2025 International Conference on Advanced Mechatronics and Intelligent Energy Systems

Research on Obstacle Avoidance and Path Planning for UAVs based on Visual Algorithms

AIPCP25-CF-AMIES2025-00084 | Article

Research on Obstacle Avoidance and Path Planning for UAVs based on Visual Algorithms

Shangjiong Lu

Southampton Ocean Engineering Joint Institute, Harbin Engineering University, Harbin, Heilongjiang province, 150001, China

lushangjiong@hrbeu.edu.cn

Abstract. With the growing demand for drone applications in special environments such as forest fires, earthquake disasters, and complex urban terrain, improving their autonomous obstacle avoidance and path planning capabilities in dynamic environments has become a core issue in current research. This paper systematically reviews the research progress of four typical visual algorithms in this field, covering target detection (YOLO series) and environmental perception and localization (SLAM). On the basis of introducing the principles of various algorithms, this paper focuses on comparing their performance in terms of real-time, environmental adaptability, modeling accuracy, and deployment efficiency, and further points out their typical advantages and disadvantages in complex dynamic scenes: For example, the YOLO series has extremely high frame rate and small target detection capabilities in dynamic target recognition, but has limited robustness under weak light or occlusion conditions; the SLAM algorithm can achieve high-precision positioning and mapping in unknown scenes, but it is prone to drift and failure when the texture is sparse or dynamic objects interfere. Finally, this paper further explores the development trend of the perception-decision integrated path planning system, and looks forward to future research directions such as multi-sensor fusion, lightweight model design, and multi-UAV collaborative obstacle avoidance.

INTRODUCTION

In recent years, global warming and frequent extreme weather have led to a significant increase in the frequency and intensity of forest fires [1]. Forest ecosystems play a key role in carbon balance and biodiversity, and their safety is related to human sustainable development [2]. Traditional ground monitoring methods, such as watchtowers and fixed-wing aircraft, are difficult to meet the rapid response needs in disaster scenarios due to terrain obstruction and limited mobility [2]. Unmanned aerial vehicles (UAVs) have demonstrated good data acquisition capabilities in complex and high-risk environments with their flexible deployment and low-risk advantages [3]. However, traditional path planning methods face severe challenges in environments with dense smoke, heat flow, and dynamic obstacles [4].

The three existing mainstream algorithms each have their advantages and disadvantages. RRT has good global search capabilities in high-dimensional space, but it responds slowly in dynamic scenarios. Liu et al. introduced Dubins curve optimization to reduce redundant nodes by 35%. Zhao et al. proposed a double sampling mechanism to reduce path cost by 38.32% and time by 71.22%, but it still requires frequent global resampling [5, 6]. A* can generate approximate optimal paths using heuristic functions, but the computational burden is large in high-dimensional space, and the path often contains multiple turns. EBS-A* proposed by Wang et al. improves planning efficiency by 278% and completely eliminates turns [7]. Guo et al. introduced Bezier curves to reduce turning points by 46.15%, but the algorithm relies on static heuristic functions and lacks the ability to adapt to instantaneous changes [8]. The APF algorithm has excellent real-time performance. Pan et al. used the rotating potential field to increase the success rate of multi-aircraft formation re-planning to 92.5%. Hao et al. combined collision assessment to improve efficiency by 24.6% and reduce energy consumption by 15.7%, but it is still difficult to maintain robustness under conditions of perception uncertainty [9, 10].

In summary, although the performance of traditional algorithms has been significantly improved in static environments, it is difficult to cope with the complex dynamic characteristics of disaster sites. Visual algorithms provide a new path. YOLOv5n achieves 102 FPS and mAP50 of 76.7% on the Jetson platform, and has high frame rate target detection capabilities [11]. NICER-SLAM reduces ATE-RMSE to 1.88 cm, and SLAM3R can maintain 20+ FPS and effectively suppresses drift by relying only on RGB video [12]. The integration of target detection and centimeter-level mapping capabilities provides a basis for real-time and accurate path reconstruction for UAVs in dynamic environments.

This paper will systematically sort out the research context of YOLO and SLAM in UAV obstacle avoidance, conduct a quantitative comparison based on a unified data set, and explore the development trend of their integration and optimization. It aims to clarify the applicable boundaries and performance bottlenecks of different visual algorithms in complex dynamic scenes, and provide methodological reference and technical support for the subsequent construction of an efficient and robust UAV path planning system.

