UAV-Based SLAM and 3D Reconstruction System

Tianwei Li
Computer Science
University College London
London, UK
ucabtl1@ucl.ac.uk

Steve Hailes
Computer Science
University College London
London, UK
s.hailes@ucl.ac.uk

Simon Julier
Computer Science
University College London
London, UK
s.julier@ucl.ac.uk

Ming Liu
Robotics and Multi-perception Lab (RAM-LAB)
ECE&CSE Department, HKUST
City University of Hong Kong
Shenzhen Research Institute
Hongkong, China
eelium@ust.hk

***Abstract*—3D reconstructing a landscape is a prevalent problem that attracts a lot of interest in recent years. This project intended to verify whether the hypothesis of a UAV-based SLAM and 3D reconstruction system is practical. A GPS-Fused SLAM system is built based on ORB-SLAM. Inverse depth is also implemented to make the system suitable for a UAV-based platform. Meanwhile, REMODE is a depth filter and is tested as not being well enough as a dense mapping module. In the end, PMVS is implemented to build a dense map of the environment which produces a reasonable result. The small-scale-scene experiments produce the total error ratio of 5.60% in the x-y plane and 6.59% in the z axis.**

Keywords—UAV, SLAM, GPS, PMVS

I. Introduction

A UAV-based SLAM and 3D reconstruction system is developed based on ORB-SLAM[1] and PMVS[2] pipeline. The aim is to build a low-price and high efficient system to 3D reconstruct the landscape and the 3D model could be used for further analysis such as warning geological disasters[3].

SLAM system, also known as Simultaneous Localisation and Mapping system[4], is able to measure the position of the robot and build a description of the explored area simultaneously implementing the information of image flow. Compared with LDS-SLAM[5], SVO[6] and others, ORB-SLAM is selected to start with because of its robustness, reliability and other all-around performance.

Mavic[7] is picked as the platform as it has remarkable manevuer. The gimbal attached to the camera offers extra 2DOF and gives the best perspective of view for recording the target area. The gimbal attached makes the image flow steadier and harder to be lost. However, it also limits the implementation of inertial data[8, 9]. The SLAM system finally fuses GPS data to solve the problem of scale ambiguity. Inverse depth[10] is also implemented to make the system more suitable for the agent of a UAV. The camera poses that the SLAM system produces are used for 3D reconstruction.

The paper stresses the following contributions:

a) A SLAM system is built based on ORB-SLAM and is adjusted to be more suitable for a platform of a quadcopter with a gimbal. The SLAM system fuses GPS data to solve the scale ambiguity.

b) More than 100 times of experiments are processed to find out the optimal value of information matrix of GPS data which is implemented to keep the balance between GPS and visual informaiton.

c) Small-scale-scene experiements are processed to evaluate the performance of the system.

The paper finished the preliminary task of reconstructing a dense 3D model for an open landscape. The next step is to develop the function of trajectory planning based on point cloud[11]. The system implements core function to handle failure matching or observation in image flow. However, the robustness to GPS signal is still a problem, which could be mitigated by DP-fusion[12]. A unified framework for planning and execution-monitoring of mobile robots proposed by Liu and his team can also be a direction for further development of an autonomous system[13].

II. Related Work

The section talks about two typical methods of measuring a landscape. The comparisons between the typical methods and UAV-based method are also discussed to see the pros and cons of the purposed method.

A. Surveying

Geodetic surveying is a classical way of measuring a landscape. The constraint of classical surveying methods is that it is limited by the access to the target points. Moreover, accessibility is extremely limited in some landscapes. Possible little hills and plants also cause trouble in finding good visual lines from reference to target points. Also, it is especially difficult in landslide areas in which fixed or stable points cannot be placed in the neighborhood of the site [14].

Surveying with GPS Techniques will allow the surveyor to work even on rainy days. Meanwhile, as such method does not need a direct line of sight between stations and targets, hills and

trees will not be a problem. Moreover, GPS surveying provides high-level precision—12 to 16 mm in the x-y plane and 18 to 24 mm in z axis [14]. However, GPS surveying also has the common drawback as all the other surveying methods, the low accessibility to the targets. Accessibility varies at each particular site but is often restricted in mountain areas because the reflector has to be moved to the target areas. Therefore, the surveying methods may not be an ideal and safe way for monitoring some landscapes as these areas might be dangerous.

