Unmanned Aerial Vehicle Visual Simultaneous Localization and Mapping: A Survey

Y Tian[1, a], **H Yue**[1, b, *], **B Yang**[2, c] **and J Ren**[3, d]

[1] Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XW, UK
[2] College of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan 030024, China
[3] National Subsea Centre, Robert Gordon University, Aberdeen AB21 0BH, UK

[a] yikun.tian@strath.ac.uk; [b, *] hong.yue@strath.ac.uk; [c]632717115@qq.com; [d]j.ren@rgu.ac.uk

Abstract. Simultaneous Localization and Mapping (SLAM) has been widely applied in robotics and other vision applications, such as navigation and path planning for unmanned aerial vehicles (UAVs). UAV navigation can be regarded as the process of robot planning to reach the target location safely and quickly. In order to complete the predetermined task, the drone must fully understand its state, including position, navigation speed, heading, starting point, and target position. With the rapid development of computer vision technology, vision-based navigation has become a powerful tool for autonomous navigation. A visual sensor can provide a wealth of online environmental information, has high sensitivity, strong anti-interference ability, and is suitable for perceiving dynamic environments. Most visual sensors are passive sensors, which prevent sensing systems from being detected. Compared with traditional sensors such as global positioning system (GPS), laser lightning, and ultrasonic sensors, visual SLAM can obtain rich visual information such as color, texture and depth. In this paper, a survey is provided on the development of relevant techniques of visual SLAM, visual odometry, image stabilization and image denoising with applications to UAVs. By analyzing the existing development, some future perspectives are briefed.

1. Introduction

Simultaneous Localization and Mapping (SLAM) means that the machine realizes environmental perception and understanding in an unfamiliar environment to complete its positioning and path planning [1]. Localization and mapping are the basic needs of humans and mobile devices. In fact, we human can perceive our movements and the environmental positioning through multimodal senses, and rely on this awareness to locate and navigate oneself in a complex three-dimensional (3D) space [2]. Although the global positioning system (GPS) is the most commonly used positioning system that is capable of providing a relatively accurate position without error accumulation, it is only suitable for outdoor scenes where the sky is not obviously blocked. In addition, the positioning error of commercial GPS systems can only reach the meter level, which is insufficient for applications that require centimeter-level errors, such as accurate alignment in construction and automatic parking. SLAM provides an attractive alternative to user-built maps, demonstrating that real-time navigation can be achieved in unknown environments even without the GPS or temporary positioning infrastructure.

SLAM technology is widely used in autonomous vehicles, service robots, augmented reality, and virtual reality. Other applications include mobile and wearable devices such as cell phones, wristbands, or Internet of Things (IoT) devices, from walking navigation to emergency response.

SLAM is essentially a state estimation problem, which is divided, according to sensors, into the LiDAR SLAM [3] and the visual SLAM (vSLAM). The research of LiDAR SLAM is relatively mature in theory and engineering. Many industries have begun to use LiDAR SLAM to carry out industrial tasks. The vSLAM uses images as the main origin of environmental perception information. To calculate the camera's posture as the main goal, a 3D map is constructed by the multi-view geometric method. The vSLAM is still in the laboratory research stage and has few practical applications.

A complete SLAM system consists of four parts: the front-end tracking, the back-end optimization, the loop detection, and the map reconstruction. The front-end tracking is a visual odometer responsible for preliminary estimation of the posture state between the camera frames and the location of map points. The back-end optimization is responsible for receiving the altitude information measured by the visual odometry front-end, in a probabilistic framework. The loopback detection is attributed to judging whether the mobile device has returned to the original position and performing a loopback to correct the predicted error. The map reconstruction is to rebuild a map that is suitable for the task requirements based on the camera's posture and the required image.

Visual odometry is a challenging and open topic in vSLAM systems. Vision-based methods are often the preferred choices because of several factors such as cost-saving, low power requirements, and useful complementary information that can be provided to other sensors e.g. inertial measurement unit and the GPS [4]. At present, advanced vSLAM has achieved quite high accuracy as validated by using both public data sets and actual physical experiments.

