Vision-Inertial Fusion Pose Estimation for UAVs in Complex Environments: A Review and Comparative Analysis

Huazhen Wei

{whz18021022657@qq.com}

Adelaide University, School of Electrical and Mechanical Engineering, Adelaide, SA 5095, Australia

Abstract. Drone attitude estimation is very sensitive to blocked satellite navigation signals and sensor errors in complex environments. So, methods for autonomous attitude estimation using onboard sensors have gotten a lot of attention. Visual sensors and inertial measurement units complement each other well in the information they provide. Because of this, visual-inertial fusion has become a key research area for drone attitude estimation. This paper reviews and compares UAV visual-inertial fusion attitude estimation methods for complex environments. It organizes typical fusion frameworks based on filtering and optimization. It focuses on the modeling ideas, IMU pre-integration methods, and the features of tightly coupled and loosely coupled fusion strategies in optimization-based methods. Then, under hard visual conditions like low texture, changing light, and dynamic scenes, it looks at how stable and practical different fusion strategies are. The paper lists the main challenges and future directions for current visual-inertial fusion methods.

Keywords: Drone, attitude control, fused attitude estimation

1 Introduction

Drones are now used in many complex jobs like surveying, inspection, environmental monitoring, and emergency rescue. The need for autonomous flight is growing. This need makes stable and reliable attitude estimation very important. In complex places like indoors, city streets, or areas with strong electromagnetic interference, satellite navigation signals are often blocked or fail. So, traditional navigation that needs external positioning does not work well here. Because of this, autonomous attitude estimation using onboard sensors is a key part of drone navigation and control systems. How well it works directly affects drone safety and mission success in complex environments.

Visual sensors and Inertial Measurement Units (IMU) are the two most common sensors on drones. Their information is very complementary. IMU updates very fast and are accurate over short times. But their measurement errors build up over time, causing drift in the attitude estimate. Visual sensors give information about the outside world and are stable over long times. But their performance is very sensitive to light changes, environmental textures, and movement. Studies show one sensor alone often cannot be both accurate and robust in complex

ICMEEA 2026, April 24-26, Singapore, Singapore

DOI 10.4108/eai.24-4-2026.2364969

environments [1]. This has pushed the development of visual-inertial fusion methods. The main methods are based on filtering or optimization. Filtering methods, like the Extended Kalman Filter (EKF), fuse information through recursive estimation. The Multi-State Constraint Kalman Filter (MSCKF) is a good example [2]. Optimization methods use a nonlinear framework to jointly estimate observations within a time window. They can be more accurate and consistent [3]. In optimization methods, IMU pre-integration is often used to handle high-frequency data without heavy computation [4]. Also, different fusion strategies work differently when visual quality is poor. Conditions like low texture, light changes, or dynamic scenes can weaken the system's geometric constraints. This hurts the stability of the fusion system [5].

Because of this, this paper focuses on vision-inertial fusion for UAV attitude estimation in complex environments. It reviews and compares related technical methods. It focuses on the modeling ideas of different fusion frameworks and strategies. The paper analyzes their performance and limits under poor visual conditions. It summarizes the main challenges in robustness and practical use faced by current methods. It also outlines future research trends. This analysis can help in choosing UAV attitude estimation methods and designing better autonomous navigation systems.

2 Classification of Visual-Inertial Fusion Pose Estimation Problems and Methods

UAV attitude estimation is a basic part of navigation and control systems. It estimates the aircraft's attitude steadily using data from multiple sensors. A single sensor often cannot be both accurate and robust in complex environments. So, fusing visual and inertial information is a key research direction. Visual sensors provide constraints from the environment to reduce long-term drift. Inertial measurement units update fast and are good for short-term accuracy. The two types of information help each other a lot.

Current visual-inertial fusion methods mainly fall into two groups: filter-based and optimization-based. Filter-based methods usually use the extended Kalman filter. They predict and update the system state step by step. A key work here is the Multi-State Constraint Kalman Filter(MSCKF)[2]. It has a lower computational cost and is easier to run in real time. But the filtering framework uses linear approximations. Also, old observation data is compressed during updates. If there are big model errors or observation noise, system consistency can be a problem.

As computers get better, visual-inertial fusion based on nonlinear optimization has become more popular. These methods treat the system states in a time window as variables to optimize together. They model visual and inertial observations as a nonlinear least-squares problem[3]. In this framework, IMU pre-integration is widely used. It turns high-frequency inertial data into constraints between keyframes. This lowers the computational cost[4]. Also, based on how closely visual and inertial data are combined, methods can be tightly coupled or loosely coupled. Their performance differs a lot in complex environments.

