector and scale factor are denoted by $\mathbf{R_m}$, t_m and s_m respectively. Here, $\mathbf{R_M}$, t_M and s_M will transform the point clouds PC^M to the point clouds PC^1. We denote the 3D corresponding points as $\{\mathbf{X^m}\}$, L is the number of the global RANSAC iterations, ϕ is the global threshold between the loop inliers and outliers, τ is the local threshold between the inliers and outliers of adjacent clusters. The output are an ordered set of rotation matrices $(\mathbf{R}_1, \ldots, \mathbf{R}_M)$ and corresponding sets of scalings $(s_1, \ldots, s_M)$ and translations $(\mathbf{t}_1, \ldots, \mathbf{t}_M)$, which determined by RANSAC as the best choices to sequentially transform all the models into a unified coordinates system with loop closure.

4. Experimental Validation

We evaluate the method on two university campus datasets — the **Campus Podium** (1669 images) and **Track Field** (2769 images), where the photos of the sites are captured using standard consumer cameras. Images in both datasets are divided into 8 image sets following a scene-space clustering scheme (Section 3.1). The number of image sets could be chose arbitrarily, provided that the sets are adjacent to at least one neighbor for fusion. Even though the computational time reduces as the number of image sets increases, the number of duplicated anchor images increases too. A sweet spot exists in between the image set division and the duplicated anchor images, which can be data specific. Our experiments suggest that 10% to 15% overlap is sufficient for our algorithm to calculate the fusion alignment robustly and accurately. SfM 3D reconstruction in each image sets is generated using VisualSFM [22]. The stitching and fusion of all sub-models yields the point clouds shown in Figure 1. The result is in correct geometrical shape of the sites. For comparison and validation, we also generate the all-in-one SfM point cloud with brute force image matching without division. The visual examination show no visible difference between the Divide-Recon-Fuse re-

Table 3. *Comparison of computational time for the two university campus datasets.* Each row lists the details of a SfM computational run.

Dataset	Clusters	#Images	#Anchors	Match pairs	Matching time	SfM (BA)
Campus Podium	Cluster 1	201	C1-C2: 14	20100	0.43 hours	55 seconds
	Cluster 2	188	C1-C2: 7	17,578	0.4 hours	45 seconds
	Cluster 3	147	C1-C2: 20	30,576	0.23 hours	35 seconds
	Cluster 4	215	C1-C2: 11	23,005	0.51 hours	42 seconds
	Cluster 5	218	C1-C2: 13	23,653	0.48 hours	57 seconds
	Cluster 6	260	C1-C2: 17	33,670	0.70 hours	80 seconds
	Cluster 7	258	C1-C2:19	33,153	0.73 hours	62 seconds
	Cluster 8	293	C1-C2: 10	42,778	0.9 hours	62 seconds
	Divide-conquer	1780	111	224,513	**4.38 hours**	438 seconds
	All(Brute force)	1,669		139,1946	**26.80 hours**	512 seconds
Track Field	Cluster 1	461	C1-C2: 11	106,030	1.84 hours	191 seconds
	Cluster 2	466	C2-C3: 9	10,8345	2.13 hours	170 seconds
	Cluster 3	415	C3-C4: 10	85,905	1.71 hours	91 seconds
	Cluster 4	359	C4-C5: 10	64,261	1.41 hours	103 seconds
	Cluster 5	290	C5-C6: 10	41,905	0.78 hours	77 seconds
	Cluster 6	276	C6-C7: 10	37,950	0.59 hours	106 seconds
	Cluster 7	272	C7-C8: 10	36,856	0.73 hours	166 seconds
	Cluster 8	311	C8-C1: 11	48,205	0.97 hours	130 seconds
	Divide-conquer	2,850	81	156,550	**10.16 hours**	1,034 seconds
	All(Brute force)	2,769		3,832,296	**67.67 hours**	887 seconds

constructed point clouds and brute force ones. Quantitative comparison results are shown in Table 3, where the number of images in each image sets, the number of anchor images in adjacent clusters, and the number of pairwise matchings in each SfM run are reported. Observe that the pairwise matching time indeed dominates the whole reconstruction process; and in comparison, the SfM Bundle Adjustment (BA) process only takes a small portion of computation time. Experiments were performed on a 3.3 GHz Intel i7-5820K machine.

Our 3D visualization is implemented using the Visualization Toolkit (VTK) [14], which supports real-time 3D rendering. The fused 3D point cloud (with size up to tens of million of points) can be visualized and interactively navigated using a standard PC. Visual examination by zooming-in to the fused region (as shown in Figure 4) suggests that sub-models are well aligned without visually observable gaps or glitches.