ANALYSIS OF MAINSTREAM VISUAL ALGORITHMS

YOLO Series Algorithms

Basic Concepts

YOLO is a single-stage target detection framework, which is widely used in real-time perception tasks of drones due to its fast reasoning speed and lightweight structure.

You Only Look Once (YOLO) is a single-stage object detection framework that compresses the inference latency to milliseconds by dividing the input image into a regular grid and regressing the bounding box and category probability at one time [13]. Thanks to its end-to-end design, lightweight convolutional backbone, and parallel decoding head, YOLO can maintain ≥60 FPS on embedded platforms such as Jetson Orin Nano, which is very suitable for drones' needs for low latency and high frame rate perception. From v1 to v8, the algorithm has introduced improvements such as the CSPDarknet backbone, Anchor-Free prediction, and automatic data enhancement, which significantly improved the speed-accuracy trade-off [13]. In tasks such as forest fire monitoring and urban canyon obstacle avoidance, YOLO can output the location information of smoke, fire points, or dynamic obstacles in real time, provide semantic constraints for downstream SLAM mapping and local path planning, and realize online decision-making and safe avoidance during flight.

Experimental Data and Performance Comparison

In order to verify the detection capability of YOLOv8 in complex environments, Table 1 shows the performance comparison results of the algorithm on the VisDrone2019 dataset. By comparing with the YOLOv8s baseline model, the improvement in indicators such as precision, recall, F1 score, and mAP@0.5 is analyzed.

TABLE 1. Comparison of detection performance of the improved YOLOv8 model on the VisDrone2019 dataset

Model	Accuracy	Recall	F1 score	mAP@0.5
UAV-YOLOv8 [4]	54.4 %	45.6 %	49.6 %	47.0 %
YOLOv8 s(baseline)	50.9 %	38.2 %	38.9 %	39.3%

UAV-YOLOv8 uses YOLOv8s as the baseline, introduces WIoU v3 loss and BiFormer sparse attention, and expands the three-scale detection to five-scale to strengthen the representation of small targets. VisDrone2019 scenes cover complex backgrounds such as urban high-rise buildings, roads, and low light at night. The improved model improves Precision and Recall by 3.5 pp and 7.4 pp, respectively, and mAP@0.5 increases by 7.7 pp. For drone inspections, this means that small targets such as pedestrians and vehicles can still be reliably captured in the gaps between high-rise buildings or in dense traffic flows, providing more accurate obstacle positioning for subsequent path planning.

UAD-YOLOv8 is based on YOLOv8n. It first deletes the high-level P5 feature layer and then uses C2f-DCNv2 to adapt deformation and occlusion. Then it replaces the standard convolution with UGDConv and Lw-Detect lightweight detection head, achieving parameters –77% and GFLOPs –34%. Despite the significant reduction in computational complexity, the model still maintains more than 80% in Precision, and mAP@0.5 is increased to 80.3%. This ensures that the drone can stably identify and track moving obstacles at a speed of ≈90 FPS in dynamic

environments such as forest areas or power inspections, laying the foundation for real-time obstacle avoidance and track replanning.

Next, in the UAD-YOLOv8 experimental dataset, this experiment used the UAVDT and VisDrone datasets, the latter of which is a large-scale drone image dataset that contains a variety of flight scenes and obstacle types. By adding deformable convolution (DCNv2) and lightweight modules (such as UGDConv), the model reduces the computational burden while improving detection accuracy. Although the recall rate is slightly lower (78.2%), the precision rate is close to 90%, indicating that the model can avoid misidentification in most scenarios and can effectively identify moving obstacles. In actual flight, real-time reasoning performance is crucial. The addition of DCNv2 and UGDConv greatly improves the efficiency of the model and meets the real-time detection requirements of drones. To further evaluate the performance of the improved YOLOv8 model in multi-class obstacle recognition, Table.2 shows the comparison of its detection effect on the UAD obstacle dataset, covering key indicators such as Precision, Recall, F1 score, and mAP@0.5.