The productivity and accuracy of GPS surveying method are similar with classic geodetic surveying. Meanwhile, GPS techniques allow larger-scale scene than classic method. Compared with classic surveying method, GPS technique does not require direct line of sight between stations which would mitigate the limitation of obstacles [14]. The disadvantage of such method is that it required high labor cost.

B. SAR Methods

Synthetic aperture radar (SAR) is another way to monitor some landscapes. It could be either airborne or ground-based. SAR is used to collect images of the target area while each pixel in the image contains both gray information and phase signal interference. Together with its high precision, the SAR system could also investigate the displacement based on the phase variance as the phase of each pixel contains the information of depth along the line of sight [15]. Compared with surveying methods, SAR methods generate more information for further analysis of landscape motion.

However, the observation of airborne SAR method on steep slopes is highly distorted as perspective deformations intrinsically influence SAR images. Such drawback can only be slightly mitigated by processing data in the ascending and descending directions. Additionally, the precision of airborne SAR methods is only reasonable in the large-scale scenes and satellite-loaded SAR systems are also limited by revisiting time. For example, the revisiting time for ERS is 35days, and JERS even needs 44 days.

Comparing to the airborne SAR systems, the ground-based system suffers from different limitations. For instance, as the base is fixed, the system is only able to cover limited area which is about $1000{,}000m^2$. The system also requires a position which should have an open and wide vision of the target area. Moreover, the installation of such system is usually complicated. Such process would take a few hours which is high labour cost.

The main advantage of the UAV-based SLAM and Reconstruction system is its low cost. Compared with other methods, such system sacrifices some accuracy to reduce its cost dramatically as its sensor which collects most of the information is a camera. A UAV-based platform also offers remarkable mobility and maneuverability. One of the challenges is the accuracy of the 3D model built by the system. Moreover, the method to analyse models at different epochs is another problem that the system would come across in the future.

III. UAV-Based SLAM and 3D Reconstruction System

The first part of this section is about the platform and the procedure of collecting and processing the data. Afterwards, the second part is about the SLAM system. The majority is about the GPS fused principle. Last but not least, the 3D reconstruction part is discussed.

A. Platform

The preliminary mission that the system is aiming to finish is to reconstruct a dense 3D model for an open landscape. The figure below shows the block diagram of the system structure that has been built so far.

The quadcopter implemented in this project is DJI Mavic Pro which is a portable and powerful personal drone. The gimbal attached mitigates the effect of airflow.

Mavic Pro offers mobile SDK instead of Onboard SDK. One application called Litchi is chosen as the mobile application and is used in this project which offers the mature solution to trajectory planning.

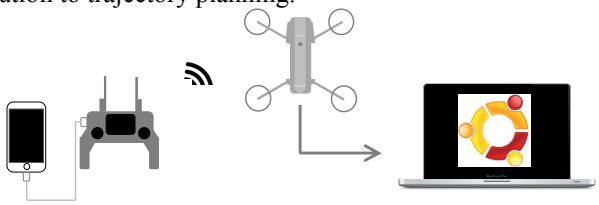

Fig. 1. The Platform of the Project which contains an iPhone, Mavic quadcopter and its controller and a Mac with Ubuntu 16.04.

The researcher sets the trajectory by Litchi in advance. Afterwards, the trajectory would be uploaded to the quadcopter through the controller. Once the Mavic Pro receives the command, it would fly to collect data following the route. The SLAM and 3D reconstruction algorithm are built in the environment of Ubuntu 16.04.

B. GPS Fused SLAM

The SLAM system is developed based on ORB-SLAM. The SLAM system is adjusted to be more robust on the platform of UAV and suitable for the camera of Mavic. For example, inverse depth is implemented.

The process of GPS data is a very common and crucial step is the system. The aerial video offers 30 frames per second while the frequency of the GPS is data is only 10 Hz. Since the frequency of the speed data achieved from the flight file is the same as the frequency of the GPS data, here only linear interpolation is implemented to let each frame have one corresponding position.