With the recent fast development of integrated circuits performance and computer vision techniques, there are huge number of applications with sensing equipment engaged with single or binocular cameras based on vSLAM. Benefited from visual-sensing equipment on huge data processing and the high flexibility, vSLAM technology can be built up with a simpler system structure, low costing and easy implementation [5]. This is why simultaneous positioning and mapping have attracted increasing research interests in recent years, which have produced immense implementation value in various applications.

2. Conventional vSLAM

Traditional vSLAM solutions are divided into feature point methods and direct methods. The characteristic point method is mainly composed of feature detector, feature descriptor, and motion estimation. The feature detectors include the point-feature detector and the blob feature detector. The motion estimation is divided into the 3D-to-3D methods, the 3D-to-2D methods and the 2D-to-2D methods. The direct method no longer extracts feature points, where the camera's posture and map structure can be directly extracted through the photometric mechanism, without calculating the feature points and descriptors.

The feature point method includes the following three implementation steps. The first is to extract stable characteristic points from each picture frame. These features are usually immutable descriptions, and the matching of adjacent frames is done through the descriptor. The second is to restore the camera posture and the coordinates of the mapping point through the epipolar geometry. In the third step, the camera posture and the mapped structure are fine-tuned by minimizing the projection error. Feature points extracted from each frame are cyclically detected or relocated through operations such as data clustering.

Although the feature-point based methods are the mainstream in traditional visual odometry, they have some shortcomings. The extraction of feature points and the computation of descriptors are very time-consuming. All other information is ignored except for the feature points. The camera sometimes moves to places that has no features, such as white walls or empty corridors that lack obvious texture information. The number of the feature points in these scenes can be obviously reduced, resulting in insufficient matching points to determine the camera motion. Since the direct method does not need to

extract feature points, nor can it characterize the global features of an image, the loop detection of the direct method is still an open topic.

The traditional vSLAM has the following problems that remain unsolved satisfactorily.

- Under adverse conditions such as poor lighting conditions or large changes in lighting, the robustness of the algorithm becomes weak.
- In the case of large movement or rapid shaking, traditional vSLAM algorithms are prone to unmatched feature points.
- Traditional algorithms cannot identify foreground objects, that is, moving objects in the scene can only be considered as "bad pixels".

The localization and mapping problem has been studied for decades, and various manual design models and algorithms are still under development. Under ideal conditions, most of the developed sensors and models can accurately estimate system conditions without time constraints and across different environments. However, factors such as imperfect sensor measurements, inaccurate system modeling, complex environments, and unrealistic constraints can affect the reliability and accuracy of manual packaging systems [6].

3. Data-driven visual odometry

With the development of deep learning technology for computer vision, many vision problems have been converted from conventional imperfect modelling to a learning problem and achieved higher breakthroughs [7]. Combining deep learning and vSLAM overcomes the limitations of manual design algorithms such as the visual odometry and scene recognition. It improves the learning ability and intelligence of robotic platforms.

The advantages of deep learning methods in this context are as follows. Firstly, deep learning algorithms utilize multi-layer neural networks to learn the hierarchical feature representations and automatically discover task-related features. They can thus help to learn a model more invariant to different situations, such as featureless areas, dynamic lightning conditions, motion blur, and precise camera calibration, which are extremely challenging in traditional modelling [8]. Moreover, high-level semantic information can be extracted through deep learning, which is particularly useful for understanding and use of semantic vSLAM and scene semantic information [9]. On the contrary, it is very difficult to describe this kind of semantic information in vSLAM using formal mathematical terms.

Secondly, deep learning algorithms are more in line with the laws of human cognition and the interaction of the environment. The learning method allows the spatial machine intelligence system to learn from experience and actively utilize new information. Constructing a standard data-driven model avoids the workforce required to specify the mathematics and physics rules [10] to solve specific domain problems before real deployment. This ability may enable learning machines to automatically discover new computing solutions in emerging scenarios including user customized modelling.

