ACCURATE MESH-BASED ALIGNMENT FOR GROUND AND AERIAL MULTI-VIEW STEREO MODELS

Yang Zhou[1,2], *Shuhan Shen*[1,2], *Xiang Gao*[1,2], *Zhanyi Hu*[1,2]

1. NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, P. R. China
2. University of Chinese Academy of Sciences, Beijing 100049, P. R. China

ABSTRACT

We propose a method for accurate alignment of ground and aerial multi-view stereo (MVS) models. We achieve this goal by reconstructing the surface meshes from MVS point clouds generated by aerial and ground images respectively, and then iteratively removing the gap between them. The key issue is how to establish reliable correspondences between two meshes. To address this issue, we introduce a new set called the skeleton facet set (SFS) to represent the locally smooth part on the mesh, and then compute the transformation matrix by comparing the depths of the facets in SFS between aerial and ground models. Experimental results show that the proposed method is able to yield accurate alignment results and is robust to noise as well.

Index Terms— Skeleton Facet Set, Model Alignment, Mesh Correspondence

1. INTRODUCTION

Image based 3D reconstruction has been a hot topic in Computer Vision for decades. Researchers have made tremendous progress in this area, and many of the existing structure-from-motion (SfM) [1, 2, 3] and multi-view stereo (MVS) [4, 5, 6] methods perform well in real world scenes. For large scale outdoor scenes, 3D models are commonly built from aerial images acquired by unmanned aerial vehicles (UAV), which often lack sufficient details. Recently, a growing number of works [7, 8, 9] combine aerial and ground images to yield complete and detailed models. However, due to the wide baselines and the varying imaging conditions between the images from different sources, it is hard to directly match aerial and ground images via image feature descriptors like SIFT [10], which finally leads to a badly aligned model.

To improve the quality of the models generated from combined aerial and ground images, we propose a novel method to align the two MVS models separately reconstructed from different sources. The inputs of our method are two point clouds generated from aerial and ground images respectively, and a list of the calibrated cameras of the aerial images. To filter gross noise and establish reliable correspondences, we respectively reconstruct the surface meshes from the two point clouds and extract the skeleton facet set (SFS) from the aerial mesh, which represents the locally smooth part on a mesh. Each facet in SFS is attached with a reference camera that represents the best view of the facet. We then iteratively compute the transformation matrix between the corresponding points in the SFS facets and the ground mesh, and align these two meshes accordingly. The rest of the paper is organized as follows. In Section 2 we discuss the existing works, and in Section 3 the proposed method is elaborated. In Section 4, we introduce a new evaluation method and evaluate the proposed method qualitatively and quantitatively.

This work was supported by the Natural Science Foundation of China under Grants 61333015, 61421004 and 61473292.

2. RELATED WORK

The main problem in large scale scene modeling from both aerial and ground images is that image features from different sources can barely be matched, due to the wide baselines, quite different viewpoints and the varying imaging conditions between them. One possible solution is to add extra images with viewpoints between aerial and ground images, as in [8]. Such images bridge the separate aerial and ground images, thus yielding more complete and detailed models. However, this is not feasible for other datasets as the ”bridge images” are hard to obtain. Another attempt of solving this matching problem on the image level is proposed in [11]. To align models built from ground images to existing geography system, Shan et al. [11] synthesize aerial views with ground images by depth-based warping, and then match the synthesized views with images retrieved from Google Maps. The accuracy of the alignment method in [11] depends on the quality of the MVS reconstruction results and the efficiency of the depth map interpolation, which makes it sensitive to noise.

To utilize the geometry information in the scene, Wu et al. [12] propose a Viewpoint-Invariant Patches (VIP) feature. A VIP descriptor is extracted from the orthographic projection of the textured local tangential plane into a virtual orthographic camera. The VIP features rely on the dominant planar geometry of the scene. Other works [13, 14] hold the same assumption, which is not suitable for many real world scenes. In this paper, we introduce the skeleton facet set (SFS) based on

the assumption that the objects in the scene is locally smooth, which is a reasonable assumption in real world scenes.

