3D RECONSTRUCTION OF UNSTABLE UNDERWATER ENVIRONMENT WITH SfM USING SLAM

Ilseo Jeon[1], Impyeong Lee [1, *]

[1] Dept. of Geoinformatics, University of Seoul, Seoul, Republic of Korea – (uosKR14, iplee) @uos.ac.kr

Commission II, WG II/9

KEY WORDS: Visual SLAM, SfM, Underwater SLAM, Underwater Photogrammetry, 3D Reconstruction

ABSTRACT:

The underwater environment has substantial properties for underwater research such as marine archaeology, monitoring coral reefs, and shipwrecks. SfM, as a major step of photogrammetry, has been widely used in the field. For a high 3D construction quality, images must have a clear visual sight environment and known orientations of the images. However, underwater images have various types of visual disturbances, but also GPS/INS, commonly used on the ground, are not accepted. Finding more feature points or using more images for SfM are solutions to the problems. However, these methods take high computational costs. An alternative to this problem is to provide the known orientations of the images. For a solution to provide known orientations of images, the presented method in this study uses visual SLAM that processes the localization of a vehicle system and mapping of surroundings. The experiment aims to verify whether SLAM improves the quality of underwater 3D reconstruction and the computation efficiency of SfM. We examine the two aqualoc datasets with the results of the number of cloud points, SfM processing time, and the number of matched images/total images and mean reprojection errors. The outcome shows SLAM-determined orientations improved the quality of 3D reconstruction and the computation efficiency of SfM with results of the increased number of point clouds and the decreased processing time.

1. INTRODUCTION

The underwater environment has substantial properties for underwater researches such as marine archaeology, monitoring coral reefs, and shipwrecks. In the second half of the twentieth century, AUVs (Autonomous Underwater Vehicles), USBL (Ultra-short Baseline), DVL (Doppler Velocity Log) are widely used for underwater researches, and they commonly exploited acoustic sensors such as sonar. Due to their disadvantage of being expensive and large (), alternatives such as ROV (Remotely Operated Vehicle) with visual cameras, which are flexible, have been utilized (Teague & Scott, 2017). Additionally, underwater researches using cameras have often involved SfM (Structure from Motion) to reconstruct the underwater environment, which is a major step in the photogrammetric method of 3D reconstruction.

To get a good quality of 3D reconstruction from SfM, the images should contain a stable surface, visually distinct objects, uniform illumination, and known intrinsic and extrinsic parameters (Dellaert et al, 2000). However, unlike the environment on the ground, GPS (Global Positioning System) for absolute locations and INS (Inertial Navigation System) for computing orientation devices are difficult to be used in the underwater environment. Besides, difficult visual challenges that are not encountered on the ground are in the water. Backscattering that changes the direction of the light ray, turbidity from the cloudiness or haziness caused by large numbers of particles and dynamism from the motion of water are commonly shown in underwater images (Ferrera et al, 2019a). These obstructions can be major hindrances to 3D construction.

Features mean interest points in an image that can be identified even if the location and scale of the object, the photometric state changes. By corresponding features between images, the location of images can be defined with geometry transformations in SfM. Features in images, in other words, are the dominant part of 3D reconstruction of SfM. Thus, in underwater images having visual disturbances, features are not guaranteed. To overcome visual disturbances for SfM, more images to increase the image overlaps or setting a higher keypoint limit (relevant to finding more feature points) for each image can solve problems. However, this increases the time complexity of SfM.

Underwater SLAM (Simultaneous Localization and Mapping) using visual cameras has also been applied contribute to underwater researches, commonly in robotics (Hidalgo and Bräunl, 2015). SLAM is an approach to localize a system with sensors that perceive unknown surroundings and to map the surroundings at the same time. SLAM can be divided according to the sensors used. Visual SLAM uses a visual camera, and its method to estimate camera poses and map the surrounding environment resembles the method of SfM. However, since it targets a real-time operation, SLAM calculates a camera frame trajectory from frames of a video; for example, if a video has 20 frames per second, 20 positions are determined for a second. The positions from SLAM are frequently computed and fast, compared to SfM.

A presented method in this paper aims to reconstruct an underwater environment by combining SfM with SLAM-determined orientation. The experiment here is to verify whether the known orientation from SLAM improves the quality of underwater 3D reconstruction and the computation efficiency of SfM. We inspect the two aqualoc sequences including visual disturbances to compare the results; i.e. the number of cloud points, processing time, mean reprojection error, and dense models depending on the setting (number of keypoint limit, number of images, and SfM-determined orientation) of SfM.