4. UAV image stabilization

Videos captured from unmanned aerial vehicles (UAV) are often susceptible to unexpected sensor movement, which can severely interfere with the detection and tracking of the targets of interest. Before 2010, there were two main methods of video stabilization, i.e. the mechanical video stabilization (MVS) and the digital video stabilization (DVS). MVS stabilizes frames of videos via a special equipment, it cannot remove all video vibration, and the cost of hardware is high. On the contrary, DVS stabilizes video frames by image processing techniques, which is effective with a relatively low cost. DVS consists of three stages: global motion estimation (GLE), motion filtering and compensation, and video completion.

In GLE, the global motion between two consecutive frames is determined by using global intensity alignment approaches [11]. Compared with global intensity alignment approaches, feature-based methods are generally faster. However, they are always vulnerable to local effects [12]. After motion capture, motion filtering methods, such as the Random Sample Consensus (RANSAC), the Least Mean Square (LMS), and the M-estimate of Sample Consensus (MSAC), are employed to eliminate the outlier

vectors to avoid misleading movement. Motion correction then uses motion vector integration, low-pass filtering, Kalman filter, etc. to distinguish the intentional motion from the unintentional motion. Finally, frame compensation moves the image frame out of the correction vector from the filtered result. The purpose of video completion is to fill in missing frames in a video, where a high-quality visual effect can be obtained if sufficient similar information is obtained.

Based on the three-stage framework of DVS, several variants are also proposed. Mai et al. [13] introduced a motion prejudge procedure before the GLE, among which the flight parameters of the drone and the control commands of the camera are used as auxiliary data to predict the motion mode of the image sequence. Wang et al. [14] introduced a linear fitting auxiliary curve-fitting method for global motion estimation to optimize the saturation of videos caused by the original curve fitting. Kejriwal et al. [15] extracted the corner points for motion capturing, and then the optical flow is utilized on features of two consecutive frames. Similar to this, Wang et al. [16] utilized the Good Features to Track (GFTT) algorithm for feature extraction, followed by the sparse optical flow method to build motion vectors. Meanwhile, a hybrid filter composed of Kalman filter and Gaussian discrete filter was introduced to smooth the motion of each frame image. Rahmani et al. [17] adopted the dense optical flow to match the features between two frames. The rigid transform motion that handles rotation and translation is also utilized to acquire a wide imagery range, which is achieved by the K-means clustering. By considering orientation estimation, Odelga et al. [18] proposed a method that uses the onboard inertial measurement unit of the quadrotor to calculate the camera rotation without the need for a special altitude and heading reference system.

5. UAV image denoising

As a low-altitude remote sensing platform, UAV images are easily affected by lightning, ground electromagnetic waves, light changes, and the UAV mechanical noise. Thus, it is necessary to apply algorithms to remove such noise from the acquired images.

In recent years, several works have been conducted on UAV image denoising. This review will focus on methods that are applied to four widely used image types, i.e. (Red, Green, Blue, RGB), infrared (IR), hyperspectral /multispectral, and synthetic aperture radar (SAR) images.

Liu et al. [19] explored the correlation among the different bands of multicomponent images regarding image denoising. In their study, the auxiliary image, taken from additional sensor with the same scene, is introduced as a prior of the partial differential equation denoising method. Experiments were performed on multi-spectral and hyper-spectral remote sensing images, where the noise is generated by a Gaussian noise.

To reduce the speckle noise observed from SAR images, Rajapriyadharshini et al. [20] proposed a clustering-based method. A noise image was first split into several disjoint local regions, each of which was then denoised by the Wiener filter and further checked for linear minimum mean square error shrinkage in a linear discriminant analysis domain. A denoised SAR image is eventually aggregately produced by the denoised patches. Wang et al. [21] proposed a high-order balanced multi-band multi-wavelet packet transform to denoise remote sensing images. The denoising performance was improved by a higher balance order or frequency band with an increased computational cost, which was tested using several kinds of mixed noise such as multiplicative noise, salt and pepper noise, and Gaussian/Poisson noise.