There is an item called “isVideo” in the flight data which indicates when the video is started to record. The time stamp when the “isVideo” is true is set as the time stamp of the first frame of the video initially.

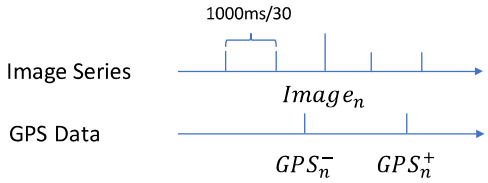

Fig. 2. For each image frame, this is a GPS datum before and after. The GPS data for the nth frame is calculated by linear interpolation.

As can be seen in the figure above, the axis on the top is the image series while the one on the bottom is the GPS data flow. $Image_n$ indicates the n^{th} frame. Since the aerial video is 30 frames per second, the time stamp of $Image_n$ is $T_n = T_1 + (n-1) \times 1000/30$, given that T_1 is the time stamp when the "isVideo" is true.

Once the time stamp T_n is found, the closest GPS data on the left and right sides will be found which are names as GPS_n^- and GPS_n^+.

The equation of GPS_n is shown as below.

$$GPS_n = GPS_n^- + (GPS_n^+ - GPS_n^-)\frac{T_n - T_n^-}{T_n^+ - T_n^-} \quad (1)$$

where T_n^+ is the time stamp of GPS_n^+ while T_n^- is the time stamp of GPS_n^-.

Once thing that needs to be noticed is that the first frame does not actually correspond to the time stamp when "isVideo" is true. There is a delay between the time when "isVideo" is true and the time when the first frame is recorded. The experiment shows that the delay is up to 3300 ms. The figure below shows speed variation of the pure vision SLAM outcome and GPS data, which proves that the timestamps of the image frames and GPS data fit each other very well.

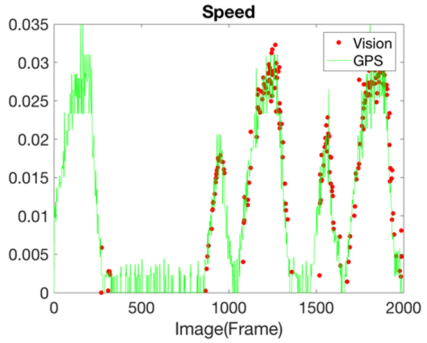

Fig. 3. The red dots indicate the speed of the camera calculated by pure vision while the green line shows the speed variation of GPS data.

In terms of GPS fusion, the key problem is how to set the error term for the GPS observation within the algorithm. The easier way to compute the error term is shown as below.

$$e = \begin{bmatrix} x_C - x_G \\ y_C - y_G \\ z_C - z_G \end{bmatrix} \quad (2)$$

where $[x_C, y_C, z_C]^T$ is the position of the camera while $[x_G, y_G, z_G]$ is the 3D coordinate derived from GPS data.

This method offers the difference in different axis. However, the GPS signal is always fluctuating even the UAV stays at the same point. Therefore, it is absolutely a terrible model.

TABLE II Accuracy of GPS Data

Vertical	+/- 0.1m (when Vision Positioning is active) or +/- 0.5m
Horizontal	+/- 0.3m (when Vision Positioning is active) or +/- 1.5m

The table above shows the specification of the accuracy of the GPS data. Therefore, the maximum error in the horizontal and vertical direction can be used to design the model. As the GPS signal is not very steady, it is very common that the GPS data contains an error. Here, a new way of computing the error is shown as below.

$$e = \begin{bmatrix} \frac{|x_C - x_G|^n}{1.5} \\ \frac{|y_C - y_G|^n}{1.5} \\ \frac{|z_C - z_G|^n}{0.5} \end{bmatrix} \quad (3)$$

where n is an exponential index which would be tested in the result section in order to find out the most suitable value.

The idea of this equation is that if the difference between the camera position and the GPS data is less than the threshold given by the specification, the system does not care. Once the error is bigger than the threshold, the system would put its concentration on it.