TABLE 2. Comparison of detection performance of the improved YOLOv8 model on the UAD obstacle dataset

Model	Precision	Recall	F1 %	mAP@0.5 %
UA D -YOLOv8 [11]	80.7 %	73.8 %	77.1 %	80.3 %
YOLOv8 n （baseline）	80.4 %	71.9 %	75.9 %	76.9%

The data is extracted from the control experiment of the UAD dataset built by the author (Table 2). The UAD dataset contains 3636 obstacle images, covering multiple types of obstacles such as trees, electric poles, vehicles, pedestrians, etc. Experimental data show that the YOLOv8 series of algorithms has high accuracy and good real-time performance in target detection tasks, especially when dealing with obstacles in dynamic environments (such as moving fire sources, smoke, etc.), which can effectively reduce the occurrence of false detection and missed detection. In the UAV obstacle avoidance and path planning tasks, YOLOv8 not only improves the accuracy of obstacle recognition but also helps UAVs plan safer and more effective flight paths through high-precision target positioning.

Limitations and Challenges

Although UAV-YOLOv8 increased mAP0.5 to 47.0% on VisDrone2019, its FLOPs increased from 28.7G to 53G, and the detection rate of extremely small and single-texture targets (such as bicycles) is still low. In low-light and backlit scenes, the decrease in feature contrast leads to a significant increase in the missed detection rate, and the mAP0.5 of small targets decreases by 8-12 pp. In addition, when inferring at high resolution, both the video memory usage and inference latency will double, making it difficult for embedded GPUs to achieve real-time processing. At the same time, algorithm training is highly dependent on large-scale, cross-view annotation data, otherwise, it is easy to overfit to a specific view.

Application of Visual SLAM in Autonomous Navigation and Positioning Mapping

Basic Concepts

Simultaneous Localization and Mapping is a key algorithm for simultaneous localization and map construction in unknown environments, and is widely used in autonomous navigation tasks of drones. Through sensors such as cameras, IMUs, or lidars, SLAM systems can estimate their own positions and build environmental models in real time during flight, providing spatial support for path planning and obstacle avoidance. Among various perception methods, visual SLAM has become a common solution in drone missions due to its lightweight and low cost. Its basic principle is to extract features from continuous image sequences, infer pose changes, and gradually generate sparse or dense environmental maps to help drones achieve accurate navigation in GPS-deficient or complex environments [14, 15].

Experimental Data and Performance Comparison

In order to systematically compare the comprehensive performance of different dense SLAM algorithms on multiple data sets, Table 3 summarizes the pose error, reconstruction quality, and real-time indicators of SLAM3R

and mainstream algorithms on 7Scenes and Replica data sets, covering key dimensions such as ATE-RMSE, Accuracy/Completeness, and FPS.

TABLE 3. Comparison of pose error, reconstruction quality, and real-time performance of dense SLAM algorithms on 7Scenes and Replica datasets

Method	7scenes_RMSE (cm)	Repliac_RMSE (cm)	7scenes_Acc/Comp	Repliac_Acc/Comp	FPS
NICER-SLAM [12]	8.55	1.88	3.65 / 4.16	3.65 / 4.16	<1
DROID-SLAM	5.66	0.33	5.66 / 11.70	5.50 / 12.29	~20
SLAM3R-NoConf	8.44	6.61	2.40 / 2.24	3.76 / 2.62	~24
SLAM3R	8.41	6.61	2.13 / 2.34	3.57 / 2.62	~24

In the comparative experiment of SLAM3R, Liu et al. selected two commonly used 3D scene datasets: 7Scenes and Replica, and compared and evaluated the current mainstream dense SLAM systems, including NICER-SLAM, DROID-SLAM, and its two self-developed versions: SLAM3R-NoConf and SLAM3R. The experiment quantitatively evaluated the pose estimation error (ATE-RMSE), the accuracy and completeness of point cloud reconstruction (Accuracy / Completeness), and the system operation efficiency (FPS), and fully verified the comprehensive performance of SLAM3R in 3D reconstruction tasks.

First, in the NICER-SLAM experiment, Zhu et al. used voxel hashing and implicit surface reconstruction. The lowest RMSE on the Replica dataset was 1.88 cm, but the FPS was always <1, which was insufficient for real-time performance [12]. Then, in the DROID-SLAM experiment, Teed & Deng used a recurrent neural network to estimate and reconstruct the scene through a joint optimization process. The RMSE in the Replica dataset was 0.33 cm, and the Completeness reached 12.29 cm, the highest among all methods. The actual GPU measurement was about 20 FPS [14]. Subsequently, the RMSE of SLAM3R-NoConf on 7Scenes and Replica was 8.44 cm and 6.61 cm, respectively, but it still maintained a high reconstruction consistency and structural coherence. FPS≈24. After further introducing the confidence gating mechanism, SLAM3R-NoConf (SLAM3R) maintained an average RMSE of 8.41 cm/6.61 cm, but the reconstruction accuracy was improved to 2.13/2.34 cm (7Scenes) and 3.57/2.62 cm (Replica), indicating that this mechanism significantly reduced the introduction of low-quality points, thereby improving the model stability without relying on global optimization.