Both filtering and optimization frameworks depend on good visual observations. When conditions have low texture, big light changes, or dynamic scenes, visual observations get worse. This weakens the system's geometric constraints. It then limits pose estimation accuracy and system stability[5].

3 Filter-Based Visual-Inertial Fusion Pose Estimation Method

The filter-based method is an early approach for visual-inertial pose estimation. The main idea is to estimate the system state over time recursively. It uses visual observations to correct predictions from the inertial data. These methods are often based on the extended Kalman filter. They have a clear structure, relatively low computational cost, and are easy to run in real time. These advantages made them popular early on[6].

In the filtering framework, the system state usually includes the vehicle's attitude, velocity, position, and sensor errors. The IMU gives high-frequency state predictions. Visual observations are used as external measurements to update these predictions. This way, filtering methods can reduce the build-up of inertial errors. It improves the stability of the attitude estimate. But a problem with traditional filtering is that adding many visual features directly makes the state very large. This hurts real-time performance and numerical stability.

The Multi-State Constraint Kalman Filter (MSCKF) was made to solve this. Its key idea is the multi-state constraint. It uses visual observations from different times to constrain the system state. It does not explicitly estimate where feature points are. This implicit way of handling features controls the state size. It adds more visual constraint information but keeps the filter running in real time. Tests show that compared to filters using only inertial data, MSCKF greatly reduces attitude drift. It gives more stable estimates for short to medium times.

A strength of MSCKF and similar filtering methods is they are simple to make and efficient. This is good for UAV platforms with little computing power or high real-time needs. Also, because filters process data step by step, their system structure is compact. They are easier to build, which was valuable in early UAV visual-inertial systems.

But tests of MSCKF also show the limits of filter-based methods in complex environments. Because they use linear approximations during updates, the estimates are sensitive to model errors and measurement noise. So, when visual observations are bad or the environment changes a lot, the filter cannot use old information well to correct estimates. This can hurt consistency. Also, because filters only update the current state at each step, they cannot use long-term constraint information well. This becomes a problem for high-precision pose estimation tasks.

Filter-based methods are important in the history of visual-inertial pose estimation. Work like MSCKF showed that adding visual information can control inertial drift. It helped later research. But as environments get more complex and the need for high precision grows, the limits of filtering methods in using information and keeping estimates consistent have become clearer. This has helped drive the development of optimization-based methods.

4 Optimization-Based Visual-Inertial Fusion Pose Estimation Method

4.1 Fundamental Principles of Optimization-Based Methods

The filter-based fusion method is an early and common technique. Its core idea is to predict the system state in a recursive framework. It then updates the inertial predictions with visual observations. It usually uses the extended Kalman filter. It has a clear structure and low computational cost. It fits UAV platforms that need real-time performance and have limited computing power.

In the filtering framework, the system state includes attitude, velocity, position, and sensor errors. The IMU gives high-frequency state predictions. Visual observations correct these predictions. This reduces the build-up of inertial errors. But putting many visual features directly into the state makes the state very large. This hurts real-time performance and numerical stability.

The Multi-State Constraint Kalman Filter is a well-known filtering method. It uses the idea of multi-time constraints. It uses visual observations across time to constrain the system state without estimating feature point positions. This controls the state size well. It also adds more visual information in real time. Compared to filters using only inertial data, MSCKF reduces attitude drift a lot. It gives stable estimates for short to medium times.

Filtering methods are useful in practice. But they have limits. For example, they are sensitive to linear approximations, noise models, and observation quality. When visual observations are poor or the environment changes a lot, they cannot use old information well. This may hurt consistency.

4.2 Keyframe-based Optimized Fusion Method OKVIS

The visual-inertial fusion method based on nonlinear optimization is now a hot research topic. The OKVIS system was first presented in 2015 by Leutenegger and others from ETH Zurich[3]. It is a key work that first showed optimization-based visual-inertial fusion is possible. OKVIS controls the state size by managing keyframes. It adds visual and inertial constraints across time within the window. This lets the system make global corrections to the pose estimate during optimization.

But OKVIS also shows the cost of optimization methods in practice. They have higher computational cost. They are more sensitive to how they are started and how parameters are set. Their performance drops a lot when visual observations are bad. The optimization framework does not automatically work well in complex environments. So, more work is needed to make them more robust and practical.

4.3 Optimized Fusion System VINS-Mono for Unmanned Aerial Vehicle Scenarios

After early work like OKVIS showed optimization works, researchers like Qin Tong made the VINS-Mono system [7]. It deals with problems like unknown scale and unstable start-up with a single camera. It was designed to meet the stability needs for real UAV flights.

