Proceedings of CONF-MPCS 2024 Workshop: Quantum Machine Learning: Bridging Quantum Physics and Computational Simulations

DOI: 10.54254/2753-8818/51/2024CH0173

Advances and challenges in UAV navigation based on visual SLAM

Jiamu Liu

School of Energy Systems, Hebei University of Technology, Lappeenranta, Finland

Jiamu.liu@student.lut.fi

Abstract. In recent years, unmanned aerial vehicles (UAVs) have seen extensive use in fields such as agriculture, search and rescue, commercial, and military operations, driving the demand for autonomous navigation capabilities. Though GPS is the traditional method used for navigation, it becomes problematic in harsh environments like deserts. The Visual Simultaneous Localization and Mapping technology offers a solution to enhance UAV navigation in complex environments by constructing maps and localizing the UAV simultaneously in real time. This paper presents state-of-the-art visual SLAM technology developed for UAV navigation regarding algorithms like Oriented FAST and Rotated BRIEF SLAM (ORB-SLAM) and Large-Scale Direct Monocular SLAM (LSD-SLAM), whereby their performance is also discussed, with its positives and negatives. In this regard, the latest progress and challenges in applications are reviewed and analyzed through relevant literature from the databases of PubMed, IEEE, and Google Scholar in the past five years. The novelty of this paper lies in the comprehensive evaluation of the application performance of different visual SLAM algorithms in UAV navigation and the proposal of future research directions.

Keywords: UAV navigation, visual SLAM, environment perception, autonomous navigation, future directions.

1. Introduction

In recent years, UAVs have developed rapidly and have been widely used in agriculture, search and rescue, commercial, military, and other fields [1]. The demand for UAVs is increasing, and so are the quality requirements. The capability of being able to navigate autonomously in complex environments has become a focal point of attention. Navigation is a key capability of UAVs, and traditional navigation relies on GPS, which is greatly limited in extreme environments such as deserts, significantly reducing the autonomous navigation capability of UAVs in these environments [2]. Visual SLAM is a technology that constructs maps and localizes itself simultaneously through visual sensors in an unknown environment. Initially applied in the field of robotics, visual SLAM enables autonomous navigation and obstacle avoidance by constructing maps and localizing in real time. By integrating SLAM technology, UAVs can overcome environmental constraints, achieving more precise navigation capabilities in extreme environments through the localization and mapping capabilities of SLAM [3].

This paper reviews the latest advancements in visual SLAM technology in the field of UAV navigation, analyzes existing major challenges, and explores future development directions. This paper systematically organizes and analyzes key methods and applications for enhancing UAV navigation

DOI: 10.54254/2753-8818/51/2024CH0173

capabilities through SLAM technology by collecting relevant literature on UAVs and SLAM from databases such as PubMed, IEEE, and Google Scholar over the past five years. The structure of this paper is as follows: first, it introduces the principles and current status of visual SLAM, then evaluates several typical SLAM algorithms and analyzes UAV navigation schemes using visual SLAM, and finally discusses the challenges faced by UAV visual SLAM, future research directions, and draws conclusions.

2. Principles and current status of UAV visual SLAM technology

The UAV visual SLAM technology uses cameras or other sensors mounted on the UAVs to collect surrounding environmental information in real time. It achieves localization and map construction through three main steps:

The first step is featuring extraction and matching the UAV camera captures images of the surrounding environment and extracts key visual feature points from these images, such as corners or edges. These feature points typically have unique descriptors used for subsequent matching and localization. Next is the pose estimation. Using position information of the feature points, it estimates the motion of the UAV relative to the last position—that is, the pose of the UAV. This process involves the calculation of the movement of the UAV by way of rotation and translation, that is, the present position and orientation, internal and external parameters of the camera, to be established. Finally, the map is updated. As the UAV moves, new feature points are combined with the existing map and the map is continuously updated. These include map expansion, correction to match real environmental changes, and updates in the position of the UAV.