The full-process GPU parallelism enables SLAM3R to achieve 24FPS real-time dense reconstruction on RTX 4090 without camera calibration, and can provide ±10 cm, ≥20 FPS positioning and mapping support for indoor drone inspections, post-disaster search and rescue, and other GPS-free scenarios, which is better than traditional graph optimization SLAM methods. SLAM3R shows extremely high practicality, especially in high-frequency dynamic scenes or resource-constrained platforms (such as drones or embedded devices).

Overall, there are many types of SLAM algorithms. The mainstream methods can be divided into sparse mapping (such as ORB-SLAM2) and dense reconstruction (such as SLAM3R). Both have their advantages in the application of UAV path planning [15]. ORB-SLAM2 adopts a sparse mapping method based on feature points. The system structure is lightweight and has strong real-time performance. It is suitable for resource-constrained platforms. Its positioning mean square error (RMSE) on the TUM dataset can reach 4.4 cm, with good accuracy and closed-loop capability [16]. In recent years, end-to-end dense visual SLAM methods such as SLAM3R have been developed. They complete local point cloud reconstruction and global fusion through deep neural networks, and achieve 2.13 cm reconstruction accuracy (Acc.) and 24 FPS real-time performance on the 7Scenes dataset. It performs better in point cloud integrity and visual understanding ability, but global drift and increased reasoning delay still occur when deployed on weak textures, high dynamic flight, and Jetson-level platforms; the robustness can be improved through lightweight posture supervision and multi-source sensor fusion [15]. In summary, different SLAM algorithms have different focuses on mapping accuracy, system efficiency, and path planning support capabilities, and need to be flexibly selected in combination with UAV application requirements.

Limitations and Challenges

In scenes with sparse textures or repeated patterns, the front-end matching robustness of the visual SLAM system decreases, which can easily lead to a cumulative trajectory drift of 9 cm within 5 minutes; high-speed maneuvering flight amplifies the timing error between the IMU and vision, causing frequent relocalization failures. In addition, most algorithms assume that the environment is static by default and cannot model dynamic targets. They need to rely on semantic occlusion culling mechanisms to maintain map consistency and navigation stability.

FUTURE DIRECTIONS

Although current visual algorithms have made significant progress in drone obstacle avoidance and path planning, they still face the triple challenges of "scarce computing power, harsh environment, and dynamic interference" in actual deployment. To address these limitations, several technical paths have shown good engineering potential in recent years and can serve as the focus of subsequent research. On the detection side, lightweight strategies have achieved phased breakthroughs. For example, EDGS-YOLOv8 reduces the model GFLOPs by 35% on the Jetson Nano platform by introducing GhostConv and a deep separable convolution structure, while maintaining the original mAP and reaching 25 FPS in actual measurements, verifying the computing power compression capability of the perception module on a low-power platform [17]. On the positioning side, multimodal information fusion is becoming a key breakthrough in improving the robustness of SLAM. LVI-Fusion tightly couples LiDAR, camera, and IMU data into a factor graph optimization framework, controls the ATE error within 10 cm in weak texture and illumination mutation scenes, and maintains the frame rate at 25 FPS, effectively improving the positioning stability in extreme environments. At the same time, the coupling of perception and mapping tasks also shows good synergistic effects [18]. The YPR-SLAM system combines YOLOv8's detection and elimination of dynamic objects with geometric constraints in the map optimization process, reducing the trajectory error by about 30% under the TUM-RGB-D dynamic sequence without affecting real-time performance, indicating that the map optimization mechanism based on detection priors can effectively suppress motion interference and improve map coherence and navigation reliability [19].

