Outdoor Mapping Framework: from Images to 3D Model

Sergiy Skuratovskyi, Ievgen Gorovyi, Vitalii Vovk and Dmytro Sharapov
It-Jim
13 Chernyshevska str., Kharkiv, 61045, Ukraine
ssnake@it-jim.com, ceo@it-jim.com, vitalii.vovk@it-jim.com, sharapov@it-jim.com

***Abstract* — 3D mapping techniques have a large variety of applications from entertainment to military and medical fields. However, there is a big challenge of obtaining well refined 3D model from a set of images without usage of depth sensors. In the paper, we analyze main components of 3D reconstruction pipeline allowing to get detailed models of outdoor objects from drones. In particular, we experiment with algorithms required for structure from motion and point cloud densification. It is demonstrated, that proper local feature extraction, matching and verification directly effect on a final model quality. Analysis of two existing 3D reconstruction frameworks (MVE and COLMAP) is conducted. Initial experimental results are shown.**

Keywords— structure from motion, multiple-view stereo, local features, features matching, point cloud.

I. INTRODUCTION

3D mapping is the process of reconstructing 3D surface appearance and structure from a set of acquired two-dimensional images [1]. The input data for 3D reconstruction process is typically acquired from various sensors. For instance, RGB-D (red-green-blue and depth) which is used in Microsoft Kinect [2]. It is applied for extraction the depth from both static and dynamic scenes in indoor [3] and outdoor environments [4]. Another example of depth sensor is Occipital's structure sensor intended for mobile devices [5].

Another type of devices specially designed for capturing depth information are time-of-flight (ToF) cameras. Comparison of RGB-D and ToF cameras may be found in [6]. Fusion of ToF and RGB cameras is implemented in Microsoft Kinect v2 [7]. ToF cameras are often used in medicine [8], robotics [9], etc.

One of the challenges in 3D mapping is to obtain well-refined model using only cheap monocular camera without additional equipment like depth sensors or active transmitter. This means that there is no initial information about depth [10] provided. Despite of high complexity, 3D reconstruction from a sequence of optical images has extremely wide range of applications [1] mainly due to low cost of such systems. Moreover, such algorithms are successfully applied in both indoor [11] and outdoor scenes [12]. It is even possible to run 3D reconstruction directly on a mobile device [13], [14].

In this paper, we analyze important steps required for computation of 3D model of real objects in outdoor scenes. We consider structure-from-motion (SfM) technique, point cloud densification step, and additional procedures applied for full 3D imaging. We demonstrate strong effect of local feature extraction and matching schemes not only on initial sparse point cloud content, but also on final 3D model quality. Two popular frameworks (MVE [15] and COLMAP [16]) are comprehensively analyzed for this purpose.

The rest of the paper is organized as follows. In Section II, we explain main principles of sparse point cloud construction (SfM) and following densification step. Section III contains experimental results and their discussion.

II. 3D MAPPING PRINCIPLES

Here we focus on offline 3D reconstruction as a problem of obtaining of 3D model from unordered sequence of 2D images acquired by drones.

A. Feature Extraction, Matching and Geometric Verification

A high-level scheme of 3D reconstruction is shown in Fig. 1. Let us discuss it step-by-step.

In order to match multiple images, various local features are utilized. Usage of keypoint descriptors is a common practice in computer vision and, in particular, in SfM problems. Such algorithms as scale-invariant feature transform (SIFT) [17] and speeded-up robust features (SURF) [18] are good tools of choice due to high robustness to extremal viewing angles and varying illumination.

After feature extraction step, each image is represented by a set of local feature descriptors. Such vectors are matched in order to determine their similarity (Fig. 1). The simplest and the most complete matching approach is exhaustive or "brute force" strategy. In this case, all keypoint descriptors are compared one-by-one. This scheme provides good results, but it is extremely slow. There are two general ways of making it faster: implement matching procedure using hardware optimization or to propose alternative efficient matching schemes.