* Corresponding author

2. METHODOLOGY

2.1 Overview of SfM with SLAM-determined orientation

In the first stage of the SfM with SLAM-determined orientation, the SLAM extracts a frame trajectory from a sequence of frames in an underwater environment. In the second stage, the frame trajectory synchronizes with input images for SfM. At the last stage, SfM generates dense point clouds with SLAM-determined orientation. The described SfM with SLAM-determined orientation is as follows (Figure 1).

Both SLAM and SfM are carried out, in a way that they correspond to features, based on them, and determine the map and pose of images. In the next section, SLAM and SfM are shown from the perspective of how the precision differs in corresponding features and determining the pose of images. Feature correspondence means a problem of finding two corresponding points in two images, where features should remain unchanged in the movement, scale, rotation, and light of the object. Depending on the type of algorithm, the degree of matching features may vary, and the better the performance, the more time it takes. Feature correspondence is solved through the three stages of detection, description and matching feature points; 1) the points likely to be feature points are detected for each image, 2) and the invariant descriptors for the various variations (movement, rotation, intensity, etc.) of each feature point are generated, 3) and the similarity or distance between the feature points of the two images is measured and matched. ORB (Rublee et al., 2011) is widely known for being fast and robust in change as fast feature correspondences are the main concern in SLAM. On the one hand, SIFT in SfM is an algorithm that extracts and matches many features which are robust in any change than ORB, but it takes more time.

Based on feature correspondences, SLAM and SfM determine the camera poses. If the complete orientation of two cameras and the perfect intersection of two rays from two images to a point in 3D are available, the projection centers from the two cameras and the point are dependent on one plane in 3D, which is called coplanarity constraint. The camera poses are calculated with the coplanarity constraint; 1) the computation based on the coplanarity constraint is achieved with the essential matrix (for a calibrated camera) or the fundamental matrix (for an uncalibrated camera). 2) Once the adjacent camera poses are established, the location of the point in 3D is calculated with triangulation. In this process, errors are leaded because rays from an image onto a point do not meet when it projects back into the image. 3) Remaining uncertainties from camera poses and points in 3D should be corrected for 3D reconstruction with higher precision. SLAM integrates pose-graph optimization and local BA (Bundle Adjustment), to make local corrections quickly and efficiently. SLAM and SfM use BA to be refined globally. SfM has a more complex computational process than SLAM because it can use (control points if possible) more images and observations (see Section 2.3.3).

2.2 Visual SLAM

Visual SLAM is a solution to determine the pose of sequence images taken on the vehicle system while simultaneously mapping the surrounding environment. Visual SLAM constitutes three steps; 1) corresponding features between images, 2) estimating the camera poses based on feature correspondences, and 3) optimizing the poses globally, all steps of which are driven almost in parallel (Figure 1).

2.2.1 Feature correspondence: Since the camera poses are determined based on the matching feature points, the correspondence problem has a significant effect on the accuracy of the position. The method used to solve the problem of feature correspondences, called a feature-based method, involves feature detection, feature description, and feature matching. One of the feature-based algorithms, commonly used in SLAM, is ORB, which is used in ORBSLAM2 (Mur-Artal et al, 2017) and OpenVSLAM (Sumikura et al, 2019). ORB combines oFAST (FAST Keypoint Orientation) that detects only peripheral pixels, which are quickly considered feature points, and an rBRIEF (Rotation-Aware Binary Robust Independent Element Features), which uses a binary detector to describe the direction of features. In this way, ORB was constructed in a similar way to SIFT, reducing computational costs by using faster and robust methods for scale and orientation in change. It takes 15.3 ms for ORB and 5228.7 ms for SIFT when extracting about 1,000 feature points from the image.

