Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD27801

Research on Intelligent Navigation and Dynamic Obstacle Avoidance of Robots Based on Visual Perception: A Review

Xiang Li

Faculty of Information Science and Technology (FTSM), National University of Malaysia (UKM), Bangi, Malaysia
lx11668899@gmail.com

Abstract. In modern autonomous robotic applications, robots are increasingly expected to operate in unstructured environments, where both navigation capability and dynamic interaction directly affect task efficiency. However, although vision-based perception has been widely studied, many existing approaches still struggle to adapt to complex and unpredictable scenarios, and their integration into real-world engineering systems is often limited. In this paper, we examine vision-driven robotic navigation and interaction methods by focusing on several core aspects, including the coordination between visual SLAM and navigation, improvements in visual object recognition, and multi-node autonomous interaction. From the analysis, it can be observed that most SLAM and navigation cooperation frameworks are mainly validated in controlled environments rather than real dynamic settings, while visual recognition methods tend to be sensitive to disturbances such as illumination variation and occlusion. At the same time, the overall integration of navigation, perception, and interaction modules in practical robotic systems is still not sufficiently developed. Based on these findings, this study aims to provide a more practical reference for designing robotic systems in low- to medium-complexity environments, while also supporting the transition of vision-based perception technologies from laboratory validation toward real industrial applications and offering directions for further system-level improvements.

Keywords: Visual Perception, Visual SLAM, Navigation Collaboration, Visual Target Recognition, Multi-node Autonomous Interaction

1. Introduction

With the rapid development of the times and the world, intelligent manufacturing and service robotics technologies are rapidly iterating, and the demand for autonomous navigation and dynamic interaction of robots has increased significantly [1,2] (for example, in unstructured scenarios, cleaning foreign objects in warehouses, and moving objects in workshops). Visual perception technology can obtain high-dimensional environmental information (such as target color, shape, and spatial position) through non-contact means, becoming the core link between robots and complex environments. Its integration with simultaneous localization and mapping (SLAM) and path planning technologies has become a key path to achieving autonomous robot operations [3,1].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD27801

Although visual perception robot navigation technology has made significant progress, there are still gaps in the existing results: First, the collaboration between visual SLAM and navigation is mostly verified in ideal structured environments, and there is a lack of systematic summary of its adaptability to unstructured scenes (such as work areas with randomly distributed obstacles) [4,5]; second, the "accuracy-efficiency" trade-off mechanism of visual recognition algorithms does not fully consider actual lighting and perspective interference, and fixed parameters (such as HSV color thresholds) are prone to failure in complex scenes [6,7]; third, there is insufficient research on the engineering connection of "navigation-recognition-interaction", and the literature mostly focuses on the optimization of a single technology, which makes it difficult to directly provide a reference for the implementation of low-cost robots [3,1].

In this paper, visual perception as applied to robotics is examined, including visual SLAM and the collaborative nature of navigation; optimizing for visual object recognition; and the use of multi-node autonomous agents. The intention of this paper is to provide a theoretical reference to robots that use visual-perception technologies in developing robotic systems in low and moderate complexity environments and assist in the transition of evolving visual-perception technologies from an academic to a commercial application.

2. Literature survey

Dynamic obstacle avoidance and intelligent navigation of robots is extremely important for their full integration into many applications including home services, warehousing and logistics, and outdoor inspection. Due to the nature of the navigation methods typically used, they are unable to produce satisfactory results in complicated environments.

2.1. Collaborative design of visual SLAM and autonomous navigation strategies

Visual SLAM (Simultaneous Localization and Mapping) is one of the primary technologies involved with providing robots the ability to autonomously navigate in previously uncharted territory. The accuracy-adaptability trade-off is widely studied in the literature. Kim et al. introduced the PDN (Perception-Driven Navigation) method, which integrates visual saliency clustering with path planning so as to equalize with SLAM exploration and revisit in structured laboratory situations. But this methodology presupposes a static environment, is liable to error in the presence of moving objects and can harm robots with false convergence on moving objects [4]. To overcome the challenges of dynamic positioning in outdoor gardens, Peng et al. [5] developed a real-time dynamic SLAM visual odometry found on instance segmentation that separates dynamic objects from static backgrounds, thereby enhancing the accuracy of positioning, providing a lightweight approach for dynamic scene adaptation.

Cong et al. extended the application of 3D vision in SLAM and improved the navigation accuracy of the dynamic scenes. This method, however, depends on dense point cloud obtained from a depth camera and is constrained by specific hardware [8]. Corresponding to this, Kumar et al. [9,10] have made a series of research on the 3D SLAM algorithm. They integrated it with a human-aided movement mechanism to improve the robustness of robot pose estimation in low-texture environment, proving a low-cost viable way to adapt to complicated scenes without strong hardware dependence.