Visual Simultaneous Localization and Mapping technology for UAVs has been progressing dynamically over the years. The latest version of the ORB series, called ORB-SLAM3, further extends the algorithm to work with monocular, stereo, and RGB-D cameras, making it even more robust and accurate in dynamic environments [4]. Visual SLAM is increasingly making use of deep learning. For example, D3VO uses deep learning to infer depth information and visual odometry to enhance the accuracy and robustness of map construction [5]. The application of deep reinforcement learning (DRL) in SLAM is also gradually increasing. Some studies use deep reinforcement learning algorithms (such as SAC) to enhance UAV autonomous navigation capabilities in dynamic environments, although these methods require longer training times, they exhibit excellent adaptability and obstacle avoidance capabilities [6].

3. Representative visual SLAM algorithms

3.1. ORB-SLAM

Oriented FAST and Rotated BRIEF SLAM (ORB-SLAM) primarily use ORB features for image feature extraction and matching. The main important attributes of ORB features are that they are computationally efficient, have very fast matching speeds, and are invariant to changes of rotation and scale. Feature calculations are smaller in the amount, hence it fits a real-time system.

ORB-SLAM first initializes the map through the extraction of feature points and their descriptors and later uses ORB features to conduct frame-to-frame matching and the estimation of camera poses. The system is implemented on a three-thread architecture: one is the tracking thread, which estimates camera pose in real time and selects keyframes; another is the local mapping thread, which locally optimizes keyframes and map points; and yet another is the loop closure detection thread, which detects and corrects loops. Finally, global optimization ensures the final consistency and accuracy of the map.

This is a feature-based SLAM approach, where ORB feature points are adopted to match frames and close loops, so high precision and robustness can be achieved. Its tri-thread architecture makes it efficient in real time and strong in parallel processing. However, ORB-SLAM suffers from its performance on low-texture scenes and during fast motion and requires high computational resources [4].

DOI: 10.54254/2753-8818/51/2024CH0173

3.2. VINS-Mono

Visual-Inertial Navigation System for Monocular (VINS-Mono) is a visual-inertial navigation technology used for robots, UAVs, and augmented reality devices. It can achieve high-precision, real-time, and robust pose estimation by fusing monocular visual and inertial data. However, its computation complexity is very high and the influencing factors from lighting and a dynamic environment are strong at the same time. With the development of both algorithms and hardware, the technology of VINS-Mono will have more extensive applications and improvements.

VINS-Mono first initializes an initial state through IMU and monocular camera and then performs joint optimization with the image feature points and IMU data in each frame for the estimation of camera pose and velocity. The system is composed of two key parts: front-end and back-end. The front-end extracts the feature points and makes frame-to-frame matching, while the back-end fuses the IMU and visual information using sliding window optimization. In the end, a global Bundle Adjustment further optimizes the pose and map.

The accurate high-precision pose estimation and mapping of a robot are achieved using the VINS-Mono by fusing IMU and monocular camera data concurrently, with high robustness to dynamic environments and high-speed motion. This improves the accuracy of the system, yet it is computationally complex, with low real-time performance and with a large demand on hardware resources [7].

3.3. LSD-SLAM

Large-scale direct Monocular SLAM (LSD-SLAM) is a direct algorithm for monocular SLAM dedicated to real-time localization and mapping in large-scale environments. The two main features of LSD-SLAM are the direct method and semi-dense mapping.

LSD-SLAM first establishes an initial map through a short video sequence initialization process, then estimates camera pose through direct image alignment between consecutive frames, and generates semi-dense depth maps in each keyframe. With the continuous integration of new observations, the depth maps are optimized, and the system constructs a semi-dense 3D map using keyframes as nodes and depth maps as edges. Finally, global optimization (such as pose graph optimization) reduces cumulative errors and ensures global map consistency.