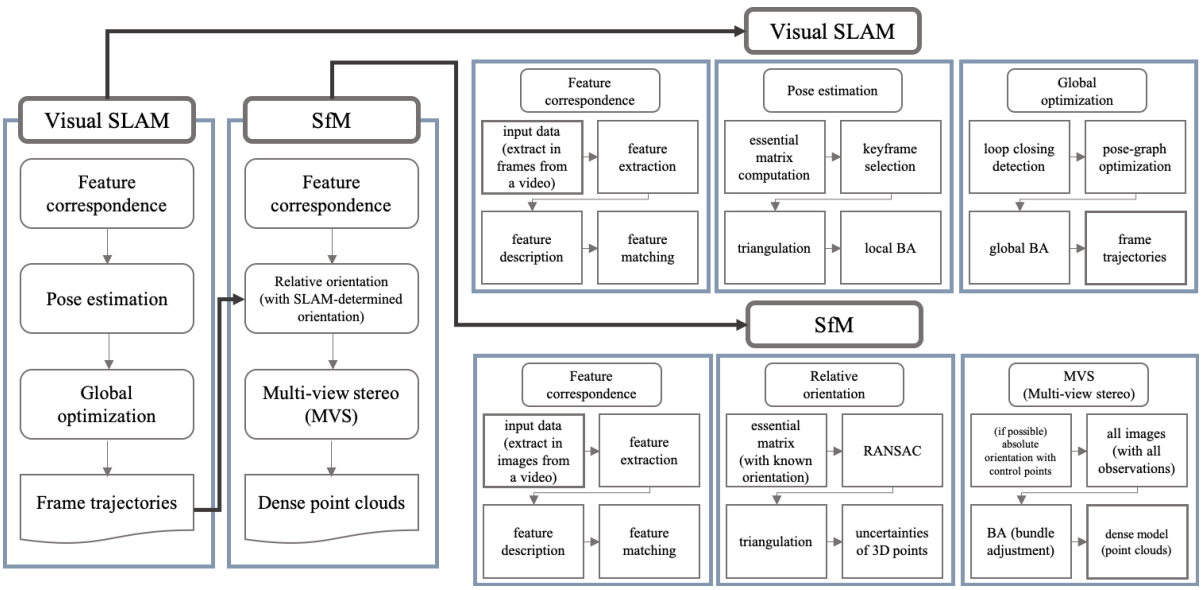

Figure 1. SfM with SLAM-determined orientation

2.2.2 Pose estimation: Once the feature points are matched between frames in the process of feature correspondences, the two adjacent camera poses can be estimated based on the matched features. For a calibrated camera, the essential or fundamental matrix indicating the relationship (R; rotation, t; transition) between two 2D images (I_{k-1}, I_k) is calculated. In the process of estimating the location of a point (assumed as a feature) in 3D by triangulation, the relative scale can be computed from the two adjacent image pairs. This process may result in drift errors (motion errors) and image reprojection errors because the line passing through a point on one image is not the same as the back-projection line. By minimizing the reprojection error of triangulated points, the proper scale set, and the camera pose can be determined. To solve this problem, keyframes that reduce the reprojection error is often used. After the keyframe selection and triangulation, errors still remain. Local bundle adjustment (local BA), one of the nonlinear optimization techniques, for the last acquired frames is widely used. Here, BA aims to minimize reprojection errors. Through the process, local drift errors of camera poses are corrected.

2.2.3 Global optimization: Global optimization is a step in which the entire camera pose wants to remain optimal based on long-term camera positioning. The remaining noises (errors) still affect camera pose decisions. To explain the uncertainties of camera poses from the previous estimation, there are linear techniques such as Kalman Filter, Particle Filter, and non-linear techniques such as pose-graph and BA. A method based on pose-graph and BA is often used in SLAM to maintain the optimality of camera poses while maintaining the efficiency of the operation time. Loop closing detection, which detects the starting point of a loop at the last point of the loop, and keyframe are elements for the optimization. After pose-graph optimization, global BA can be applied in the gross. The final results of SLAM are a map of the surrounding environment and a frame trajectory. In this study, information about camera poses (frame trajectories) is used for SfM.

2.3 Structure from Motion

SfM is a process of 3D reconstruction of objects from the motion of the camera. In particular, SfM, as an early stage in the photogrammetric process, is a major step in determining the reliability of 3D reconstruction. SfM takes a substantial amount of time for feature correspondence because it aims to match more features, for high-precision mapping, than visual SLAM that solves a real-time problem. The process in SfM, similarly in SLAM, consists of 1) feature correspondences, 2) relative orientation of images and 3) multi-view stereo to cover the whole surface of the objects and make higher precision of 3D reconstruction (Figure 1).

