Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/115/2025.18520

Robotics Navigation Based on YOLO and Depth Camera

Boxuan Zhang[1,a,*]

[1]*Department of University of Birmingham, Birmingham, England*
a. 1152778726@qq.com
**corresponding author*

Abstract: In recent years, significant advancements have been made in robotics, yet challenges remain in navigation, particularly for autonomous vehicles and bipedal robots. This paper provides a comprehensive review of three critical components in robotic navigation: YOLO neural networks, depth cameras, and A* path planning. Existing studies often address these aspects independently, lacking an integrated approach. Here, we systematically summarize and analyze the effectiveness of these techniques in navigation tasks. Through a literature review of recent publications, we compare and categorize various methods to assess trends in YOLO research and explore potential for integrated research in navigation. Furthermore, we analyze the strengths and limitations of each method in dynamic environments. The findings suggest that while YOLO and depth camera-based systems excel in real-time object detection and spatial awareness, they face challenges related to light sensitivity and high computational demands. Future research directions are proposed to enhance adaptability in complex environments, improve efficiency, and support cost-effective navigation solutions in robotics.

Keywords: YOLO neural networks, Depth camera, Path planning, Navigation.

1. Introduction

In recent years, the field of robotic navigation has developed rapidly, not only in the domain of autonomous vehicles but also in areas like robotic walking and drone recognition. For example, Tesla has launched the Autopilot assisted driving feature. Although the current models are equipped with the necessary hardware for autonomous driving, full self-driving has not yet been achieved. In the field of high-performance robots, Unitree Robotics has developed the Unitree G1, which integrates an intelligent vision system with powerful AI algorithms, allowing G1 to continuously evolve to adapt to diverse tasks and environments. However, due to the immature software ecosystem, G1's navigation and operational capabilities in complex or dynamic environments still require further optimization and validation. In the field of drone recognition, DJI drones have also made significant advancements by combining multiple sensor systems to capture real-time environmental images and process them to achieve high accuracy in obstacle avoidance and object recognition. Robotic navigation involves multiple steps, including depth camera recognition, neural network computation, and path planning, with each step using different methods that exhibit significant efficiency differences in recognizing and responding to different objects [1]. Therefore, finding a solution that can effectively integrate these technologies is crucial for achieving efficient autonomous navigation. Depth cameras provide rich spatial information, helping robots understand their surroundings. A*

DOI: 10.54254/2755-2721/115/2025.18520

path planning is widely used to find the shortest path between points in a known environment, while the YOLO algorithm is a powerful real-time object detection method that enables robots to quickly recognize and respond to nearby objects [2]. Although these technologies each have unique advantages, they have mostly developed independently, which limits their potential due to a lack of systematic integration. For instance, depth camera image analysis is effective for detailed mapping and obstacle detection, but it struggles in environments with complex lighting conditions. A* path planning is a powerful path optimization algorithm but requires a well-defined map, which is challenging to obtain in dynamic environments. YOLO 3D detection excels in identifying objects but faces challenges in maintaining accuracy when objects are partially obscured or moving rapidly. This paper aims to systematically summarize and discuss the effectiveness of depth cameras, A* path planning, and YOLO neural networks in robotic navigation while exploring potential integrated research directions in the future. The goal is not only to evaluate the strengths and limitations of each method but also to explore how they can complement one another to address current challenges in robotic navigation. This paper first introduces the background. It then discusses the current status and research progress of depth cameras, A* path planning, and YOLO neural networks in robotic navigation. Following that, it analyzes the advantages and disadvantages of these technologies in both independent and integrated applications, with a comparative analysis based on literature reviews and experimental data. Finally, the paper proposes future research directions and integrated application strategies to enhance the performance and adaptability of robotic navigation systems in complex dynamic environments.