LSD-SLAM uses the direct method to match and optimize image intensity values without requiring feature point extraction, and generates semi-dense maps, improving computational efficiency and robustness in low-texture scenes, suitable for real-time operation in large-scale environments. However, LSD-SLAM is sensitive to lighting changes and motion blur, and long-term operation may produce cumulative errors that need correction through global optimization [8].

3.4. RTAB-Map

Real-Time Appearance-Based Mapping (RTAB-Map), as an advanced SLAM technology, integrates visual appearance information and pose graph optimization, supporting multi-sensor fusion, providing powerful real-time localization and environmental perception capabilities, and offering critical support for autonomous navigation and task execution in complex environments.

RTAB-Map first captures image and depth information using stereo or RGB-D cameras, then extracts and matches feature points using the visual bag-of-words model. This system is separated into front-end visual odometry and back-end loop closure detection and pose graph optimization: front-end on-the-fly pose estimation and local map updating, while the back-end ensures the consistency of the map with global pose-graph optimization through loop-closure detection. A final 3D dense map is then produced.

RTAB-Map, based on the visual bag-of-words model and loop closure detection, has good loop closure detection capabilities and global consistency, suitable for large-scale 3D mapping environments. Its multi-sensor fusion improves robustness but has high computational overhead, relatively weak real-time performance, and high hardware resource requirements [9].

DOI: 10.54254/2753-8818/51/2024CH0173

3.5. Comparison of representative visual SLAM algorithms

In the numerous implementations of visual SLAM algorithms, ORB-SLAM, VINS-Mono, LSD-SLAM, and RTAB-Map have advantages and disadvantages, suitable for different application scenarios. Table 1 is a comparison and analysis of these SLAM algorithms.

Table 1. Comparative analysis of visual SLAM algorithms.

Algorithm	Characteristics	Advantages	Disadvantages	Suitable Scenarios
ORB-SLAM	Uses ORB features for feature extraction and matching, three-thread architecture	High computational efficiency, fast matching speed, strong real-time performance, strong parallel processing capability	Poor performance in low-texture scenes and fast-motion scenarios, high computational resource requirements	Applications requiring high real-time performance and abundant computational resources
VINS-Mono	Fuses monocular vision and inertial data, tight coupling method	High robustness and accuracy, good adaptability in dynamic environments and fast motion	High computational complexity, weak real-time performance, high hardware resource requirements	Complex environment localization and navigation
LSD-SLAM	Direct method using image intensity values for matching and optimization, generates semi-dense maps	High computational efficiency, robustness in low-texture scenes, suitable for large-scale real-time operations	Sensitive to lighting changes and motion blur, long-term operation may produce accumulated errors	Real-time operation in low-texture, large-scale environments
RTAB-Map	Visual-based real-time 3D mapping and localization, supports multi-sensor fusion	Strong loop closure capability and global consistency, suitable for large-scale 3D mapping	High computational overhead, weaker real-time performance, high hardware resource requirements	Scenarios requiring multi-sensor fusion and complex environment perception

3.6. Future technological development of visual SLAM

3.6.1. Unsupervised learning visual SLAM. A general, typical traditional visual SLAM system greatly depends on labelled data, i.e., manually annotated maps for the localization and mapping processes. Unsupervised learning visual SLAM is reducing this dependency by learning and inferring a trajectory of the camera's motion and scene structure from unlabeled data in a self-supervised or weakly supervised way.

3.6.2. Visual SLAM technology in high-dynamic environments. Visual SLAM systems face some problems in high-dynamic environmental conditions, especially very bright lights, shadows, and fast dynamic objects. The technology works under such environments in an effort to increase robustness and performance for accurate estimation of camera motion and scene structure. The system should quickly adapt to dynamic lighting and scenes of high speed, hence raising the environmental perception capabilities by UAVs.