Xu et al. [22] proposed a method focused on the nonlocal means (NLM) algorithm to denoise dual image patches. The noise is produced by additive white Gaussian noise with zero mean and a variety of standard deviations. For the same denoising target, Penna et al. [23] proposed two methods, based on the stochastic distances and a priori NLM, respectively. Kim et al. [24] adopted an adaptive noise filtering method based on background registration and robust principal component analysis to deal with unfixed pattern noise in UAV infrared images.

Some works are also reported on the sparse representation based denoising methods. In [25], combining unidirectional total change and sparse representation regularization are used for the de-streaking and de-noising of images. To improve both the automation level and the (corresponding)

denoising results, Cerra et al. [26] improved the unmixing-based denoising method by adjusting the contribution of each spectral band to the final result as well as its automation degree (dataset available at: https://www.enmap.org/). Xu et al. [27] transformed the original SAR image to the logarithmic SAR image domain, followed by image denoising using a non-local sparse model and iterative regularization techniques. Liu et al. [28] further exploited features from similar reference images from different bands and sensors for effective data fusion and sparse representations.

To further improve the efficacy of denoising and feature extraction, a number of new techniques can be further applied. Padfield et al. [29] proposed sparse learning and genetic feature selection for effective signal classification. In [30], superpixel-adaptive singular spectrum analysis is proposed for efficient feature extraction in hyperspectral images (HSI). Sun et al. [31] proposed Gumbel-softmax trick to enable energy-constrained concrete autoencoders for dimensionality reduction in HSI. In [32], a multi-level residual feature fusion network for medical applications is proposed. In Ren et al [33], a multitask learning U-net is proposed for effective segmentation of cardiac images. Liu et al. [34] proposed energy-aware correlation filter for visual tracking in dealing with dark objects. In Sun et al. [35], deep fusion of local spectral and multi-scale features is introduced for efficient classification of HSIs. In addition, topological optimization of the deep learning architectures can be used [36].

6. Conclusions and discussions

In this review, we summarize the research on UAV based path planning using vSLAM with environmental perception and understanding. It is worth noting that the videos captured by UAVs are usually susceptible to unexpected sensor movements, which can seriously interfere with the detection and tracking of the objects of interest. Replacing the homography transformation with a deep-learning based reconstruction method can actually improve the quality of UAV captured image. As a low-altitude remote sensing platform, UAV images are easily affected by lightning, ground electromagnetic waves, light changes, and the UAV mechanical noise. It is necessary to denoise UAV images in a more efficient way i.e. by accelerating with a GPS and more effective models using the emerging deep learning and computer vision approaches.

References

[1] Cadena C, Carlone L, Carrillo H, Latif Y, Scaramuzza D, Neira J, Reid I and Leonard J 2016 *IEEE Trans. Robot.* **32** 1309-32
[2] Fetsch C, Turner A, DeAngelis G and Angelaki D 2009 *J. Neurosci.* **29** 15601-12
[3] Pan Y, Xiao P, He Y, Shao Z and Li Z, 2021 Versatile LiDAR SLAM via multi-metric linear least square *Preprint arXiv:2102.03771*
[4] Yang B, Xu X, Ren J, Wang B, Cheng Y, Ling W 2022 *Pattern Recognition Letters* **153** 126-35
[5] Sünderhauf N *et al*., 2018 *Int J. Rob. Res.* **37** 405-20
[6] McCormac J, Handa A, Davison A and Leutenegger S, 2017 *Proc. Int. Conf. ICRA* (Singapore) pp 4628-35
[7] Ma L, Stückler J, Kerl C and Cremers D, 2017 *Proc. Int. Conf. IROS* (Vancouver, Canada) pp 598-605
[8] Brendan McMahan H, Holt G, Sculley D, Young M, Ebner D, Grady J and Nie L, 2013 *ACM SIGKDD Int. Conf. KDD* (Chicago, USA) pp 1222-30
[9] Zhang Z and Scaramuzza D, 2018 *Proc. Int. Conf. IROS* (Madrid, Spain) pp 7244-51
[10] Grupp M, evo: Python package for the evaluation of odometry and SLAM URL https://michaelgrupp.github.io/evo/
[11] Buyukyazi T, Bayraktar S and Lazoglu I, 2013 *Proc. RAST* (Istanbul, Turkey) pp 121-6
[12] Yu J, Luo C, etc, 2015 *Proc. Int. Conf. CCCV* (Berlin, Germany) pp 180-9
[13] Mai Y, Zhao H and Guo S, *2012 Proc. Int. Conf.* CSEE (Hangzhou, China) pp 586-9
[14] Wang L, Zhao H, Guo S, Mai Y and Liu S, 2012 *Proc. IGARSS* (Munich, Germany) pp 4391-4
[15] Kejriwal L and Singh I 2016 *Procedia Comput. Sci.* **93** 359-66
[16] Wang J, Qi J, Liu J and Jiang Y, 2020 *Proc. IICSPI* (Chongqing, China) pp 261-4