There is also an adaptation on the Lightweight SLAM solutions: for instance, Gmapping algorithm, which is a Rao-Blackwellized particle filter algorithm that uses 2D lidar and a monocular camera sensor to capitalize on its compatibility with standard mobile platforms, is well-known

DOI: 10.54254/2755-2721/2025.LD27801

enough now. The entire map building quality in restricted passages and multi-obstacle situations, however must be strengthened [3]. The reference [11] employs static-dynamic-expansion layer of hierarchical cost map structure to realize dynamic obstacle priority coverage, which can be considered as an orthogonal solution of visual dynamic target filtering logic with multi-dimensional hierarchical processing. The approach to couple TSP path planning with dynamic prediction also offers a referral path for task level optimization of path replanning for multi-foreign object elimination in tight channel scenarios, which can further improve the applicability of navigation tasks [12].

In order to further prove that lightweight SLAM is feasible on the industrial field, given the narrow processing channel of industrial products in hardware manufacturing, the experiments were conducted on the TurtleBot3 Waffle Pi simulating an industrial environment with tight passages-width and scattered static obstacles. Exploiting a well-known algorithm, the Gmapping algorithm, for combined mapping, localization and path planning, the experiments proved that the method effectively fenced off passage walls line and static obstacles and reduced the map errors caused by moving interferences. This confirms the applicability of lightweight SLAM to industrial narrow-channel navigation, which is of great constructive guidance to practical engineering.

Figure 1 represents the left half of the industrial narrow channel simulated map generated by employing the Gmapping algorithm. It has channel boundaries, with the location of the blue foreign object ball, and the static obstacle ice cream cone locations. Left: Real-time navigation window of the TurtleBot3 robot. It shows the robot as it follows the planned path while avoiding motion interference with visual dynamic filtering logic. These two figures show that lightweight SLAM can reconstruct high precision maps and robust navigation in narrow channel environments.

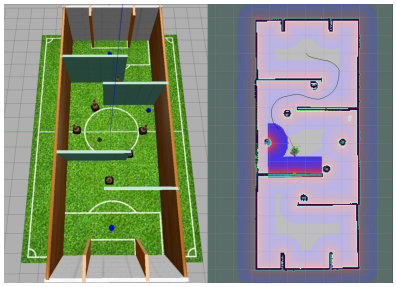


Figure 1. Narrow channel SLAM mapping and navigation results

2.2. Visual feature extraction and target recognition optimization

Visual target recognition is essential for robots to sense the environment and find targets/obstacles. This is essentially a problem of being sufficiently robust in the presence of complicated disturbances. Among traditional algorithms, Sun et al. What's more, it is capable of enhancing the feature-matching stability in low-texture scenes by means of adaptive ORB feature extraction and a multi-stage matching strategy. Nevertheless, this model applies to a firm HSV color threshold resulting in a lot of undetected targets in bright or shady regions [6]. Mousavi et al. introduced a randomized sampling feature selection algorithm for real-time constraints and demonstrated significantly reduced feature screening time while applicable for mobile robot platforms. But the recognition performance is poor when the target is partly occluded [13].

Deep learning technique opens up a new way for this area. Some works apply convolutional neural networks (CNNs) to perform end-to-end target detection and increase environmental robustness (by pre-training on large image datasets) [14]; Misir et al. [7] also established the

DOI: 10.54254/2755-2721/2025.LD27801

effectiveness of CNNs by performing end-to-end detection of blue spherical targets in real-time dynamic scenes with an obstacle avoidance success rate of more than 90%, signaling towards practical task integration. Among them, some integrate attention mechanisms to improve CNN models and mitigate influence of background distraction to recognition results [9].

It should be noted that the above research also studied the breakthrough from the angle of decoupling of multimodal features and cross-spectral fusion: the dual-branch multi-scale spatiotemporal network introduced in [15] reinforces the weak texture targets' robustness by spatiotemporal feature decoupling; the thermal-visible light fusion framework is the optimal way to reduce the recognition blind spot issue under stable light/shadow by cross-modal complementarity, it can serve as a reference for the uplifting of the dimensions of visual perception [16].

A corresponding tradeoff between robustness, computational burden, and environment adaptability arises if these techniques have to be integrated systematically: Conventional feature-based methods (e.g., KTBER_AORB [6], random sampling [13]) possess lightweight computation for deployment on resource-limited platforms but the fragile performance against dynamic noise (e.g., illumination changes, partial occlusions) is mainly due to fixed parameters and handcrafted features. Deep learning methods (e.g. CNNs [7], attention-augmented networks [9]) exhibit good generalization and robustness with large-scale pre-training but are computationally intensive and hence their use in typical robotic systems is limited. Multimodal fusion (e.g., spatiotemporal decoupling [15], thermal -visible fusion [16]) offers a trade-off: cross-sensor complementarity improves robustness to extreme conditions (e.g., dark, fully occluded) with a mild computational cost and also generalizes well to different environments.