VINS-Mono uses a sliding window optimization framework. It makes the system much more reliable for real flight tasks. It does this by using inertial data to find the scale, improving the marginalization strategy, and adding failure detection and re-start methods. Tests show the system can give continuous, stable pose estimation on public datasets and real UAV platforms [7].

Later research started to focus on working well in complex environments. The goal is to make optimization-based systems better for dynamic scenes and poor observations [8]. Some work has reduced the effect of dynamic interference on the system. It does this by making constraints more robust or adding ways to suppress problems [9][10]. But when visual observations are bad for a long time or environmental interference continues, the performance of traditional feature-based methods still falls. Also, high computational cost limits their use on platforms with little computing power [9][11].

4.4 The Role of IMU Pre-Integration in Optimization-Based Fusion Methods

High-frequency inertial data is key for making optimization frameworks work better. But it also brings modeling and computational challenges. Putting each inertial measurement directly into the optimization would make the state too large. This is not practical.

The main idea of IMU pre-integration is to combine inertial data between keyframes. It adds this as relative motion constraints to the optimization. This greatly lowers the computational cost [4]. This idea gives a solid base for using inertial data in a discrete optimization framework. It also helped it be used widely in systems like OKVIS and VINS-Mono [3][7].

Current pre-integration depends on error state modeling and linearization. If not done right, it can hurt estimation consistency [8]. In complex environments, pre-integration helps keep pose estimation going when vision is briefly bad. But it cannot remove the long-term build-up of inertial errors. It is still limited by sensor performance [1]. Later research has improved pre-integration theory using Lie group manifolds. This has made it more consistent and stable under high-dynamic conditions [12].

5 Comparative Analysis of Fusion Methods and Discussion of Key Issues

Different methods often work well in ideal conditions. But their performance differences show up more as conditions get complex.

The fusion frameworks themselves are fundamentally different. Filtering methods are based on recursive estimation. They are clear and work in real time well. But old observation data is compressed during updates. So, early errors that get into the system are hard to correct fully. Optimization methods estimate states for multiple time steps together in a time window. Old observations keep acting as constraints during estimation. This uses information more efficiently. But this advantage needs accurate error modeling and linearization. If the modeling is wrong, it can cause biases [13].

In recent years, research has looked at the consistency problem of traditional fusion methods. The FEJ2 method notes that bad Jacobian linearization in nonlinear systems can break estimation consistency. Fixing the initial estimation point can help this [14]. Also, from an error state view, some studies use a right-invariant error representation. This makes attitude error propagation fit the system's geometry better. It improves the stability of visual-inertial systems over long runs [15]. These studies show that how well fusion methods work depends not just on the framework choice, but also deeply on how error states are modeled.

In practice, optimization-based visual-inertial fusion has long had a problem. The covariance results are not reliable enough. Because of sliding windows and marginalization, the system often cannot get good uncertainty estimates. Chen and others recently made a covariance recovery method to fix this [16]. It lets the state uncertainty in a sliding window framework be estimated well. This makes the system easier to understand and more reliable in practice. Based on this, the same team also suggested decoupling error states from system states at the modeling level [17]. This unifies the view of error representation for filtering and optimization methods. It gives a new way to check the consistency of different fusion algorithms.

Besides theory and practical understanding, poor visual conditions in complex settings remain a key limit on fusion system performance. When there is not enough texture, big light changes, or many moving objects, visual constraints based on geometric features become unstable. To fix this, researchers are trying more adaptive modeling. Adaptive VIO methods let the system change the visual model while running through online learning. This improves how well the system adapts to changing environments [13]. Other methods add neural radiance fields to the visual-inertial framework. They use dense scene representations to help when visual constraints are weak in low-texture areas. This shows that learned geometric representations can help in complex environments [18].

As more fusion algorithms for complex environments appear, system evaluation has become more important. The SMARTNav dataset gives a common test platform for visual-inertial state estimation in complex conditions [19]. It makes comparing how well different methods work in dynamic scenes and poor conditions more repeatable and fair. Such datasets show a change in research focus. They also provide a key base for designing and evaluating future algorithms.

In short, work on visual-inertial fusion pose estimation is changing. It is moving from improving single algorithms to also thinking about error modeling consistency, system robustness, and adapting to complex environments. Future research must balance theory and practical use more closely. It should also explore combining geometric models and learned information better. This will improve how well drones can work on their own in complex places.