To directly align two 3D models, 3D feature points [15, 16] are extracted either from mesh or point cloud. Rusu et al. [15] propose a 3D feature descriptor for point cloud, called Fast Point Feature Histograms (FPFH). But the K Nearest Neighborhood (KNN) based schema used in FPFH is highly sensitive to the point density, which limits its applicability in MVS point clouds. Zaharescu et al. [16] propose a method to extract mesh feature points which often appear on the edges or the corners of the model. But this is not suitable for our task since our aerial meshes are rough and the edges and the corners are often inaccurate in outdoor scenes. In this paper, we directly align aerial and ground 3D models. The corresponding points are computed in depth maps with facets in SFS, regardless of the roughness of the aerial mesh.

The main contributions of our work are 1) proposing a new pipeline for the alignment of aerial and ground MVS models, and 2) introducing a new way to establish mesh correspondences.

3. MESH-BASED ALIGNMENT

The inputs of our method are two point clouds generated separately from aerial images and ground images, and a list of the calibrated cameras of the aerial images. We assume that the two models has already been approximately aligned (for example, using GPS). To filter gross noise and establish reliable correspondences, we reconstruct the surface meshes from the input point clouds with [17], and extract the skeleton facet set (SFS) from the aerial mesh. A reference aerial camera is attached to each facet in SFS, which represents the best view of the facet. Then, we eliminate the gross gap and align the two meshes by iteratively comparing the corresponding points on the projection of the facets in SFS to the reference cameras. The output of our method is the similarity transformation matrix from the ground model to the aerial model. The pipeline of the proposed method is shown in Fig. 1. In the following subsections, we will detail each part of the proposed method.

3.1. Skeleton Facet Set

We define the skeleton facet set (SFS) as a set of facets which locate at locally smooth part of the mesh and can be viewed by at least one camera with a relatively good angle. For each facet in SFS, as shown in the Fig. 2, the angle between its normal $\vec{N}$ and that of all its adjacent facets $\vec{N}'$ should be smaller than a threshold θ_a; the angle between its normal $\vec{N}$ and the line of sight from the center of its reference camera to the facet center $\overrightarrow{Cf}$ should be greater than a threshold θ_v; and the size of the facet projection ϕ on its reference camera should be greater than a threshold ϕ_s. The reference camera of a facet in SFS is a camera that satisfies ϕ_s-constraint and has highest scores in θ_v-constraint. We observe that the edge- and corner-part of the rough aerial mesh is often inaccurate and unreliable, comparing to the locally smooth part. Therefore, the aerial mesh is represented by SFS in the following process, and the correspondences are computed between SFS and the ground mesh.

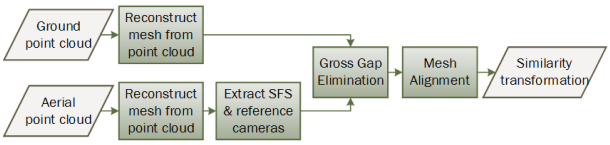

Fig. 1. The pipeline of the proposed method.

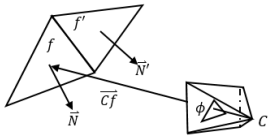

Fig. 2. The constraints of a candidate facet in SFS.

Algorithm 1 Point Pairs Acquisition.

Input: Aerial and ground meshes, SFS and reference cameras.

Output: Depth map pairs (D_a, D_g) and point pairs (P_a, P_g).

1: Pairs of aerial and ground depth maps (D_a, D_g) are obtained by projecting aerial and ground meshes into the reference camera associated with each facet in SFS.
2: For each pixel p_a in the aerial depth map, if it lies in the projection of a facet in SFS, we find its corresponding pixel p_g in the ground depth map which has the same image location as p_a, and back-project p_a and p_g into 3D space to form a 3D point pair (P_a, P_g).
3: Return (D_a, D_g) and (P_a, P_g).

3.2. Gross Gap Elimination

To ease the mesh alignment process , we eliminate the gross gap between the aerial and the ground meshes by iteratively computing the translation between them and translating the ground mesh accordingly. In this process, the correspondences between the aerial and ground meshes are not necessarily reliable, thus, the point pairs are computed using Algorithm 1. In each iteration, we obtain the corresponding point pairs (P_a, P_g) and give each point pair a weight w . If P_g is in a facet facing to a wrong direction, i.e. the angle between the normal of this facet and the optical axis of the camera is smaller than $90°$, we increase w by w_e . Then, the translation between the two meshes is computed as $\vec{T}_g = \frac{\sum_i w_i (P_{a,i} - P_{g,i})}{\sum_i w_i}$, and we translate the ground mesh by $\vec{T}_g$. This iteration is repeated until $|\vec{T}_g|$ is smaller than a threshold T.