Such a trade-off calls for scenario-dependent tuning — an area filled mainly by the experimental demonstration of dynamic parameter tuning, which reconciles the traditional/speedy efficiency with the deep/sturdy robustness on the old platforms.

As shown in Figure 2, this study simulated industrial blue sphere recognition and conducted experimental verification on the TurtleBot3 platform. The visual detection logic was optimized by dynamically adjusting the color space parameters (such as the HSV range) and contour recognition threshold corresponding to the blue target. The left-hand interface (such as the RViz visual topic window) displays the adjusted configuration status. In the real-time camera image, the blue sphere target is accurately marked. Experimental results further verify that this solution does not require high-computing power modules and can be adapted to conventional computing platforms.

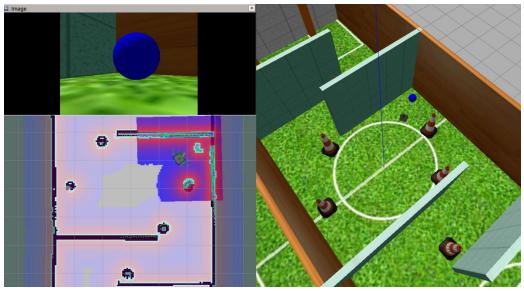


Figure 2. Visual recognition effect of the sphere

2.3. Multimodal perception fusion and dynamic obstacle avoidance

The perception blind spots of a single sensor can easily lead to the failure of the robot's obstacle avoidance or deviation in task coordination in complex scenarios. Multimodal perception fusion has become the core technical direction of dynamic obstacle avoidance and task coordination by

DOI: 10.54254/2755-2721/2025.LD27801

complementing the performance advantages of different sensors (such as visual target recognition capabilities and lidar's depth perception capabilities). Existing research focuses on the two core aspects of "perception accuracy" and "platform adaptability". Misir et al. proposed an improved MobileNetV2 model to fuse the distance information of the ultrasonic sensor with the visual image. They assisted the obstacle avoidance decision by superimposing warning signs on the image. This dual-modal solution performed stably in medium and short-range obstacle detection, but the ultrasonic reflection angle easily affected the recognition of low obstacles [7]. Yang et al. pointed out that "vision + LiDAR" fusion can effectively compensate for the shortcomings of vision in depth judgment and improve the accuracy of obstacle positioning in complex environments through point cloud data. However, its point cloud processing process requires high computing power and is more suitable for high-performance robot platforms [17].

In terms of cooperative reasoning and on-line optimization in multimodal fusion, a technical scheme based on multi-source sensor fusion and trajectory planning for wheeled robots in unstructured environments is developed in this paper. The key is to realize the construction of an environmental perception model by fusing visual with lidar data to close the loop of "real-time perception-dynamic trajectory adjustment". The core insight that the perception quality determines the trajectory optimality theoretically supports the existence of the relation between the sensor data quality and the obstacle avoidance efficiency in multimodal schemes. Meanwhile, its lightweight optimization strategies can also serve as guidelines for adapting classical platforms [18]. Soualhi et al. [19,17] further generalize this collaborative framework through merging motion perception with deep reinforcement learning, developing a visual control scheme to reduce response latency in dynamic scenes - an approach that enhances the connection between real-time environmental perception and agile decision-making in fast-changing environments. The LQR control based dynamic allocation framework that follows turns the task priority logic of human machine interaction into robot scene, giving an engineering decision reference for bridging "the obstacle avoidance action and the target related tasks" (such as avoiding obstacles first then perform the target related operation), which is especially suitable for multitask parallel scenarios [20].

Meanwhile, the "vision + infrared" fusion method has also gained public attention. The solution employs infrared sensors to detect low obstacles that are visually hard to recognize, thus adding one more dimension to perception. However, the triggering mechanism of multi-sensing cooperation (when to turn on infrared and when to rely on vision) is still mainly tailor-made in current study without general process for conventional robots. It can be further optimized according to the specific scenario requirements [21]. From the perspective of current research, in the area of robot obstacle avoidance, multimodal perception fusion currently has the following kind of exploration: dual-modal fusion (such as vision + ultrasound, vision + LiDAR) has been tested to work in some practical scenarios, and can also provide theoretical and engineering references for multi-modal collaborative logic and light-weight multi-modal adaptation in the related literatures. These attempts involve technical characteristics of various sensor pairs and provide preliminary optimization strategies based on the requirements of adoption for a generic robotic platform and hence, form a foundation to achieving practical multi-modal technology.