2. Literature Review

2.1. Path Planning

Path planning plays a crucial role in fields such as robotic navigation, autonomous driving, and drone control. The purpose of path planning is to find an optimal path from a starting point to a destination for a robot, autonomous vehicle, or drone while avoiding obstacles. Path planning is generally divided into global path planning and local path planning, and what we require is real-time obstacle avoidance in dynamic environments, also known as local path planning. This paper focuses on the A* algorithm and Dijkstra's algorithm [3]. The A* algorithm is a classic path planning algorithm based on graph search, which finds an optimal path from the starting point to the target point among nodes or grids in a graph. It combines heuristic and cost-based search methods. Dijkstra's algorithm, also known as the greedy algorithm, is another path planning algorithm used to calculate the shortest path in a weighted graph. It only applies to non-heuristic functions.

The A* algorithm uses a heuristic function to guide the search direction, effectively reducing the search space and improving search efficiency. In contrast, Dijkstra's algorithm does not use a heuristic function, making it less efficient in complex spaces. Therefore, A* is more suitable for finding optimal paths when the target is known, while Dijkstra is better suited for single-source shortest path searches. The A* algorithm is widely applied in autonomous driving and drone control, because autonomous vehicles need to achieve real-time obstacle avoidance and find the optimal path in complex road environments [4]. Dijkstra's algorithm, on the other hand, is suitable for static environments, such as indoor robotic navigation, like warehouse robots [5,6,7].

2.2. Space Detection

Space detection is also a key technology in fields such as robotic navigation and autonomous driving, primarily used to identify and perceive navigable areas and critical obstacle locations in the environment. Its purpose is to provide robots or autonomous vehicles with a clear environmental map, allowing them to automatically plan paths and avoid obstacles in complex environments. The process

DOI: 10.54254/2755-2721/115/2025.18520

of space detection can be divided into three steps: Data Collection; Data Processing and Analysis; Map Generation. The models we commonly use are LiDAR Space Detection and Intel RealSense Camera. In autonomous driving, where LiDAR generates precise environmental models, enabling vehicles to recognize roads, pedestrians, and other real-world obstacles, thereby achieving accurate obstacle avoidance and path planning[8]. Depth-camera-based indoor robot navigation has already been applied in large, well-defined indoor environments, and there are also applications in augmented reality (AR) [9].

Today's space detection technology has significant advantages in environmental modeling and object recognition within known environments, especially in robotic navigation and autonomous driving. Using technologies such as LiDAR and depth cameras, these fields can detect the environment and generate a complete environmental grid map, aiding in the navigation of robots and autonomous vehicles. However, current space detection technology still has limitations. For instance, LiDAR has high speed and large detection areas but is costly; time-of-flight (ToF) imaging is slower, has higher processing costs, covers a larger area, and has moderate accuracy; structured light has the highest accuracy but machine learning shows advantages in imaging area and multi-unit applications. Future improvements in space detection may include reducing costs and improving image accuracy.

2.3. Vision Navigation

Visual navigation refers to the use of cameras or other sensors to capture image data of the environment. Through image processing and computer vision algorithms, robots, drones, or autonomous vehicles are enabled to recognize paths and avoid obstacles. In 1979, J.J. Gibson proposed the theory of optical flow, which provided fundamental concepts for visual navigation and laid the foundation for computer vision [10]. David Marr's multi-stage visual processing theory further solidified the basis for visual navigation by incorporating steps such as feature extraction, edge detection, and 3D reconstruction [11], establishing a layered approach to information processing for subsequent navigation systems.

As shown in figure 1, YOLOv3 performs well overall, especially in detecting small objects. Although YOLOv2 achieves higher accuracy in certain large object categories, YOLOv3 demonstrates more stability in F1 scores. As shown in figure 2, YOLOv5 provides better detection accuracy across most object categories, making it suitable for applications requiring higher precision, while YOLOv3 continues to perform stably and balanced in large object detection.