C. 3D Reconstruction System

Two types of dense mapping algorithm called REMODE and PMVS have been tried in this project which would be discussed one by one as follows.

REMODE (REgularized MOnocular Depth Estimation) [16] is operating based on the camera poses which are given by the front or back ends. Unlike local mapping, the system cannot take each pixel as the feature point. Therefore, matching becomes a very important step in dense mapping. In order to locate the position of the pixels in other images, the epipolar line searching and block matching techniques are implemented. Once the corresponding pixel positions in different images are achieved, the method of triangulation can be used to figure out the depth. What's different is that the algorithm will do multiple times of triangulation to converge the depth which is called depth filter.

REMODE turned out to be not suitable or not good enough for this project. The block matching algorithm gets struggled in the scene full of plants. There are a lot of features which are not distinct. Such circumstances would make NCC (Normalized Cross Correlation) hard to find out a peak which corresponds to the right match.

Another problem of REMODE is that it needs too many frames to converge the pixels on the single image, which is very inefficient. For example, for a single image, the algorithm needs more than 200 images to estimate the depth of each pixel on that image. 200 images mean that the UAV needs to hover at the same place for more than 6.67 seconds, which is unacceptable for a UAV whose max single flight time is 20 min.

PMVS (Patch-based Multi-view Stereo Software) [2] is able to reconstruct a 3D model based on a set of images, and the corresponding camera poses. The software will only

reconstruct rigid structure. Therefore, non-rigid objects like moving people or waving leaves will not be paid attention to in the scene of the target area. Only an interface is built between the SLAM system and PMVS software.

IV. EXPERIMENT

The system has been tested given shift-rotation adjusted GPS data to solve the problem of scale ambiguity. More than 100 times of experiments are operated to find out the best setting to figure out the balance between image information and GPS information.

The experiments below test whether the system could offer good performance if the GPS data is only shifted. Such test could verify if the system is rotation invariant. Three rounds totaling 72 iterations of experiments were completed to get the representative data.

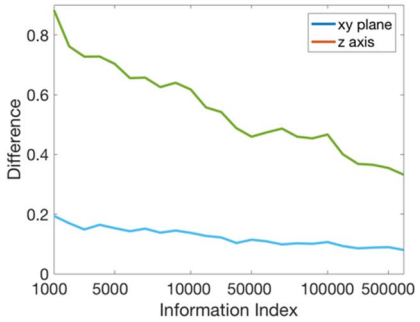

Fig. 4. Translation Difference between GPS data and Outcomes of System given Different Information Index.

As can be seen in the figure above, the difference generally decreases with the increase of information index. The bigger the information index is, the more confident the optimizer is for the GPS data. The difference in the x-y plane is bigger than the one in z axis because the system is more tolerant of the error in the x-y plane.

The figure below shows the performance of the pose estimation of the system gives different information index. The system achieves the smallest difference when the information index is set to be $2e+04$ or $7e+04$. The system also achieves good results when the information indiex is set to be bigger than $4e+05$. However, the event of tracking lost may happen given such setting. Therefore, the optimal information index is $2e+04$ or $7e+04$.

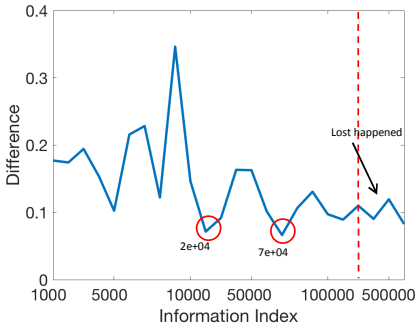

Fig. 5. Difference between ORB-SLAM Result and System Outcome Given Different Information Index.

All the experiments so far are operated given no ground truth data. The parameters are adjusted based on more than 100 times of tests. The information matrix is hard to be adjusted because it is not only about the covariance of a GPS signal but also the balance between a single GPS observation and a bunch of map point observations.

After the experiments of SLAM system, it is time to assess the performance of dense mapping module. Given no ground truth of any environment, several settings are build based on some objects of which sizes can be measured.