978-1-7281-1714-0/19/$31.00 ©2019 IEEE

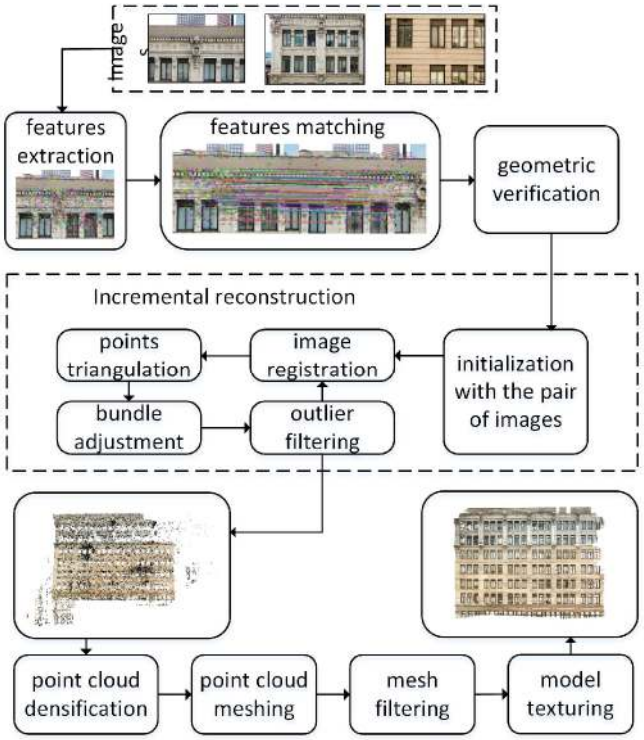

Fig. 1. 3D mapping pipeline.

The strategy of matching in MVE includes matching a small number of features extracted from low-resolution images, forming so-called pre-bundle with image candidates. Then full-resolution matching is accomplished. COLMAP is designed to process large datasets and provides several several matching strategies. Vocabulary tree matching uses pre-trained vocabulary for determining visually similar images to the current one, thus "brute force" feature matching is performed only for a limited sub-set. Three pre-trained vocabulary trees for different sizes of input dataset (1000, 10000, and 100000 images) are available, however custom vocabulary creation is supported as well. Another approach is sequential matching. It is assumes that a set of images is ordered, and consecutive frames are overlapped, thus, there is a higher probability of visual similarity. The idea of loop detection is implemented by matching every N-th image to the candidates from the vocabulary tree. Spatial matching utilizes global positioning system (GPS) information in order to match only images obtained from spatially close locations. Transitive matching improves already existing matching graph by matching images, connected through the third one.

Importantly, that matched image pairs are passed through the geometric verification (Fig. 2). Different transformations are available: homography, essential or fundamental matrix. If a valid transformation maps sufficient number of points from one image to another with acceptable reprojection error, the image pair is considered as geometrically verified.

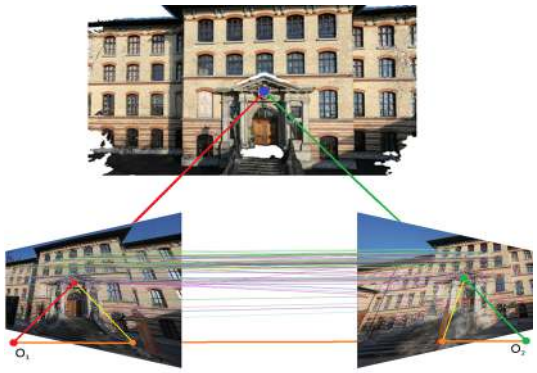

Fig. 2. Image matching and geometric verification (schematically).

B. *Incremental Sparse Point Cloud Reconstruction*

Both MVE and COLMAP implement the same scheme of sparse point cloud reconstruction called incremental SfM. It starts from determining an initial image pair. It is a crucial step, as the reconstruction may not recover from bad initialization. The pair can be chosen automatically considering the results of geometric verification, the angle between viewing directions and the position of the pair in dense part of the correspondences graph. However, there is an option of custom initialization as well. Having this pair of images, initial 3D points are triangulated (two-view reconstruction). After forming initial point cloud, the main loop is started. It includes four steps (Fig. 1).

Image registration represents adding a new image to the existing bundle. The optimal selection of an image to be added is based on specially developed heuristic algorithms [16]. They are based on 2D-3D point correspondences and the relations with already registered images. Thus, initial camera pose is estimated. Then new the point cloud is extended by 3D points triangulation.