Visual navigation has made significant progress in various fields, especially in autonomous driving, drone navigation, and robotics applications. Improvements in ORB-SLAM have notably enhanced the precision and real-time capabilities of visual navigation systems. In recent years, with the integration of deep learning, visual navigation has achieved more complex scene understanding and applications. Visual navigation now incorporates deep learning techniques such as Fast R-CNN and YOLO for object detection in visual navigation [13, 14]. For instance, Keisuke Tateno's article "CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction" combines deep learning with SLAM to merge the advantages of depth prediction and SLAM, resulting in more precise and dense single-view scene reproduction, laying a foundation for future scenarios and supporting data for path planning [15]. The advantages of visual navigation technology include low cost, rich information, and strong adaptability. Compared to LiDAR, cameras provide more comprehensive environmental information, capturing details and colors of objects, which offer extensive visual data for path planning and object recognition. However, visual navigation is highly dependent on lighting conditions, with performance decreasing under low light or extreme weather conditions.

DOI: 10.54254/2755-2721/115/2025.18520

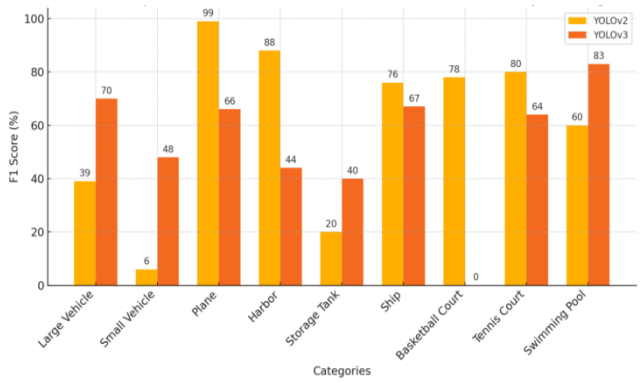

Figure 1: Score Comparison between YOLOv2 and YOLOv3 for Different Object Categories

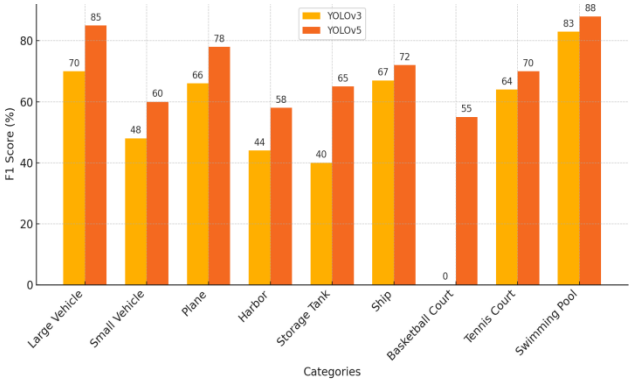

Figure 2: Score Comparison between YOLOv3 and YOLOv5 for Different Object Categories

3. Results and discussions

3.1. Advantages and Limitations of YOLO and Depth Camera-Based Navigation

YOLO can recognize multiple targets in a very short time, making it particularly suitable for navigation scenarios that require rapid responses. Depth cameras provide precise distance information of objects within a scene, enabling the navigation system to understand the absolute position and three-dimensional distribution of objects. The combination of YOLO and depth cameras enables the navigation system to achieve precise obstacle avoidance and path planning in complex environments. The rapid detection capabilities of YOLO and the real-time depth updates from depth cameras allow the navigation system to efficiently handle constantly changing obstacles in dynamic environments. The combination of these two will require us to overcome many more obstacles.Depth cameras perform inconsistently in complex lighting environments, such as strong or weak light, particularly those that rely on infrared light, which tend to perform poorly under strong illumination. The combination of YOLO and depth cameras demands high computational resources. Efficient operation of the YOLO model typically relies on robust GPU support to maintain real-time performance. YOLO is designed with speed in mind, adopting an end-to-end regression approach that transforms the entire detection process into a single regression problem. Although this method significantly increases detection speed, it falls short in precision compared to region proposal-based methods such as Faster R-CNN.