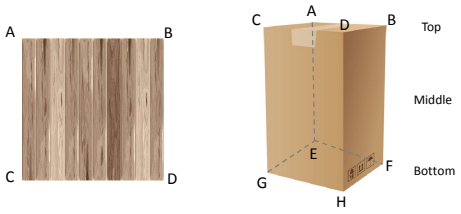

Fig. 6. Top: Objects Used in the Small-Scale Scene. Bottom: The labelling method for table and boxes.

As can be seen in the figure above, three types of items are used in the experiment which are a table, small and big boxes. The table below shows the mean size of each item by three times of measurements.

TABLE III Items Size

Item	x(cm)	y(cm)	z(cm)
Table	59.77	59.80	70.90
Big Box	47.27	38.10	57.40
Small Box	37.33	24.57	20.57

These items would be put in the scene. However, the assessment doesn't care about the position of each item. Once the 3D model is reconstructed, the 3D coordinate of the corners of each item would be found out. As the surface of the table is rectangular and the boxes are cubic, the 3D coordinate of the corners would be used to calculate the length of the edge. Each length would be compared with the length of the corresponding edge to measure the performance of 3D reconstruction.

There are a table, a small box and one composition in the scene. As can be seen in the figure below, the photo on the left side is the experiment scene while the image on the right side is the 3D reconstruction result. The model is reasonable. However, it is still necessary to assess the 3D reconstruction module by statistics.

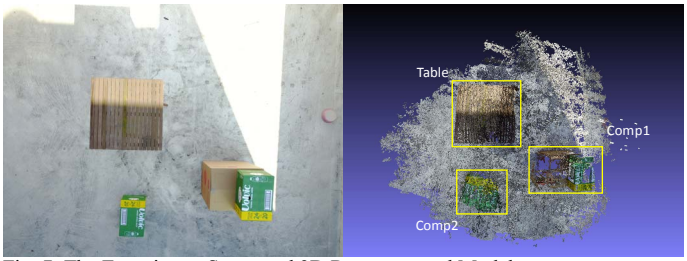

Fig. 7. The Experiment Scene and 3D Reconstructed Model.

Once the reconstruction is finished, the 3D coordinate of each corner would be found out. Afterwards, the length of each edge would be calculated.

TABLE IV Error Calculation of Item Table

Table	3D	Ground Truth	Absolute Error	Relative Error
AB	0.4275	0.5977	0.1702	0.0130
BD	0.4434	0.5980	0.1546	0.0355
CD	0.4256	0.5977	0.1721	0.0103
CA	0.3774	0.5980	0.2206	0.0588

The ground truth value of each edge is the length measured before. The equations shown below are used to calculate the errors.

$$e_a = |l_{3D} - l_{gt}| \quad (4)$$

$$e_r = |s \times l_{3D} - p_{gt}| \quad (5)$$

where l_{3D} indicates the length of a particular edge in the 3D model while l_{gt} is the length of the corresponding edge in reality. e_a and e_r are the absolute and relative error respectively.

The mean error of each edge is 18cm which contains the error caused by a scale problem. As mentioned before, the GPS data is implemented to solve the problem of scale ambiguity. Such method could achieve a relatively good performance in a large-scale scene. As the experiment is set to be a small-scale scene, the influence of the GPS data error is amplified. Once the scale value is added to the model, the relative error is quite small. The error ratio of the table is 4.92%, which proves that the relative accuracy of the 3D reconstruction module in the x-y plane is quite high.

The big and small boxes are basically cubes. Each cube has 12 edges which are divided into three parts. The top and bottom parts measure the performance in the x-y plane as the middle part measures the performance in the z axis.