2.3.1 Feature correspondence: SIFT is a key technique to match features in SfM. In visual SLAM, feature detection and description were operated by separate algorithms, but SIFT has both detection and description parts. In particular, for scale, SIFT solves the problem by building a pyramid structure that includes multi-scale images. SIFT descriptor contains a solution to the problem of rotated images. Before setting the feature description, the algorithm obtains a gradient histogram of the pixel set. In this histogram, it finds the direction with the greatest value as the dominant orientation. One way to solve the feature correspondence problem is to check all features as much as possible. Another way is to use as many images for SfM as possible, given logistical constraints, which is highly recommended as this optimizes the ultimate number of keypoint matches and system redundancy (Westoby, et al., 2012). However, this approach increases computational complexity $O(N_I^2 N_{F_i}^2)$, where N_I is the number of images and N_{F_i} is the number of features (Schönberger, et al., 2016).

2.3.2 Relative orientation: Once the features are matched between images, a mathematical solution can determine at which orientation the images were taken. Relative orientation starts with calculating the essential or fundamental matrix to determine the geometric relationship between the two cameras. Typically, 8-point, 7-point, and 5-point algorithms are known as solutions to calculate essential matrices. In case of relative orientation with known rotations, the degree of freedom of essential matrix can be reduced by two. By reducing the number of unknowns to be estimated, the required computation time decreases. There are no guarantees that the features extracted in one image at this stage are the same for the counterpart image. RANSAC (Fischler, et al., 1981) in this problem can be used. RANSAC is to sample random points, determine them as inliers, set the range of inliers, and determine the target points that have many inlier points. Given the relative orientation of images, in triangulation, the points in 3D are computed, where uncertainties of these 3d points are obtained.

2.3.3 Multi-view stereo: Multiple images are needed to cover the whole environment for the estimated poses of all the adjacent images. This is to adjust the remaining uncertainties caused by long correspondences between images. The standard approach for this is BA, where it adjusts orientations, calibration parameters, and point locations with initial guess in the previous step. All observations and the uncertainties and correlations are exploited in BA for the correction. In the aerial bundle adjustment, control points are commonly assumed to be provided. However, if images without control points are provided, such as images taken in an underwater environment, BA corrects the geometry up to a similarity transform because the reference frame is not defined. Instead, the observations of camera pose determined by SLAM is used in this study.

3. EXPERIMENTAL RESULTS

3.1 Experimental Overview

The purpose of this experiment is to examine how the SLAM trajectory insertion influences the 3D reconstruction results of underwater images that have multiple visual disturbances. The aqualoc dataset provided by http://www.lirmm.fr/aqualoc/ is an underwater dataset to contribute to the localization of underwater vehicles for navigating close to the seabed (Ferrera et al, 2019b). Two sequences of the total 17 sequences provided from the aqualoc dataset were selected, which shows frequent visual disturbances at different depths. The sequence of A and

B are divided again according to the following conditions: the number of key point limit which means the upper limit of key points for each image, the number of images for the same video, and whether to process 3D reconstruction with the trajectory from visual SLAM. At the different settings for the experiment, we compare the results of SfM by the number of cloud points, SfM processing time, number of matched images/total images, and mean reprojection errors. Furthermore, dense point clouds of SfM are used for analyzing the quality of 3D reconstruction.

3.2 Description of Dataset and Experimental Method

3.2.1 Dataset: Archaeological site #2 (A) and #8 (B) acquired by the same camera sensor include visual disturbances such as turbidity, backscattering, sandy clouds, and dynamics around 270m and 390m depth. Most of the images in site #2 (A) include homogeneous textures of the seabed. In particular, site #2 (A) has a greater degree of disturbances than site #8 (B), making it difficult to perform feature correspondences for SLAM and SfM. In site #8 (B), amphorae are stacked in piles, which gives sufficient features, but, in the rest of them, uniform textures of the seabed come into the sight.

The camera for acquiring the dataset has a resolution of 986×604px and monochromatic images. The duration of videos is over 7’ 30’’ and under 8’ 00’’. The data description is in table 1, and the samples are shown in figure 2.