It is important to emphasize, that bundle adjustment (BA) is mandatory step. There are two types of commonly applied BA algorithms. Local BA starts after each registration and affects only 3D points corresponding to the 2D points array from the newly added image. Also filtering procedure is performed after BA. It is based either on triangulation angle and the position in front of the camera (MVE), or additionally uses reprojection error value as an indicator (COLMAP).

Global BA starts after registering a certain number of images (10% of all for example) or adding a certain amount of 3D points into the point cloud. Importantly, that both camera intrinsic and extrinsic parameters are refined as well within the same pipeline.

C. *Point Cloud Densification*

Point cloud obtained after SfM is quite sparse and not enough to build the mesh from it. Densification techniques of extending it are a bit different in MVE and COLMAP. MVE uses estimation of depth and normals for 3D points to apply region growing procedure [19]. The information about depth and normals is used as an initial approximation in the point's

neighborhood. An optimization is accomplished through the matching with correspondent views. In COLMAP, authors solve the problem of estimating the depth map of a reference image with given M source images with known homography from reference to the source. Generalized expectation maximization algorithm (GEM) with estimating the visibility distribution and patch-match scheme is applied. COLMAP also supports densification of several sparse maps simultaneously with subsequent fusion of outputs.

Dense point cloud is the input data for meshing procedure. One of the most popular ideas here is Poisson surface reconstruction [20]. Another good idea is combining the surface with different scale of mesh depending on the details existing in particular regions [21]. Additional mesh filtering is an important step, which removes geometrically inconsistent planes. Interested reader may find more details in references.

III. Experimental Results

The datasets for 3D reconstruction are quite different depending on the purpose and creation principles. Eth3D provides a number of open indoor and outdoor datasets, including raw and undistorted images, and benchmark sparse point cloud data [22]. However, each of them contain less than 100 images, which is not enough for demonstration of some specific cases in sparse point cloud reconstruction. Still, we used the "facade" dataset containing 76 images for the initial comparison of the techniques (Fig. 3a).

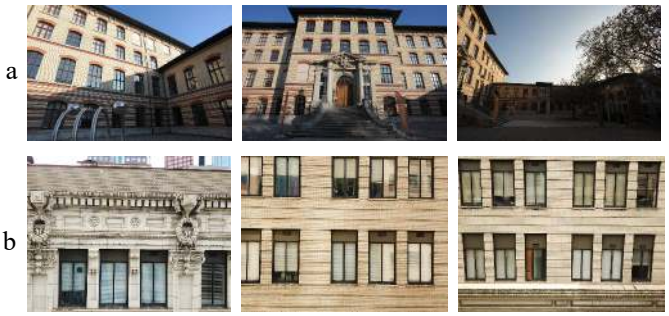

Fig. 3. Images from Eth3D facade dataset (a), and building facade dataset (b).

Another dataset we used is a custom one and acquired with a drone camera. It also represents building facade and contains 256 images (Fig. 3b).

We ran both MVE and COLMAP frameworks on CPU Intel Core I5-7600K, 2.8 GHz with 16 Gb RAM and GPU GTX 750 Ti 4Gb. MVE was run on CPU only, while COLMAP's feature extraction, matching, and MVS required GPU computations.

We obtained quite similar results on Eath3D's facade dataset (Fig. 4). Though this dataset has only 76 images, it provides good coverage of the courtyard of the building and strong relations between images. Thus, we obtained quite complete sparse point cloud (Fig. 4, up), corresponding dense point cloud (Fig. 4, central) and good object shape (Fig. 5, down).

a b

Fig. 4. 3D reconstruction results of Eth3D facade dataset by MVE (a) and COLMAP (b): sparse point cloud, dense point cloud, 3D model (from up to down).

Our dataset of building facade is much larger. The facade was completely covered during acquisition. However, in some regions the relation between images is not consistent. The repetitive pattern on images is created here not only by small bricks, but also by large windows (see Fig. 3 for example), and, unlike Eth3D façade, there are only few of good details on some images. This leads to the errors of MVE reconstruction (Fig. 5). There is only small part of the building represented by the sparse cloud (Fig. 5, left). An explanation here is a bad selection of image pair for incremental SfM init. Some images were registered, but then the relation with others was lost. Thus, no camera parameters were estimated for the most of images. As a result, we have an incomplete model with some artifacts on the top (Fig. 5, right).