TABLE V Error Calculation of Object Big and Small Boxes

	Small Box	3D	Ground Truth	Absolute Error	Relative Error
Top	AB	0.1516	0.2457	0.0941	0.0072
	BD	0.2615	0.3733	0.1118	0.0380
	CD	0.1222	0.2457	0.1235	0.0535
	CA	0.2518	0.3733	0.1215	0.0227
Mid	AE	0.1166	0.2057	0.0891	0.0040
	BF	0.1368	0.2057	0.0689	0.0403
	DH	N/A	0.2057	N/A	N/A
	CG	0.0898	0.2057	0.1159	0.0443
Bot	EF	0.1580	0.2457	0.0877	0.0028
	FH	N/A	0.3733	N/A	N/A
	GH	N/A	0.2457	N/A	N/A
	EG	0.2169	0.3733	0.1564	0.0322
	Average			0.1077	0.0272
	Big Box	3D	Ground Truth	Absolute Error	Relative Error
Top	AB	0.2938	0.4727	0.1789	0.0106
	BD	N/A	0.3810	N/A	N/A
	CD	N/A	0.4727	N/A	N/A
	CA	0.2489	0.3810	0.1321	0.0106
Mid	AE	0.2566	0.5740	0.3174	0.0328
	BF	0.2381	0.5740	0.3359	0.0110
	DH	N/A	0.5740	N/A	N/A
	CG	0.2335	0.5740	0.3405	0.0218
Bot	EF	0.2806	0.4727	0.1921	0.0313
	FH	0.2590	0.3810	0.1220	0.0265
	GH	0.2864	0.4727	0.1863	0.0222
	EG	0.2335	0.3810	0.1475	0.0137
	Average			0.2170	0.0201

The table above shows measurements for each edge. The error ratio of the small box in the x-y plane is 8.43% while the error ratio in the z axis is 14.36%. The error ratio of the big box in the x-y plane is 4.49%, and the one in the z axis is 3.81%.

The total error ratio is calculated by weight which is defined by the size of each object. The total error ratio is 5.60% in the x-y plane and 6.59% in the z axis.

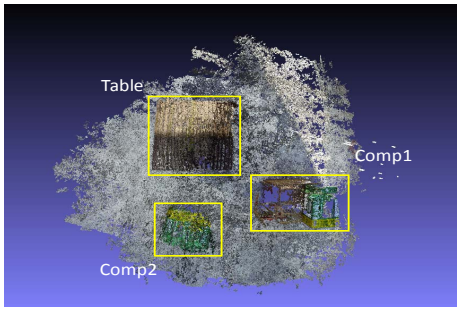

Fig. 8. The Point Cloud of the Experiment Scene.

One thing that can be noticed is that the model of Comp2 which is a small box is wrong as it looks like there are two boxes overlapped. This is because the quadcopter follows the trajectory which is shown as below to get a denser map. The dense mapping module reconstructed the small box twice by mistake.

By single lawnmower pattern, the system can build a 3D map of the scene and avoid the mistake as shown above. However, the outcome may suffer the problem of being incomplete which is shown as below. Therefore, further research is needed to improve the accuracy of pose estimation and 3D reconstruction.

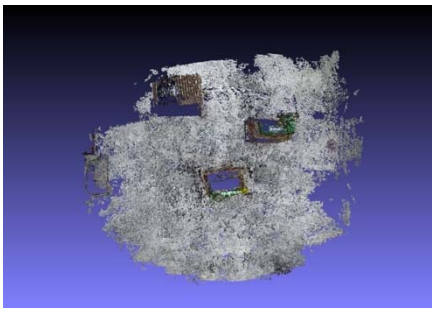

Fig. 9. The Point Cloud of the Experiment Scene by a single Lawnmower Pattern.

V. CONCLUSION

The results prove that the UAV-based SLAM and reconstruction system is feasible. A SLAM system is built based on ORB-SLAM and is adjusted to be more suitable for a quadcopter. The parameterization implements inverse depth to deal with infinite-depth pixels. The SLAM system is able to process high-resolution images and fuses GPS data to solve the problem of scale ambiguity. More than 100 times of experiments are processed to find out the optimal value of information matrix of GPS data. Moreover, the mathematical model of GPS could be optimized by using Cauchy.

PMVS is implemented for 3D reconstruction. An interface is built to connect the SLAM system and the PMVS software. Several small-scale-scene experiments are also processed to measure the performance of the reconstruction module.