DOI: 10.54254/2753-8818/51/2024CH0173

4. Visual SLAM for UAV navigation

This chapter will detail the practical application and advances of Visual SLAM technology over the last few years, and especially its wide use in UAV navigation. It introduces some popularly used SLAM algorithms, like ORB-SLAM, VINS-Mono, LSD-SLAM, RTAB-Map, their performance in different environments, and some advantages these methods have. Secondly, it presents case studies of the multi-sensor fusion and semantic map-based navigation systems to analyze how technologies have furthered these components' autonomous navigation capabilities while enhancing robustness.

4.1. Practical application of visual SLAM algorithms in navigation systems

One of the most widely used feature-based SLAM algorithms in UAV navigation is ORB-SLAM. In 2019, Gómez-Ojeda et al. tested and demonstrated ORB-SLAM's practical application for UAV autonomous navigation. They proposed to use ORB-SLAM in high-precision map building and localization within an indoor flight environment. Through experiment, the result shows that ORB-SLAM can maintain the localization in a stable and high-precise manner in challenging environments, whereas loop closure detection with ORB-SLAM effectively reduces accumulated errors so that UAVs can navigate with high precision during long flights [10].

The VINS-Mono system is an SLAM-based system, which can tightly integrate the data of the IMU and monocular camera with high accuracy and robustness. Qin and Shen demonstrated UAV navigation by VINS-Mono in 2020 and obtained high-precision localization and path planning of UAVs in a complex outdoor environment [7]. The inclusion of IMU data thus enabled VINS-Mono to provide stable attitude estimation, even when visual information was lost, ensuring its reliability in dynamic and fast-moving environments.

LSD-SLAM is a direct monocular SLAM algorithm able to generate semi-dense 3D maps. In 2021, Zhou et al. explored the application of LSD-SLAM in autonomous flight of UAVs. They tested LSD-SLAM in large outdoor spaces, and in such a situation, the direct method supported better performance that proved to keep the UAV in a stabilized navigation state in low-textured environments [8]. In general, LSD-SLAM is susceptible to changes of illumination and motion blur, and this part was improved in terms of the adaptability of the algorithm in complex environments.

RTAB-Map is a SLAM algorithm based on the visual bag-of-words model and loop closure detection. It finds application in 3D mapping for large-scale environments. Recently, in 2022, Labbe and Michaud applied RTAB-Map to the UAV navigation system. They generated dense 3D maps with a stereo camera capturing images and depth information in such a complex mixed environment [9]. RTAB-Map's loop closure detection feature ensured that the map was consistent and accurate, such that the UAV could offer relatively stable navigation performance during a long flight. Multi-sensor fusion further improved the robustness of the system and showed good performance in complex environments.

4.2. Navigation system case studies

The multi-sensor fusion-based visual navigation system mainly integrates data from cameras, LiDAR, IMU, and some other sensors in order to bring a better robustness and accuracy of the data. Experimental tests in complex indoor environments have shown that such systems enhance the stability of UAV navigation, thereby effectively compensating for the weakness associated with single-sensor data [11].

The paper of Jiedong Zhuang presents an effective way for cross-view matching in geolocating UAVs by using semantic maps to enhance navigation capabilities. To solve this problem, the author has proposed a multi-scale block attention (MSBA) network architecture for the extraction of features across different views and a multibranch structure for exploiting subtle inter-view relationships. Experimental results proved that this method achieves over 10% improvement in accuracy compared to the state-of-the-art methods in the latest dataset of geolocation for UAVs and, at the same time, reduces inference time by 30% [12].

DOI: 10.54254/2753-8818/51/2024CH0173

Table 2. Comparative analysis of navigation system types.

Navigation System Type	Characteristics	SLAM Integration Point	Advantages	Disadvantages	Suitable Scenarios
Multi-Sensor Fusion-Based Visual Navigation System	Integrates multi-sensor data (camera, LiDAR, IMU, etc.)	Combines with SLAM algorithms to improve localization and mapping accuracy	Enhances robustness and accuracy, adapts to different environments, strong navigation stability	High computational complexity, high hardware resource requirements	Complex indoor and outdoor environments
Semantic Map-Based Visual Navigation System	Combines semantic map concept to improve UAV navigation capabilities	Uses SLAM to generate maps and integrates semantic information	Improves navigation accuracy and inference speed	Relies on the accuracy of semantic information, high training data requirements	Geolocation, wide range of applications

Table 2 shows a comparative analysis of the characteristics, integration points with SLAM, advantages, disadvantages, and applicable scenarios of two types of navigation systems.