Fig. 5. MVE 3D reconstruction results for building facade: sparse point cloud and the resulting model.

As MVE and COLMAP share the same SfM scheme, COLMAP has similar reconstruction drawbacks. However, it contains two features for handling the above challenge. Firstly, it is modified algorithm for initial image pair selection [16]. And, most importantly, COLMAP supports reconstruction of several sparse point clouds (Fig. 6a demonstrates some examples, in general there are 14 models). These models are fused within sparse point cloud construction in the case of hierarchical scheme of models relation usage. Alternatively, the fusion is performed after point cloud densification.

Fig. 6b (left and central images) demonstrates the importance of the prior information for the model quality. Model on the left image was obtained, allowing the algorithm to refine camera parameters for each image separately. For the model from the central image, we set the condition that the same camera was used to acquire all images. Both models are

complete, but for the left one model fusion was failed, while for the right one fusion was successful.

Right image in Fig. 6b represents the model obtained under the same conditions, as the model in central image. For the right model, spatial feature matching strategy was used, while the left and the central models were obtained with exhaustive matching. Spatial matching for this dataset is 4 times faster that exhaustive matching. But it leads to missing some matches, which results in less complete model. Hence, one should carefully use the feature extraction parameters, matching modes and reconstruction pipeline for getting acceptable 3D model quality.

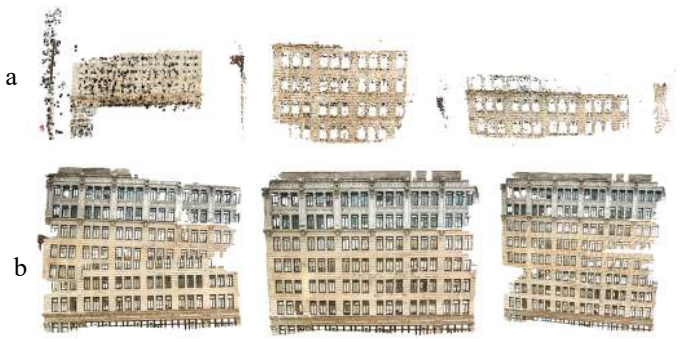

Fig. 6. COLMAP 3D reconstruction: sparse point clouds (a) and reconstruction results (b)

CONCLUSIONS

3D reconstruction from a set of optical images is quite a challenging problem. We performed comparative analysis of two popular frameworks for this purpose. In-depth analysis of feature extraction and matching steps was performed. Technical requirements in terms of system parameters were determined during experiments. Quality of sparse point cloud, dense point cloud and final 3D model were compared and discussed. It was shown, that a proper 3D mapping system configuration provides an acceptable model quality. In the near future, we are planning to deploy a fully custom 3D pipeline containing the benefits of each of discussed framework.

REFERENCES

[1] Z. Ma, S. Liu, "A review of 3D reconstruction techniques in civil engineering and their applications," Advanced Engineering Informatics, vol. 37, 2018, pp. 163-174, doi: 10.1016/j.aei.2018.05.005.

[2] L. Cruz, D. Lucio, L. Velho, "Kinect and RGBD images: challenges and applications," Conference on Graphics, Patterns and Image Tutorials (SIBGRAPI), 2012, No. 13193529, doi: 10.1109/SIBGRAPI-T.2012.13.

[3] P. Henry, M. Krainin, E. Herbst, X. Ren, D. Fox, "RGB-D mapping: using Kinnect-style depth cameras for dense 3D modeling of indoor environment," The International Journal of Robotics Research, vol. 31, No. 5, 2012, pp. 1-17, doi: 10.1177/0278364911434148.

[4] S. M. Abbas, A. Muhammad, "Outdoor RGB-D SLAM performance in slow mine detection," 7th German Conference on Robotics (ROBOTIK), 2012.

[5] M. Kalantari, M. Nechifor, "Accuracy and utility of the structure sensor for collecting 3D indoor information," Geo-spatial Information Science, vol. 19, No. 3, 2016, pp. 202-209, doi: 10.1080/10095020.2016.1235817.