	Archaeological site #2 (A)	Archaeological site #8 (B)
Depth	≈270m	≈380m
Resolution	986×604px	
Focal length	6mm	
Duration	7’29’’	7’49’’
Visual disturbances	turbidity, backscattering, sandy clouds dynamics	turbidity, backscattering, dynamics

Table 1. Data description of archaeological site #2 and #8

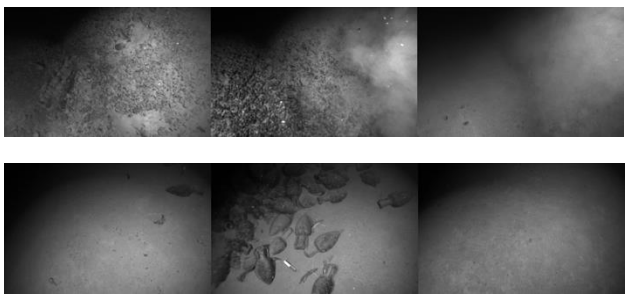

Figure 2. Samples of archaeological site #2 (top) and #8 (bottom)

3.2.2 Experimental method: The experiment set four conditions according to the following reasons. For A-1, a key point limit of 80,000 set as a default setting of SfM. The higher key point limit of 140,000 set for A-2 to see the quality of SfM by extracting more features. The number of images for A-3 sets every-half second to increase overlap, while the images for 3D reconstruction of A-1, A-2, and A-4 are extracted from each video every second. A-4 is processed with SLAM-determined orientations under the condition of 80,000 key point limit and every second frame. For B-1, B-2, and B-3, the settings are the same as A-1, A-2, and A-3. On the other hand, B-4 uses every half-second frame to see how much the result would be improved by more images and SLAM-determined orientations. The results for each of them were compared with the number of point clouds (SfM had generated up to the dense model), processing time, number of matched images out of total images, mean reprojection errors. The experiment utilized OpenVSLAM for visual SLAM and Photoscan (2014) for SfM.

3.3 Results of SfM with SLAM

3.3.1 Results from SfM with SLAM-determined orientation: Results of the experiment are shown in Tables 2 and 3. For A-1 setting 80,000 of key point limit, 245 images out of 450 images were matched for generating cloud points. Comparing A-2 with A-1, increasing the key point limit to 140,000 resulted in having more cloud points with increasing processing time. The number of images of A-3 has almost doubled, extracting images every-half second. Thus, cloud points have remarkably increased, but the processing time has increased nearly four times than A-1. Moreover, a mean reprojection error of A-3 has also increased by about 0.07 pixels. On the one hand, for A-4 that has the frame trajectories from SLAM, the number of cloud points has also increased with minimal increasing time, compared to A-1 under the same conditions of the number of images and key point limit. Since the feature tracking of OpenVSLAM was lost in the middle, 74 locations of images can be used in A-4. However, through the results of A-4, the trajectory from SLAM efficiently helped to increase the number of cloud points with minimal increasing time and almost invariant reprojection pixel errors.

B-1 and B-2 showed similar results as A-1, A-2, producing the increased number of clouds points by the different settings for the key point limit. Although the number of images was doubled for B-3, the number of cloud points did not increase unlike A-3, which could implicit that increasing the number of images is not always a solution to solve the 3D reconstruction problem for the unstable visual environment. Since OpenVSLAM was able to process for the whole frames, the entire trajectory was used for B-4. The results of B-4 has increased the number of point clouds as did A-4. Compared with the B-3, less time, the increased matched images, and the increased point clouds have revealed in B-4.

	Archaeological site #2 (A)			
	A-1	A-2	A-3	A-4 (with SLAM)
Key point limit	80,000	140,000	80,000	80,000
Number of images	450	450	899	450
Matched images /Total images	245/450	430/450	898/899	435/450
Number of cloud points	274,634	428,156	478,297	454,925
SfM processing time(s)	13’ 16’’	18’ 58’’	69’ 7’’	16’ 6’’
Mean reprojection error(px)	0.3885	0.3796	0.4598	0.3705

Table 2. Results of SfM with SLAM (with dataset A)

	Archaeological site #8 (B)			
	B-1	B-2	B-3	B-4 (with SLAM)
Key point limit	80000	140000	80000	80000
Number of images	470	470	940	940
Matched images /Total images	368/470	470/470	728/940	835/940
Number of cloud points	546,956	771,698	606,884	1,772,008
SfM processing time	15’ 27’’	22’ 39’’	67’ 32’’	50’ 3’’
Mean reprojection error(px)	0.4072	0.4137	0.4727	0.4450

Table 3. Results of SfM with SLAM (with dataset B)