Since our method can work on unorganized image sets, it is very flexible on the integration with other SfM methods. The input can be collected at different time and weather. Images can be acquired using arbitrary imaging devices or aerial drones. Our method can be integrated straightforwardly with new 3D reconstruction methods that provide correspondences between the 2D-image-features and 3D-point clouds in the reconstruction. Our divide-conquer and fusion framework can scale up to arbitrarily large sets based on two assumptions: (1) division and reconstruction from scene sections are viable, and (2) fusion via matching the anchor image feature correspondences are viable. For a very large dataset, the final fusion step can be performed out-of-the-core.

5. Conclusion

We presented an efficient large-scale photometric 3D SfM reconstruction pipeline using a divide-conquer and fusion framework. We described a novel approach of adding a few duplicate selection of "anchor images" to adjacent image sets prior to reconstruction, and we showed how to use the anchor images to facilitate effective stitching and 3D model fusion. The proposed method is versatile and highly applicable and integrable with other 3D reconstruction methods.

Future works include a quantitative evaluation comparing our resulting model and the brute-force SfM result, due to the fact that large scale 3D reconstructions normally lack ground truth data. For man-made structures such as the buildings in an urban scene, all-in-one SfM typically does not generate useful results which can serve as the groundtruth, due to the frequent repeated and symmetric patterns and structures. Also, surface mesh generation can improve the quality of the reconstruction of each scene section, thus improving the final fused model.

References

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. In *ICCV*, pages 72–79. IEEE, 2009.

[2] B. Bhowmick, S. Patra, A. Chatterjee, V. M. Govindu, and S. Banerjee. Divide and conquer: Efficient large-scale structure from motion using graph partitioning. In *ACCV*, pages 273–287. Springer, 2014.

[3] M.-C. Chang, N. H. Trinh, B. C. Fleming, and B. B. Kimia. Reliable fusion of knee bone laser scans to establish ground truth for cartilage thickness measurement. In *SPIE Medical Imaging*, pages 76232M–76232M. International Society for Optics and Photonics, 2010.

[4] B. Dalgarno, J. Hedberg, and B. Harper. The contribution of 3D environments to conceptual understanding. 2002.

[5] D. W. Eggert, A. Lorusso, and R. B. Fisher. Estimating 3-d rigid body transformations: a comparison of four major algorithms. *Machine Vision and Applications*, 9(5-6):272–290, 1997.

[6] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, et al. Building rome on a cloudless day. In *ECCV*, pages 368–381. Springer, 2010.

[7] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *PAMI*, 32(8):1362–1376, 2010.

[8] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-view stereo for community photo collections. In *ICCV*, pages 1–8. IEEE, 2007.

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

[10] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *CVPR*, pages 3121–3128. IEEE, 2011.

[11] K. Matzen and N. Snavely. Nyc3dcars: A dataset of 3d vehicles in geographic context. In *ICCV*, pages 761–768, 2013.

[12] P. Musialski, P. Wonka, D. G. Aliaga, M. Wimmer, L. Gool, and W. Purgathofer. A survey of urban reconstruction. In *Computer graphics forum*, volume 32, pages 146–177. Wiley Online Library, 2013.

[13] F. Remondino. Heritage recording and 3D modeling with photogrammetry and 3D scanning. *Remote Sensing*, 3(6):1104–1138, 2011.

[14] W. J. Schroeder, K. M. Martin, and W. E. Lorensen. The design and implementation of an object-oriented toolkit for 3D graphics and visualization. In *Visualization*, pages 93–ff. IEEE Computer Society Press, 1996.

[15] I. Simon, N. Snavely, and S. M. Seitz. Scene summarization for online image collections. In *ICCV*, pages 1–8. IEEE, 2007.

[16] N. Snavely and Y. Furukawa. State of the art 3D reconstruction techniques: Very large scale 3D reconstruction and the role of priors, 2014. CVPR.

[17] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. In *TOG*, volume 25, pages 835–846. ACM, 2006.

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

[19] A. Tremeau and P. Colantoni. Regions adjacency graph applied to color image segmentation. *TIP*, 9:735–744, 2000.

[20] O. Untzelmann, T. Sattler, S. Middelberg, and L. Kobbelt. A scalable collaborative online system for city reconstruction. In *ICCV Workshops*, pages 644–651, 2013.

[21] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, and J. Tardós. A comparison of loop closing techniques in monocular slam. *Robotics and Autonomous Systems*, 57(12):1188–1197, 2009.

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

[23] J. Xiao and Y. Furukawa. Reconstructing the worlds museums. *IJCV*, 110(3):243–258, 2014.

[24] Y. Yang, M.-C. Chang, P. Tu, and S. Lyu. Seeing as it happens: Real time 3D video event visualization. In *ICIP*, pages 2875–2879. IEEE, 2015.