[17] Rahmaniar W and Rakhmania A E 2017 *J. Robotics Contr.* **2**(4) 234-9
[18] Odelga M, Kochanek N and Bluthoff H, 2017 *Proc. RED-UAS* (Linkoping, Sweden) pp 210-5
[19] Liu P, Huang F, Li G and Liu Z 2012 *IEEE Geosci. Remote. Sens. Lett.* **9** 358-62
[20] Rajapriyadharshini R and Benadict Raja J, 2015 *Proc. ICIIECS* (Coimbatore, India) pp 1-6
[21] Wang H, Wang J, etc, 2016 *EURASIP J. Adv. Signal Process* **2016** 10
[22] Xu S, Zhou Y, Xiang H and Li S 2017 *IEEE Geosci. Remote. Sens. Lett.* **14** 2275-9
[23] Penna P and Mascarenhas N 2018 *Comput. Geosci.* **111** 127-38
[24] Kim B, Kim M and Chae Y 2018 *Sensors* **18** 1
[25] Chang Y, Yang H and Liu H 2014 *IEEE Geosci. Remote Sens. Lett.* **11** 1051-5
[26] Cerra D *et al*. 2015 *Proc. WHISPERS* (Tokyo, Japan) pp 1-4
[27] Xu B, Cui Y, Li Z and Yang J 2015 *IEEE Geosci. Remote. Sens. Lett.* **12** 1635-9
[28] Liu P, Wang M, Wang L and Han W 2019 *IEEE JSTARS* **12** 660-74
[29] Padfield N, Ren J, Murray P and Zhao H 2021 *Neurocomputing* **463** 566-70
[30] Sun G, Fu H, Ren J, Zhang A, Zabalza J, Jia X and Zhao H 2021 *IEEE Trans. Cybernetics* **1** 2168-75
[31] Sun H, Ren J, Zhao H, Yuen P and Tschannerl J, 2021 Novel Gumbel-Softmax trick enabled concrete autoencoder with entropy constraints for unsupervised hyperspectral band selection, *IEEE Trans Geosci. Remote Sens.* D **1558** 644 (*Preprint* gr-qc/3075663)
[32] Fang Z, Ren J, MacLellan C, Li H, Zhao H, Hussain A and Fortino G 2021 *IEEE Trans. Mol. Biol. Multi-Scale Commun.* **14**(8) 1-11
[33] Ren J, Sun H, Zhao H, Gao H, Maclellan C, Zhao S and Luo X, 2021 Effective extraction of ventricles and myocardium objects from cardiac magnetic resonance images with a multi-task learning U-Net, *Pattern Recognit. Lett*. D **167** 8655 (*Preprint* gr-qc/S0167865521003846)
[34] Liu Q, Ren J, Wang Y, Wu Y, Sun H and Zhao H 2021 *Pattern Recognit.* **112** 107766
[35] Sun G, Zhang X, Jia X, Ren J, Zhang A, Yao Y and Zhao H 2020 *Int. J. Appl. Earth. Obs. Geoinf.* **91** 102157
[36] Fang Z, Ren J, Marshall S, Zhao H, Wang S and Li X 2021 *Pattern Recognit.* **109** 107608