3.3.2 Dense model of SfM with SLAM-determined orientation: The dense model results from dataset A and B are shown below. The dense model of A-3 is at the top of figure 3, which includes twice images as extracted in every half-second from a video. The bottom of figure 3 shows the result of A-4 reconstructed with SLAM-determined orientation. Since most of the frames contained the uniform seabed, the dense model of A-3 and A-4, outliers, which are assumed as sandy clouds, from A-3 (downward) are visible compared with A-4. The dense model of B-3 and B-4 in figure 4 shows the qualitative 3D reconstruction of SfM. Frames from both B-3 and B-4 are extracted every-half second, every 10 frames. Point clouds of B-3 appear to be less dense compared to B-4. B-4 with SLAM-determined orientation led to the shape of the amphora with a clearer curve than B-3.

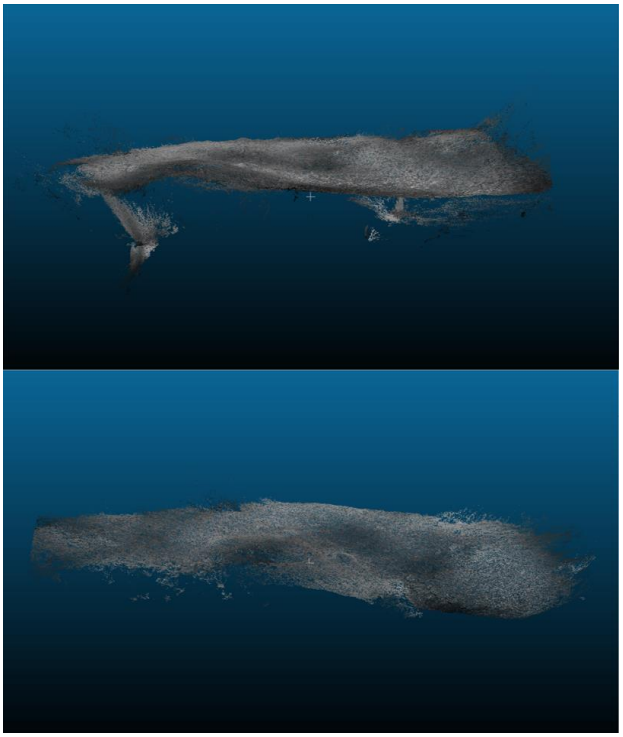

Figure 3. Dense models of A-3 (top) and A-4 (bottom)

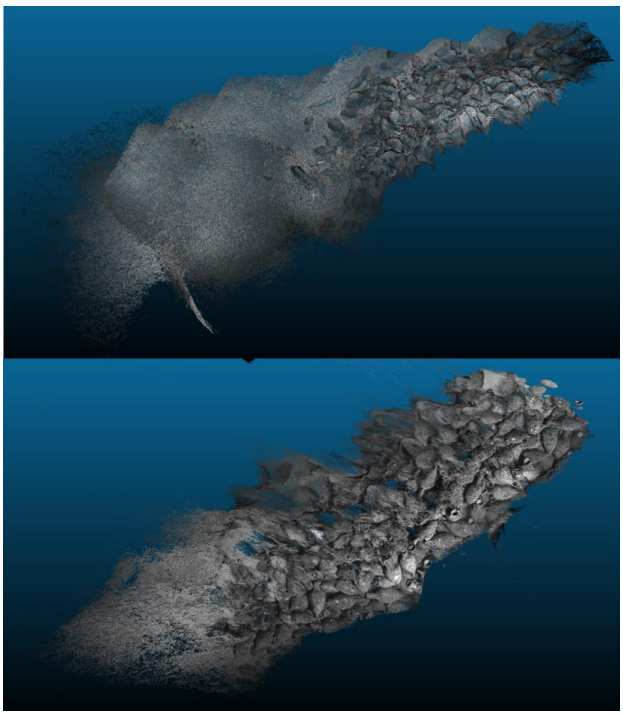

Figure 4. Dense models of B-3 (top) and B-4 (bottom)