CONCLUSION

This review focuses on the "detection-localization-avoidance" link and presents the latest progress and bottlenecks of two representative algorithms. Different algorithms have their advantages and are suitable for different mission requirements and flight environments. In UAV-YOLOv8, with the help of WIoU v3 and BiFormer, the mAP@0.5 of VisDrone2019 was increased by 7.7 pp to 47.0%; UAV-YOLOv8 achieved 80.3% mAP with only 0.68M parameters and 5.3GFLOPs through layer deletion and Ghost convolution, and the frame rate was ≈90 FPS. Dense visual SLAM-3R relies on the I2P-L2W two-stage network and confidence gating to achieve 6.61 cm ATE-RMSE and 24FPS in Replica, refreshing the "centimeter-level-real-time" balance. In summary, current visual algorithms have basically met the needs of centimeter-level positioning and hundred-frame-level detection in GPS-free scenarios such as forest inspections and indoor search and rescue, but there are still risks of recall attenuation and global drift in low-light, sparse textures, and Jetson-level platforms. In the future, it should focus on hardware perception compression, multimodal robustness enhancement, online adaptive learning, and DRL-coupled planning to further approach the optimal frontier of accuracy- computing power -environmental robustness.

Future research can focus on three aspects: lightweight models, multi-source fusion, and task coupling: by introducing Ghost/RepConv and pruning technology, the YOLOv8 structure can be further compressed and the small target detection performance can be maintained; in terms of SLAM, the integration of sparse attention Transformer and multi-modal perception such as event cameras and LiDAR is expected to achieve stable deployment of SLAM-3R on embedded platforms; at the system level, embedding the semantic information output by YOLO into the graph optimization process can simultaneously improve the map expression and positioning stability. These paths jointly promote the perception- mapping -planning integrated framework, providing efficient and robust support for multi-UAV collaborative obstacle avoidance.

REFERENCES

1. Z. Shen, S. Li, W. Lin, and Z. Du, IEEE Transactions on Intelligent Transportation Systems **24**, 6202-6223 (2023).
2. Encyclopédie de l'Environnement, "Role of forests in the planet's carbon balance" (2025).
3. S. P. Haeri, A. Razi, S. Khoshdel, F. Afghah, J. L. Coen, L. O'Neill, and K. G. Vamvoudakis, "A comprehensive survey of research towards AI-enabled unmanned aerial systems in pre-, active-, and post-wildfire management" (2024).
4. G. Wang, Y. Chen, P. An, H. Hong, J. Hu, and T. Huang, Sensors **23**, 7190 (2023).
5. S. Liu, Z. Zhao, J. Wei, and Q. Zhou, Sensors **24**, 6948 (2024).
6. C. Zhao, H. Yang, L. Jiang, Q. Wang, and Y. Chen, Electronics **12**, 2847 (2023).
7. H. Wang, S. Lou, J. Jing, Y. Wang, W. Liu, and T. Liu, PLOS ONE **17**, e0263841 (2022).

8. H. Guo, Y. Li, H. Wang, C. Wang, J. Zhang, T. Wang, and F. Yang, Computers and Electronics in Agriculture **227**, 109596 (2024).
9. J. Pan, J. Li, H. Wang, Y. Li, L. Zhang, and H. Zhao, Electronics **11**, 4049 (2022).
10. G. Hao, Q. Lv, Z. Huang, H. Zhao, and W. Chen, Aerospace **10**, 562 (2023).
11. M. Huang and F. Qian, Applied Sciences **13**, 6580 (2023).
12. Y. Liu, S. Dong, S. Wang, Y. Yin, Y. Yang, Q. Fan, and B. Chen, "SLAM3R: Real-time dense scene reconstruction from monocular RGB videos," arXiv:2412.09401 (2024).
13. G. Jocher, A. Chaurasia, and J. Qiu, "YOLOv5 and YOLOv8 Technical Report," Ultralytics (2023).
14. C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, IEEE Transactions on Robotics **32**, 1309-1332 (2016).
15. R. Mur-Artal and J. D. Tardós, IEEE Transactions on Robotics **33**, 1255-1262 (2017).
16. R. Xie, Z. Meng, L. Wang, H. Li, K. Wang, and Z. Wu, IEEE Access **9**, 24884-24896 (2021).
17. M. Huang, W. Mi, and Y. Wang, Drones **8**, 337 (2024).
18. Z. Liu, Z. Li, A. Liu, K. Shao, Q. Guo, and C. Wang, Remote Sensing **16**, 1524 (2024).
19. X. Kan, G. Shi, X. Yang, and X. Hu, Sensors **24**, 6576 (2024).