5. Challenges and future directions

5.1. Challenges

5.1.1. Computational resource limitations. The computational complexity of UAV visual SLAM algorithms is high, requiring extensive processing of visual and sensor data to achieve real-time localization and mapping. Current UAV platforms are often resource-constrained, particularly in terms of computational power and battery life, making it difficult to meet these high computational demands. For instance, algorithms like ORB-SLAM and VINS-Mono require complex image feature extraction and matching, as well as tightly coupled optimization of IMU data. These processes demand efficient computational resources, limiting the application of these algorithms on resource-constrained embedded platforms.

5.1.2. Environmental adaptability. UAVs face numerous challenges in different environments, such as lighting changes, low-texture scenes, dynamic objects, and adverse weather conditions, which can affect the performance and robustness of visual SLAM algorithms. For example, LSD-SLAM performs poorly under lighting changes and in low-texture scenes, and ORB-SLAM encounters difficulties in fast-motion and low-texture environments. These problems have an influence on the ability of the UAVs to adapt in the changing environment, which in turn impacts the final result in terms of accuracy and stability in navigation and localization.

5.1.3. Sensor data quality. Visual SLAM technology is based on high-quality sensor data, including camera images and IMU data. Under practical applications, the sensor data is vulnerable to noise, distortion, and transmission time and can lower data quality. For example, the noise and drift of IMU data affect the fusion accuracy of visual and inertial data, thereby lowering localization and mapping

DOI: 10.54254/2753-8818/51/2024CH0173

performances by VINS-Mono. Cameras may produce blurred images during motion, affecting image feature extraction and matching, thereby impacting the overall performance of SLAM algorithms.

5.2. Future directions

To address these challenges, future research can explore the following directions:

5.2.1. Hardware acceleration and low-power design. This method utilizes dedicated hardware accelerators (such as GPUs and FPGAs) and optimized embedded system designs to improve computational efficiency and reduce power consumption, meeting the resource constraints of UAV platforms. The existence of dedicated hardware can help the processing speed of SLAM algorithms greatly, in turn, largely improve real-time performance and computation efficiency. Besides, low-power design might further enhance flight time for UAVs and thus increase their practicability in real-world applications. However, designing and optimizing hardware architectures is complex, and developing efficient embedded systems requires significant engineering efforts.

5.2.2. Unsupervised learning and adaptive algorithms. Visual SLAM algorithms based on unsupervised learning and adaptive methods are developed to enhance adaptability and robustness in different environments, reducing reliance on labelled data and manual parameter tuning. Unsupervised learning methods can automatically extract environmental features, and adaptive algorithms can adjust SLAM parameters according to environmental changes, improving system robustness. Nonetheless, the effectiveness of unsupervised learning algorithms depends on large amounts of data, and ensuring algorithm stability in different environments remains a challenge.

5.2.3. Multi-sensor fusion. Multi-Sensor Fusion integrates data from multiple sensors (such as LiDAR, depth cameras, IMUs, etc.) to enhance system robustness and data quality, improving UAV navigation capabilities in complex environments. This direction faces challenges such as the complexity of sensor data fusion algorithms and the difficulty of synchronizing data from different sensors in time and space.