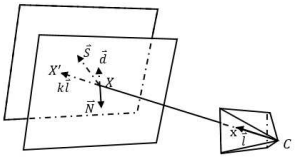

Fig. 3. Sketch map for the geometry in QEDST.

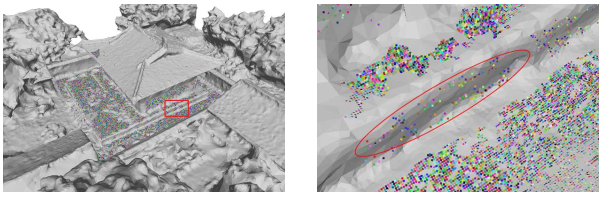

Fig. 4. Left: The temporal alignment result of NanChan Temple dataset. Right: closeup of the rectangle in the left figure. The supportive point pairs on the horizontal facets are much denser than those on the vertical facets.

3.3. Mesh Alignment

After the process of gross gap elimination, we progressively refine the alignment result by computing the similarity transformation between the point pairs and updating the ground mesh accordingly. In each iteration, we acquire the corresponding point pairs (P_a, P_g) using Algorithm 1, and discard a point pair if P_g lies in a facet facing to a wrong direction; then a RANSAC based approach [18, 19] is employed to compute the similarity transformation S between point pairs, which consists of a rotation matrix R, a scale factor s and a translation $\vec{t}$; then we transform the ground mesh by S. This iteration is repeated until S approximates an identity matrix.

However, we find that the above process may not yield accurate result if the supportive point pairs are unevenly distributed. As shown in Fig. 4, supportive point pairs on the horizontal facets are much denser than those on the vertical facets, which may results in that the gap between vertical facets remains when this process ends. To tackle this problem, at the end of each iteration in the above process, if the rotation angle rotation angle $A_R = arccos(|\frac{Tr(R)-1}{2}|)$ and the scale factor s meet the terminate conditions while the translation $\vec{t}$ does not, we define the energy E_t on the gap between the two meshes and further align them by minimizing E_t. E_t is the weighted sum of the distances between all the corresponding facets between SFS and the ground mesh. In the implementation, E_t is approximated to the sum of the distances between the corresponding points and minimized with an iterative approach. In each iteration of the minimizing process of E_t, we obtain the corresponding point pairs (P_a, P_g) using Algorithm 1, and discard a point pair if P_g is in a facet facing to a wrong direction or the angle between the normals of the corresponding facets ($\vec{N}_a$ & $\vec{N}_g$) is greater than a threshold θ_t . Then P_g is shifted to P'_g so that $\overrightarrow{P_a P'_g}$ is parallel to $\vec{N}_a$. Finally, we sort the shift vectors in candidate shift vector set Ω_s ($\Omega_s = \{\vec{S}_i | \vec{S}_i = P_{a,i} - P'_{g,i}, \forall i\}$) by their norms in a decreasing order, and test $\vec{S} \in \Omega_s$ one by one using the Quick Energy Decreasing Shift Test (QEDST; refer to Section 3.4) until we find the first $\vec{S}$ that passes QEDST, denoted by $\vec{S}^\star$, and translate the ground mesh by $\vec{S}^\star$. This iteration is repeated until $|\vec{S}^\star|$ is smaller than a threshold T_s or no $\vec{S} \in \Omega_s$ passes QEDST.

3.4. Quick Energy Decreasing Shift Test

In this section, we proposed the Quick Energy Decreasing Shift Test (QEDST) to judge wether a shift vector is an energy decreasing vector. Given a shift vector of the ground model $\vec{S} \in \Omega_s$, for each pixel x in a ground depth map, its corresponding 3D point is shifted from X to X$'$, as shown in Fig. 3. We denote the vector from X to X$'$ as $k\vec{l}$, in which $\vec{l}$ is the normal vector from camera center C to the 3D point X, and k is the depth increment along the direction $\vec{l}$. Assuming that the shift $\vec{S}$ is small and the local area of X is approximated to a plane with a normal $\vec{N}$, we have