As PMVS is not a real-time system, it is impossible to do real time trajectory planning based the outcome of PMVS. It is also difficult to do trajectory planning based on the sparse local map. However, as Mavic contains the function of front-side collision avoidance based on vision, a trajectory planning which is intended to get better measurements is possible according to the local map. The algorithm could lead the quadcopter to get a better perspective of the points where the covariance is very big.

With such methodology, a step-by-step UAV semantic SLAM system could be developed with autonomous trajectory planning.

In classic SLAM system, textures are necessary for feature matching. However, the calculation of the key point, its corresponding descriptor and feature matching process take the majority of computation of a SLAM system. Once a man-made (normally geometric structure) object can be recognized, a structure type of variable could describe the object which could save much space and computation cost. Meanwhile, SLAM system may even able to recognize non-static environment with known structure.

Even if an autonomous SLAM is built someday, there is still more research that must be done. For example, suppose two 3D models M_{t1} and M_{t2} of one target area recorded at different time are built. The system should be built to compare the two models and analyse the models to make some prediction for the future variation.

ACKNOWLEDGMENT

This work is supported by Shenzhen Science, Technology and Innovation Commission (SZSTI) JCYJ20160428154842603 and JCYJ20160401100022706; partially supported by the Research Grant Council of Hong Kong SAR Government, China, under project No. 16206014 and No. 21202816; National Natural Science Foundation of China No. 61403325, awarded to Prof. Ming Liu. Ming Liu is also with the City University of Hong Kong Shenzhen Research Institute.

REFERENCES

[1] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics,* vol. 31, no. 5, pp. 1147-1163, 2015.

[2] Y. Furukawa and J. Ponce, "Patch-based multi-view stereo software (pmvs-version 2)," *PMVS2, University of Washington, Department of Computer Science and Engineering. Web. Downloaded from on May,* vol. 14, 2012.

[3] F. Guzzetti, A. Carrara, M. Cardinali, and P. Reichenbach, "Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy," *Geomorphology,* vol. 31, no. 1, pp. 181-216, 1999.

[4] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: a survey," *Artificial Intelligence Review,* vol. 43, no. 1, pp. 55-81, 2015.

[5] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *European Conference on Computer Vision*, 2014, pp. 834-849: Springer.

[6] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, 2014, pp. 15-22: IEEE.

[7] D. Brown, "Mavic Pro vs. Phantom 4," 2017.

[8] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular SLAM with map reuse," *IEEE Robotics and Automation Letters,* vol. 2, no. 2, pp. 796-803, 2017.

[9] T. Qin and S. Shen, "Robust Initialization of Monocular Visual-Inertial Estimation on Aerial Robots," in *Proc. of the IEEE/RSJ Int. Conf. on Intell. Robots and Syst., Vancouver, Canada*, 2017.

[10] J. Civera, A. J. Davison, and J. M. Montiel, "Inverse depth parametrization for monocular SLAM," *IEEE transactions on robotics,* vol. 24, no. 5, pp. 932-945, 2008.

[11] M. Liu, "Robotic online path planning on point cloud," *IEEE transactions on cybernetics,* vol. 46, no. 5, pp. 1217-1228, 2016.

[12] M. Liu, L. Wang, and R. Siegwart, "DP-Fusion: A generic framework for online multi sensor recognition," in *Multisensor Fusion and Integration for Intelligent Systems (MFI), 2012 IEEE Conference on*, 2012, pp. 7-12: IEEE.

[13] M. Gianni *et al.*, "A Unified Framework for Planning and Execution-Monitoring of Mobile Robots," *Automated action planning for autonomous mobile robots,* vol. 11, p. 09, 2011.

[14] J. A. Gili, J. Corominas, and J. Rius, "Using Global Positioning System techniques in landslide monitoring," *Engineering geology,* vol. 55, no. 3, pp. 167-192, 2000.

[15] D. Tarchi *et al.*, "Landslide monitoring by using ground-based SAR interferometry: an example of application to the Tessina landslide in Italy," *Engineering geology,* vol. 68, no. 1, pp. 15-30, 2003.

[16] M. Pizzoli, C. Forster, and D. Scaramuzza, "REMODE: Probabilistic, monocular dense reconstruction in real time," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, 2014, pp. 2609-2616: IEEE.