6. Conclusion

This paper reviewed the applications of visual SLAM technology in UAV visual navigation, focusing on the performance and advantages of classical algorithms like ORB-SLAM, LSD-SLAM, VINS-Mono, and RTAB-Map in practical scenarios. These algorithms have made significant progress in enhancing UAV autonomous navigation capabilities but still face challenges in environmental adaptability, computational resource requirements, and sensor data quality. Future directions shall concentrate on improving algorithm robustness in highly dynamic environments and optimizing visual SLAM systems' accuracy and real-time performance using deep learning techniques. Additionally, multi-sensor fusion and hardware acceleration designs will improve computational efficiency and energy consumption, further promoting the widespread application of visual SLAM technology in fields such as military, industrial monitoring, and logistics delivery. These studies will provide essential support for UAV autonomous navigation and task execution in complex environments, driving the further development and application of UAV technology.

References

[1] Lu, C., Nnadozie, E., Camenzind, M. P., Hu, Y. and Yu, K. (2024). Maize plant detection using UAV-based RGB imaging and YOLOv5. *Front. Plant Sci.*, 14, 1274813. https://doi.org/10.3389/fpls.2023.1274813. PMID: 38239212; PMCID: PMC10794460.

[2] Xu, Y., Wei, Y., Wang, D., Jiang, K. and Deng, H. (2023). Multi-UAV path planning in GPS and communication denial environment. *Sensors (Basel)*, 23(6), 2997. https://doi.org/10.3390/s23062997. PMID: 36991708; PMCID: PMC10057094.

DOI: 10.54254/2753-8818/51/2024CH0173

[3] Chen, C. L., He, R. and Peng, C. C. (2022). Development of an online adaptive parameter tuning vSLAM algorithm for UAVs in GPS-denied environments. *Sensors (Basel)*, 22(20), 8067. https://doi.org/10.3390/s22208067. PMID: 36298416; PMCID: PMC9610021.

[4] Campos, C., Elvira, R., Rodríguez, J. J. G., Montiel, J. M. M. and Tardós, J. D. (2021). ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multi-map SLAM. *IEEE Trans. Robot.*, 37(6), 1874–1890. https://doi.org/10.1109/TRO.2021.3075644.

[5] Yang, N., Rosa, S. Z., Wang, L., Von Stumberg, L. and Cremers, D. (2020). D3VO: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* (pp. 1281–1292). https://doi.org/10.1109/CVPR42600.2020.00139.

[6] Han, J., Sun, T. and Liu, L. (2021). A deep reinforcement learning approach for autonomous UAV navigation in dynamic environments. *Sensors*, 21(1), 252. https://doi.org/10.3390/s21010252.

[7] Qin, T., Li, P. and Shen, S. (2018). VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Trans. Robot.*, 34(4), 1004–1020. https://doi.org/10.1109/TRO.2018.2853729.

[8] Zhou, B., Zhang, T. and Li, Z. (2020). Improved LSD-SLAM algorithm for monocular vision positioning of drones. *Sensors*, 20(21), 6128. https://doi.org/10.3390/s20216128.

[9] Labbe, M. and Michaud, F. (2022). RTAB-Map as an open-source lidar and visual SLAM library for large-scale and long-term online operation. *J. Field Robot.*, 39(2), 167–192. https://doi.org/10.1002/rob.22045.

[10] Gomez-Ojeda, R., Moreno, F.-A., Zuñiga-Noël, D., Scaramuzza, D. and Gonzalez-Jimenez, J. (2019). PL-SLAM: A stereo SLAM system through the combination of points and line segments. *IEEE Trans. Robot.*, 35(3), 734–746. https://doi.org/10.1109/TRO.2019.2899783.

[11] Chang, Y., Cheng, Y., Murray, J., Huang, S. and Shi, G. (2022). The HDIN dataset: A real-world indoor UAV dataset with multi-task labels for visual-based navigation. *Drones*, 6, 202. https://doi.org/10.3390/drones6080202.

[12] Dong, C. et al. (2019). A novel hierarchical control strategy for biped robot walking on uneven terrain. In *Proc. IEEE-RAS 19th Int. Conf. Humanoid Robots (Humanoids)* (pp. 140–145). https://doi.org/10.1109/Humanoids43949.2019.9035039.