$$|\vec{d}| = |k\vec{l} \cdot \vec{N}| = |\vec{S} \cdot \vec{N}| \tag{1}$$

where $|\vec{d}|$ is the distance between the two planes. Since k is the depth increment of x in the ground depth map, the depth error between the aerial and ground depth maps becomes $D_a(\mathrm{x}) - D_g(\mathrm{x}) - k$ ($D(\mathrm{x})$ is the depth along the viewing direction). Considering that $k\vec{l} \cdot \vec{N}$ and $\vec{S} \cdot \vec{N}$ have the same signs, the energy at x is defined as

$$\begin{aligned} E_{\mathrm{x}}(\vec{S}) &= |(D_a(\mathrm{x}) - D_g(\mathrm{x}) - k)\vec{l} \cdot \vec{N}| \\ &= |(D_a(\mathrm{x}) - D_g(\mathrm{x}))\vec{l} \cdot \vec{N} - \vec{S} \cdot \vec{N}| \end{aligned} \tag{2}$$

Thus, the total energy of the gap between the two meshes is $E_t(\vec{S}) = \sum_{f \in \mathrm{SFS}} \sum_{\mathrm{x} \in \Omega_f} E_{\mathrm{x}}(\vec{S})$. If $E_t(\vec{S}) < E_t(0)$, the shift vector $\vec{S}$ passes QEDST.

4. EVALUATION

4.1. Evaluation Method

In this section, we propose a new evaluation method to quantitatively evaluate the alignment accuracy of aerial and ground models, which takes into consideration the differences of coverage and point density between aerial and ground MVS point clouds. We perform our evaluation on point clouds. The reason is that the distance between points is much easier to compute than distance between triangulated meshes, and the noise and errors introduced in the meshing process could be avoided. In order to make the alignment accuracy evaluation robust to coverage differences and point densities, for each point p in the ground point cloud, we find its nearest point p' in the aerial

Fig. 5. The result of our method and Shan et al. [11] on NanChan Temple dataset (1st row) and FoGuang Temple dataset (2nd row). From left to right: the aerial point cloud, the ground point cloud, the initial status of the alignment, the final results of the alignment and the evaluation of our method (4th column) and [11] (5th column).

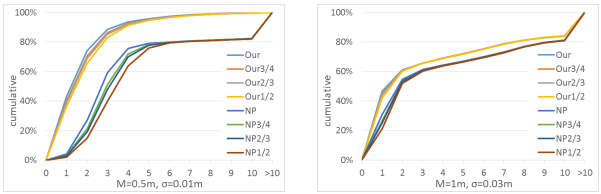

Fig. 6. The performance of our evaluation method (Our) and the nearest point method (NP) on simulation dataset (left) and NanChan Temple dataset (right).

point cloud. Then we discard p as outliers if $|\vec{d}|$ ($\vec{d} = p' - p$) is greater than a threshold M. For the remaining $\vec{d}$, we project each one of them to the normal direction (can be computed with PCL [20]) of p' and store the distance $|\vec{d'}|$. Finally, a Cumulative Error Curve (CEC) is drawn from $|\vec{d'}|$s with a given standard deviation σ.

To verify the effectiveness of the proposed evaluation method with different point densities, we compare the CECs of our evaluation method and the nearest point method in [9, 11] (with the threshold M) on a simulation dataset (detailed in the supplementary material) and our own dataset, NanChan Temple. The result is shown in Fig. 6. To test the robustness of the two evaluation methods, we randomly downsample the aerial point cloud three times, each with a ratio of 3/4, 2/3, and 1/2 respectively. The CECs of our evaluation method vary little. Therefore, our evaluation method is more robust to the point density comparing with the nearest point method.

4.2. Evaluation Results

We compare our method with the state-of-the-art method [11] on two datasets, NanChan Temple and FoGuang Temple. NanChan Temple dataset contains 700 aerial images and 1708 ground images, and FoGuang Temple dataset contains 150 aerial images and 972 ground images. The aerial images have a resolution of 4912×3264 and the ground ones of 5760×3840. To reduce the memory cost, we generate and downsample the dense point clouds from resized images with both width and height are $1/4$ of the original ones. Finally, the amounts of points of the aerial model and the ground model are 7.1M and 7.3M respectively in NanChan Temple dataset, and 7.9M and 7.3M respectively in FoGuang Temple dataset. The point clouds and the results are shown in Fig. 5. The final results of the alignment and the CECs show that our method yields more accurate results than [11]. Besides, in FoGuang Temple, the result shows that our method greatly improves the alignment, while [11] improves little. The reason might be that the noise is heavy due to the occlusion and the large amount of vegetation in this scene. More details about the datasets and the results could be found in the supplementary material.