4. CONCLUSIONS

A clear visual sight and known orientations of cameras are conditions for obtaining high-quality 3D reconstruction of SfM. In an underwater environment, visual disturbances and difficult usage of GPS/INS deteriorate the 3D reconstruction quality of SfM. This study intended to solve the 3D construction problem from these problems with SLAM-determined orientations. Visual SLAM, a solution to localization and mapping with a visual camera, provided frequent positioning decisions for all frames of the video, and the frame trajectories applied to SfM. The sequence of A and B are selected for this experiment. The two sequences were divided under certain conditions: number of keypoint limit, number of images for the same video, and whether the SLAM-determined orientation is inserted. In the final section, through the results of the number of cloud points, SfM processing time, and the number of matched images/total images and mean reprojection errors, we could see that the insertion of SLAM-determined orientation increases the number of points and the time efficiency. For the future work of this study, visual SLAM integrated with IMU and depth sensor, called Visual-inertial SLAM, will be combined for underwater 3D reconstruction, and the results of point clouds with known scales of objects will be included to assess trajectories from just SfM and SfM with known orientations of visual SLAM.

ACKNOWLEDGEMENTS

This research was supported by the project named “UAV-based Marine Surveillance System Development” which is a part of the project named “Wide Integrated Surveillance System of Marine Territory” funded by the Ministry of Maritime Affairs and Fisheries. (20150356)

REFERENCES

AgiSoft PhotoScan Professional. 2016. http://www.agisoft.com/downloads/installer/ (May 3 2020).

Dellaert, F., Seitz, S.M., Thorpe, C.E. and Thrun, S., 2000, June. Structure from motion without correspondence. In Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662) (Vol. 2, pp. 557-564). IEEE. doi.org/ 10.1109/CVPR.2000.854916

Ferrera, M., Moras, J., Trouvé-Peloux, P. and Creuze, V., 2019. Real-time monocular visual odometry for turbid and dynamic underwater environments. Sensors, 19(3), p.687. doi.org/10.3390/s19030687

Ferrera, M., Creuze, V., Moras, J. and Trouvé-Peloux, P., 2019. AQUALOC: An underwater dataset for visual–inertial–pressure localization. The International Journal of Robotics Research, 38(14), pp.1549-1559. doi.org/10.1177/0278364919883346

Fischler, M.A. and Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6), pp.381-395. doi.org/10.1145/358669.358692

Harris, C.G. and Stephens, M., 1988, August. A combined corner and edge detector. In Alvey vision conference (Vol. 15, No. 50, pp. 10-5244). doi.org/10.5244/C.2.23

Hidalgo, F. and Bräunl, T., 2015, February. Review of underwater SLAM techniques. In 2015 6th International Conference on Automation, Robotics and Applications (ICARA) (pp. 306-311). IEEE. doi.org/10.1109/ICARA.2015.7081165

Lowe, G. (2004). SIFT-the scale invariant feature transform. Int. J, 2, 91-110.

Mur-Artal, R. and Tardós, J.D., 2017. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Transactions on Robotics, 33(5), pp.1255-1262. doi.org/10.1109/TRO.2017.2705103

Rosin, P.L., 1999. Measuring corner properties. Computer Vision and Image Understanding, 73(2), pp.291-307. doi.org/10.1006/cviu.1998.0719

Rublee, E., Rabaud, V., Konolige, K. and Bradski, G., 2011, November. ORB: An efficient alternative to SIFT or SURF. In 2011 International conference on computer vision (pp. 2564-2571). Ieee. doi.org/10.1109/ICCV.2011.6126544

Schonberger, J.L. and Frahm, J.M., 2016. Structure-from-motion revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4104-4113). doi.org/ 10.1109/CVPR.2016.445

Sumikura, S., Shibuya, M. and Sakurada, K., 2019, October. OpenVSLAM: A Versatile Visual SLAM Framework. In Proceedings of the 27th ACM International Conference on Multimedia (pp. 2292-2295). doi.org/10.1145/3343031.3350539

Teague, J. and Scott, T., 2017. Underwater photogrammetry and 3D reconstruction of submerged objects in shallow environments by ROV and underwater GPS. Journal of Marine Science Research and Technology, 1(005).

Westoby, M.J., Brasington, J., Glasser, N.F., Hambrey, M.J. and Reynolds, J.M., 2012. ‘Structure-from-Motion’photogrammetry: A low-cost, effective tool for geoscience applications. Geomorphology, 179, pp.300-314. doi.org/10.1016/j.geomorph.2012.08.021

Gao, X., Zhang, T., Liu, Y. and Yan Q. 2017. 14 Lectures on Visual SLAM: From Theory to Practice. Publishing House of Electronics Industry.