[6] G. Alenya, S. Foix, C. Torras, "Using of ToF and RGBD cameras for 3D robot perception and manipulation in human environment," Intelligent Service Robotics, vol. 7, iss. 4, 2014, pp. 211-220, doi: 10.1007/s11370-014-0159-5.

[7] S. Zennaro, M. Munaro, S. Milani, P. Zanuttig, A. Bernardi, S. Ghidoni, E. Menegatti, "Performance evaluation of 1st and 2nd generation Kinect for multimedia applications," IEEE International Conference on Multimedia and Expo, 2015, No. 15346920, doi: 10.1109/ICME.2015.7177380.

[8] S. Haase, C. Forman, T. Kilgus, R. Bammer, L. Maier-Hein, J. Hornegger, "ToF/RGB fusion for augmented 3D endoscopy using a fully automatic calibration scheme," Informatic Actuell, Bildverarbeitung fur die Medizin, 2012, pp. 111-116, doi: 10.1007/978-3-642-28502-8_21.

[9] Z. Tasneem, D. Wang, H. Xie, S. J. Koppal, "Directionally controlled time-of-flight ranging for mobile sensing platforms," Robotics: Science and Systems, 2018, doi: 10.15607/rss.2018.xiv.011.

[10] F. Li, E. Li, M. J. Shafiee, A. Wong, J. Zelek, "Dense depth map reconstruction from sparse measurements using a multilayer conditional random field model," Conference on Computer and Robot Vision, 2015, No. 15323600, doi: 10.1109/CRV.2015.20.

[11] A. Arnaud, J. Christophe, M. Gouiffes, M. Ammi, "3D Reconstruction of indoor building environments with new generation of tablets," ACM Conference on Virtual Reality Software and Technology, 2016, pp. 187 – 190, doi: 10.1145/2993369.2993403.

[12] K. Tuite, N. Snavely, D.-Y. Hsiao, A. M. Smith, Z. Popovic, "Reconstructing the world in 3D: bringing games with a purpose outdoors," International Conference on the Foundations of Digital Games, 2010, pp. 232-239, doi: 10.1145/1822348.1822379.

[13] E. Nocerino, F. Poiesi, A. Locher, Y. T. Tefera, F. Romondino, P. Chippendale, L. Van Gool, "3D reconstruction with a collaborative approach based on smartphones and a cloud-based server," ISPRS International Workshop on LowCost 3D – Sensors, Algorithms, Applications, vol. XLII-2(W8), 2017, pp. 187-194, doi: 10.5194/isprs-archives-XLII-2-W8-187-2017.

[14] P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Saurer, M. Pollefeys, "Live metric 3D reconstruction on mobile phones," IEEE International Conference on Computer Vision (ICCV), 2013, No. 14144910, doi: 10.1109/ICCV.2013.15.

[15] S. Fuhrmann, F. Langguth, M. Goesele, "MVE – a multi-view reconstruction environment," Eurographics Workshop on Graphics and Cultural Heritage, 2014, doi: 10.2312/gch.20141299.

[16] J. L. Schonberger, J.-M. Frahm, "Structure-from-motion revisited," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, No. 16526742, doi: 10.1109/CVPR.2016.445.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, No. 2, 2004, pp. 91-118, doi: 10.1023/B:VISI.0000029664.99615.94.

[18] H. Bay, T. Tuytellars, L. Van Gool, "SURF: Speeded up robust features," European Conference on Computer Vision (ECCV), 2006, pp. 404-417, doi: 10.1007/11744023_32.

[19] M. Goesele, N. Snavely, B. Curless, H. Hoppe, S. M. Seitz, "Multi-view stereo for community photo collections," IEEE International Conference on Computer Vision, 2007, No. 9848978, doi: 10.1109/ICCV.2007.4408933

[20] M. Kazhdan, M. Bolitho, H. Hoppe, "Poisson Surface Reconstruction," Eurographics Symposium on Geometry Processing, 2006, pp. 61-70, doi: 10.2312/SGP/SGP06/061-070.

[21] P. Mucke, R. Klowsky, M. Goesele, "Surface reconstruction from multiple-resolution sample points," Vision, Modeling and Visualization, 2011, pp. 105-112, doi: 10.2312/PE/VMV/VMV11/105-112.

[22] https://www.eth3d.net/datasets