5. CONCLUSION

In this paper, we propose a novel accurate mesh-based method for aerial and ground MVS models. In contrast to the existing methods that use a once-for-all strategy, our method iteratively removes the gap between the aerial model and the ground model. The reliable correspondences between two models are obtained by projecting the locally smooth part of the aerial mesh and the whole ground mesh to the reference cameras and comparing the depths. Then, the similarity transformation is robustly computed from them. The experimental results demonstrate the accuracy and robustness of our method.

Our method has two limitations. One is that it is unable to align models with large gaps. In such cases, GPS information or other method should be adopted to provide a proper initial status. The other is that the aligning process is time-consuming. It takes about 20 iterations in total to align two MVS models, depending on the parameters. We will investigate these problems in our future work.

6. REFERENCES

[1] Hainan Cui, Shuhan Shen, Wei Gao, and Zhanyi Hu, "Fusion of auxiliary imaging information for robust, scalable and fast 3d reconstruction," in *ACCV 2014*, pp. 227–242. Springer, 2015.

[2] Hainan Cui, Shuhan Shen, Zhanyi Hu, et al., "Efficient large-scale structure from motion by fusing auxiliary imaging information," *Transactions on Image Processing (TIP)*, vol. 22, pp. 3561–3573, 2015.

[3] Tianwei Shen, Siyu Zhu, Tian Fang, Runze Zhang, and Long Quan, "Graph-based consistent matching for structure-from-motion," in *ECCV 2016*. Springer, pp. 139–155.

[4] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *ECCV 2016*. pp. 501–518, IEEE.

[5] Alex Locher, Michal Perdoch, and Luc Van Gool, "Progressive prioritized multi-view stereo," in *CVPR 2016*. pp. 3244–3252, IEEE.

[6] Fabian Langguth, Kalyan Sunkavalli, Sunil Hadap, and Michael Goesele, "Shading-aware multi-view stereo," in *ECCV 2016*. Springer, pp. 469–485.

[7] Alexandru N Vasile, Luke J Skelly, Karl Ni, Richard Heinrichs, and Octavia Camps, "Efficient city-sized 3d reconstruction from ultra-high resolution aerial and ground video imagery," in *International Symposium on Visual Computing*. Springer, 2011, pp. 347–358.

[8] Qi Shan, Riley Adams, Brian Curless, Yasutaka Furukawa, and Steven M Seitz, "The visual turing test for scene reconstruction," in *2013 International Conference on 3D Vision-3DV 2013*. IEEE, 2013, pp. 25–32.

[9] András Bódis-Szomorú, Hayko Riemenschneider, and Luc Van Gool, "Efficient volumetric fusion of airborne and street-side data for urban reconstruction," *arXiv preprint arXiv:1609.01345*, 2016.

[10] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91—110, 2004.

[11] Qi Shan, Changchang Wu, Brian Curless, Yasutaka Furukawa, Carlos Hernandez, and Steven M Seitz, "Accurate geo-registration by ground-to-aerial image matching," in *2014 2nd International Conference on 3D Vision*. IEEE, 2014, vol. 1, pp. 525–532.

[12] Changchang Wu, Brian Clipp, Xiaowei Li, Jan-Michael Frahm, and Marc Pollefeys, "3d model matching with viewpoint-invariant patches (vip)," pp. 1–8, 2008.

[13] Jean-Michel Morel and Guoshen Yu, "Asift: A new framework for fully affine invariant image comparison," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 438–469, 2009.

[14] Mayank Bansal, Kostas Daniilidis, and Harpreet Sawhney, "Ultra-wide baseline facade matching for geo-localization," in *European Conference on Computer Vision*. Springer, 2012, pp. 175–186.

[15] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, 2009, pp. 3212–3217.

[16] Andrei Zaharescu, Edmond Boyer, Kiran Varanasi, and Radu Horaud, "Surface feature detection and description with applications to mesh matching," pp. 373–380, 2009.

[17] Michal Jancosek and Tomas Pajdla, "Exploiting visibility information in surface reconstruction to preserve weakly supported surfaces," *International Scholarly Research Notices*, vol. 2014, 2014.

[18] Martin A Fischler and Robert C Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[19] Shinji Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, no. 4, pp. 376–380, 1991.

[20] "Point cloud library," `http://pointclouds.org/`.