The integration of YOLO and depth cameras in navigation systems demonstrates strong real-time processing capabilities, accurate depth perception, and adaptability to the environment, providing efficient processing methods for various navigation tasks. However, limitations still exist in adaptability to lighting conditions, computational demands, and precision in localization. Future research directions could include enhancing adaptability to ambient light, optimizing computational resource needs, and improving dynamic object detection and long-distance navigation to further enhance the system's performance and applicability.

3.2. Comparative analysis supported by experimental data

According to experimental data from existing literature, significant differences exist in the performance of YOLOv3 and YOLOv5 across various object detection categories. Overall, YOLOv5 exhibits higher precision and recall rates in detecting small objects (such as small vehicles and swimming pools), especially in complex backgrounds where YOLOv5 significantly outperforms

DOI: 10.54254/2755-2721/115/2025.18520

YOLOv3. On the other hand, YOLOv3 has higher recall rates for larger object categories (such as ports and basketball courts), showing stable performance suitable for detecting larger objects. YOLOv5 covers a broader range in most categories, demonstrating a balanced performance in precision and real-time capabilities, while YOLOv3 still maintains good detection precision in some large object detection. These charts clearly illustrate the applicability of the two algorithms in different detection tasks, providing a basis for choosing the appropriate model.

4. Future Research Direction

4.1. D Object Detection and Tracking in Dynamic Environments

Building on YOLO, the integration of 3D information from depth cameras can develop real-time detection and tracking methods suited for dynamic environments. Future research could focus on integrating YOLO detection capabilities with time-series data from deep learning and then utilizing motion models to enhance the stability of the model's detection in dynamic environments.

Due to the high computational costs associated with processing 3D data using YOLO and depth cameras, future improvements could involve refining the structure of YOLO and the data processing methods of depth cameras to make them more lightweight. Techniques such as optimizing convolutional structures and compressing model parameters could be employed to develop lightweight 3D detection models suitable for embedded and mobile devices.

Depth cameras provide precise distance and depth information for path planning. Future research could explore how to integrate YOLO's object detection results with depth information for automated path planning and obstacle avoidance. Techniques that combine three-dimensional spatial information with path planning algorithms aim to achieve precise navigation, obstacle avoidance, and dynamic path optimization.

4.2. Visual Comparison

To more intuitively display the advantages and disadvantages of different technologies, we have compiled a comparison chart (Table 1) that includes LiDAR, ToF, Structured Light, Kinect Fusion, and Intel RealSense depth cameras. The chart summarizes their strengths and weaknesses in terms of imaging speed, accuracy, and adaptability to lighting conditions. This visualization helps readers quickly grasp the performance differences of these technologies across various application scenarios.

Table 1: Compare among Depth Camera

Technology Type	*Imaging Speed*	*Detection Accuracy*	*Light Adaptability*	*Cost*	*DPR*
Lidar	*Fast*	*High*	*Strong*	*High*	*High*
ToF	*Medium*	*Medium*	*Weak*	*Medium*	*Medium*
Structured Light	*Slow*	*High*	*Medium*	*Low*	*Medium*
Kinect Fusion	*Fast*	*Medium*	*Weak*	*Medium*	*High*
Intel RealSense	*Medium*	*Medium*	*Strong*	*Medium*	*Medium*

5. Conclusion

The primary aim of this article is to introduce the advantages and feasibility of a robotic navigation system based on YOLO and depth cameras. Based on current achievements, the system has demonstrated impressive rapid detection and obstacle avoidance capabilities in dynamic environments, providing reliable support for robotic navigation. However, this research also has limitations due to the environmental perception capabilities of depth cameras and their computational resource requirements, which may restrict their application. Future research directions have been

DOI: 10.54254/2755-2721/115/2025.18520

identified to address these limitations, specifically improving light adaptability and reducing computational costs to enhance detection accuracy in complex environments. It is hoped that through the technological optimization of YOLO and depth cameras, they can be better applied in robotic navigation.