6 Conclusion

This paper reviews and compares methods for visual-inertial fused UAV pose estimation in complex environments. It looks at the features of visual and inertial sensors and how they complement each other. It also sums up the main research progress of current visual-inertial fusion pose estimation methods. The paper focuses on the performance differences between filtering and optimization methods in complex environments and why these differences happen. As computers get better, optimization-based visual-inertial fusion methods have become better at using information and giving consistent estimates. They are now the main approach in this field. But current methods still have many problems with poor vision and error modeling consistency. The overall system stability and robustness need to be improved more.

The summary and comparison of typical methods and key works in this paper help us understand when different fusion strategies work and their limits. This gives a reference for choosing UAV attitude estimation methods and designing systems in future work. Research on visual-inertial fusion for attitude estimation is also likely to move forward. It may lead to better robust error modeling, adaptive methods, and combining learned information with geometric models. This could further improve how well UAVs can fly on their own in complex environments.

References

[1] Titterton, D. H., Weston, J. L.: Strapdown inertial navigation technology (2nd ed.). The Institution of Electrical Engineers, London (2004)

[2] Mourikis, A. I., Roumeliotis, S. I.: A multi-state constraint Kalman filter for vision-aided inertial navigation. In: Proceedings of the 2007 IEEE International Conference on Robotics and Automation (ICRA), pp. 3565–3572. IEEE (2007)

[3] Leutenegger, S., Lynen, S., Bosse, M., Siegwart, R., Furgale, P.: Keyframe-based visual-inertial odometry using nonlinear optimization. The International Journal of Robotics Research. 34(3), 314–334 (2015)

[4] Forster, C., Carlone, L., Dellaert, F., Scaramuzza, D.: IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation. In: Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), pp. 4934–4941. IEEE (2015)

[5] Kerl, C., Sturm, J., Cremers, D.: Dense visual SLAM for RGB-D cameras. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2100–2106 (2013)

[6] Martinelli, A.: Vision and IMU data fusion: Closed-form solutions for attitude, speed, absolute scale, and bias determination. IEEE Transactions on Robotics. 28(1), 44–60 (2012)

[7] Qin, T., Li, P., Shen, S.: VINS-Mono: A robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics. 34(4), 1004–1020 (2018)

[8] Huang, G.: Visual-Inertial Navigation: A Concise Review and Perspectives. Annual Review of Control, Robotics, and Autonomous Systems. 5, 261–294 (2022)

[9] Samadzadeh, A., Nickabadi, A.: SRVIO: Super Robust Visual Inertial Odometry for Dynamic Environments and Challenging Loop-closure Conditions. arXiv preprint arXiv:2201.05386 (2022)

[10] Li, J., Pan, X., Huang, G., Zhang, Z., Wang, N., Bao, H., Zhang, G.: RD-VIO: Robust Visual-Inertial Odometry for Mobile Augmented Reality in Dynamic Environments. IEEE Transactions on Visualization and Computer Graphics (TVCG). (2024)

[11] Che, R., Hu, H., Wang, N., Li, Z.: Visual-Inertial Odometry with deep structural priors. Measurement. 259(Part A), 119369 (2026)

[12] Forster, C., Carlone, L., Dellaert, F., Scaramuzza, D.: On-Manifold Preintegration for Real-Time Visual-Inertial Odometry. IEEE Transactions on Robotics. 33(1), 1–21 (2017)

[13] Pan, Y., Zhou, W., Cao, Y., Zha, H.: Adaptive VIO: Deep Visual-Inertial Odometry with Online Continual Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18019–18028 (2024)

[14] Chen, C., Yang, Y., Geneva, P., Huang, G.: FEJ2: A Consistent Visual-Inertial State Estimator Design. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). IEEE (2022)

[15] Yang, Y., Chen, C., Lee, W., Huang, G.: Decoupled Right Invariant Error States for Consistent Visual-Inertial Navigation. IEEE Robotics and Automation Letters. 7(2), 1627–1634 (2022)

[16] Chen, C., Peng, Y., Huang, G.: Fast and Consistent Covariance Recovery for Sliding-window Optimization-based VINS. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). IEEE (2024)

[17] Chen, C., Peng, Y., Huang, G.: Visual-Inertial State Estimation with Decoupled Error and State Representations. In: International Workshop on the Algorithmic Foundations of Robotics (WAFR) (2024)

[18] Katragadda, S., Lee, W., Peng, Y., Geneva, P., Chen, C., Guo, C., Li, M., Huang, G.: NeRF-VINS: A Real-time Neural Radiance Field Map-based Visual-Inertial Navigation System. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). IEEE (2024)
[19] Mirtajadini, S. H., Mersha, A. Y., Aldibaja, M.: SMARTNav: A Visual-Inertial Dataset for Reliable Robotic State Estimation. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences. X-2/W2-2025, 149–155 (2025)