Environmental perception and path planning based on deep learning, by integrating spatial information from depth cameras and efficient path planning algorithms, have achieved a balance of high precision, real-time performance, and adaptability to complex scenes. Future research could focus on reducing costs, enhancing environmental adaptability, and expanding path planning algorithms. Future studies might further explore various directions, such as increasing adaptability to lighting conditions and enhancing the environmental robustness of the YOLO and depth camera systems. Integrating data from multiple sensors (such as LiDAR and cameras) could improve the robustness of navigation systems in low-light and dynamically complex environments. Further optimization of depth cameras and the YOLO architecture to maintain efficient performance on devices with limited computational resources is one of the prospective research directions.

References

[1] *Izadi, Shahram, et al. "Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera." Proceedings of the 24th annual ACM symposium on User interface software and technology. 2011.*

[2] *Hart, Peter E., Nils J. Nilsson, and Bertram Raphael. "A Formal Basis for the Heuristic Determination of Minimum Cost Paths." IEEE Transactions on Systems Science and Cybernetics, vol. 4, no. 2, 1968, pp. 100-107. https://doi.org/10.1109/TSSC.1968.300136.*

[3] *Warren, Charles W. "Fast path planning using modified A* method." [1993] Proceedings IEEE International Conference on Robotics and Automation. IEEE, 1993.*

[4] *Dolgov, Dmitri, et al. "Practical search techniques in path planning for autonomous driving." Ann Arbor 1001.48105 (2008): 18-80.*

[5] *Li, Xiao-Hui, Miao Miao, Biao-Jian Ran, Yi Zhao, and Gang Li. "Obstacle Avoidance Path Planning for UAV Based on Improved A* Algorithm." Computer Systems & Applications, vol. 30, no. 2, 2021, pp. 255-258.*

[6] *Zhao, Xuanxuan. Research on Indoor 3D Navigation Model and Path Generation Algorithm. Diss., Shandong University of Science and Technology, 2016.*

[7] *J. Wang, H. Huang, J. Li, L. Jiang, J. Li and F. Jiang, "AGV path planning algorithm based on improved Dijkstra algorithm," 2024 6th International Conference on Internet of Things, Automation and Artificial Intelligence (IoTAAI), Guangzhou, China, 2024, pp. 568-574, doi: 10.1109/IoTAAI62601.2024.10692705. Collis, R. T. H. "Lidar." Applied optics 9.8 (1970): 1782-1788.*

[8] *Biswas, Joydeep, and Manuela Veloso. "Depth camera based indoor mobile robot localization and navigation." 2012 IEEE International Conference on Robotics and Automation. IEEE, 2012.*

[9] *Oufqir, Zainab, Abdellatif El Abderrahmani, and Khalid Satori. "ARKit and ARCore in serve to augmented reality." 2020 International Conference on Intelligent Systems and Computer Vision (ISCV). IEEE, 2020.*

[10] *Gibson, James J. The ecological approach to visual perception: classic edition. Psychology press, 2014.*

[11] *Marr, David. Vision: A computational investigation into the human representation and processing of visual information. MIT press, 2010.*

[12] *Harris, Chris, and Mike Stephens. "A combined corner and edge detector." Alvey vision conference. Vol. 15. No. 50. 1988.*

[13] *Bin Issa, Razin, et al. "Double deep Q-learning and faster R-Cnn-based autonomous vehicle navigation and obstacle avoidance in dynamic environment." Sensors 21.4 (2021): 1468.*

[14] *Dos Reis, Douglas Henke, et al. "Mobile robot navigation using an object recognition software with RGBD images and the YOLO algorithm." Applied Artificial Intelligence 33.14 (2019): 1290-1305.*

[15] *Tateno, Keisuke, et al. "Cnn-slam: Real-time dense monocular slam with learned depth prediction." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.*