3. Future directions

Combined with the current research status of visual perception-driven robot navigation and obstacle avoidance technology, future research can focus on breakthroughs in the four core directions of "unstructured scene adaptation, complex interference robustness, multimodal standardization, and system-level integration." To address the limited positioning accuracy of visual SLAM in

DOI: 10.54254/2755-2721/2025.LD27801

unstructured environments, lightweight fusion of vision with low-cost lidar and infrared sensors can be promoted to supplement depth and low-obstacle information. Simultaneously, lightweight motion prediction modules can be integrated to minimize the interference of dynamic objects on map construction. Regarding visual target recognition, it is necessary to focus on developing lightweight deep learning models suitable for conventional platforms and to deepen the cross-spectral feature fusion and cross-scene parameter adaptation mechanisms to solve the problem of recognition failure caused by light fluctuations and occlusion. In multimodal fusion, a standardized framework for sensor priority decision-making should be established, shifting to feature-level fusion to reduce computing power and adapt to low-cost robots. Furthermore, an end-to-end integrated framework for "navigation-recognition-interaction" should be developed and validated with low-cost hardware solutions to advance the technology from module optimization to engineering implementation.

4. Conclusion

Visual perception technology is a key aspect for robot interaction with complex environments. Its fusion with SLAM navigation, target recognition and multi-modal fusion has become a key approach for robots to realize autonomous navigation and dynamic obstacle avoidance. This article systematically reviews the research progress, focusing on three core technology modules to summarize the current status: As far as visual SLAM and autonomous navigation cooperation is concerned, a mature framework consists of "high precision 3D solution" and " lightweight 2D solution" has already build up in the existing research. The high precision scheme can significantly improve the positioning accuracy of the dynamic scenes, and the lightweight solution can be fully adapted to the traditional robot platforms. This study further confirmed the feasibility of lightweight SLAM in the simulation of industrial narrow channel scenes based on experiments with low-cost robots, which serves as a practical reference for applications in low-medium complexity scenes. In visual object recognition optimization, classic algorithms provide recognition stability for low-texture scenes through feature matching techniques and deep-learning based techniques significantly improved adaptability to environment and robustness. Investigations such as cross-spectral fusion and dynamic parameter modulation offer a variety of viable avenues to enhance or customize the recognition logic of conventional platforms. In the field of multimodal perception fusion, the dual-modal solution has been verified to be effective in many specific scenarios. Related research on multi-source fusion framework and task priority decision logic provides a solid theoretical support for multimodal collaboration, "vision + infrared" Solutions such as these have further enriched the robot's perception dimension, accumulating a sufficient technical foundation for obstacle avoidance and task collaboration in complex scenarios. In summary, current robot navigation and obstacle avoidance technologies based on visual perception have achieved outstanding results in structured scene adaptation, single-function optimization, and engineering exploration. By sorting out the technical context and practical value, this review provides a clear direction for subsequent research, helping this technology steadily transition from laboratory verification to low-cost industrial applications, and laying a solid theoretical and practical foundation for designing robot systems in low- and medium-complexity scenarios.

References

[1] Yang, J., Wang, C., Jiang, B., Song, H., & Meng, Q. "Visual Perception Enabled Industry Intelligence: State of the Art, Challenges and Prospects". IEEE Transactions on Industrial Informatics, 2021, 17 (3), 2204 - 2219.https: //ieeexplore.ieee.org/document/9106415

DOI: 10.54254/2755-2721/2025.LD27801

[2] T. Wang, J. Fan, P. Zheng, R. Yan, L. Wang. "Vision-Language Model-Based Human-Guided Mobile Robot Navigation in an Unstructured Environment for Human-Centric Smart Manufacturing". Engineering, 2025. https: //doi.org/10.1016/j.eng.2025.04.028

[3] Shahria, M. T., Sunny, M. S. H., Zarif, M. I. I., Ghommam, J., Ahamed, S. I., & Rahman, M. H. "A Comprehensive Review of Vision-Based Robotic Applications: Current State, Components, Approaches, Barriers, and Potential Solutions". Robotics, 2022, 11 (6), 139. https: //www.mdpi.com/2218-6581/11/6/139

[4] Kim, A., & Eustice, R. M. "Perception-driven navigation: Active visual SLAM for robotic area coverage". In 2013, IEEE International Conference on Robotics and Automation (ICRA) (pp. 3196 - 3203). IEEE. https: //ieeexplore.ieee.org/document/6631022

[5] J. Peng, Q. Yang, D. Chen, C. Yang, Y. Xu, Y. Qin. "Dynamic SLAM Visual Odometry Based on Instance Segmentation: A Comprehensive Review". Computers, Materials & Continua, 2024, 78(1): 168-196. https: //doi.org/10.32604/cmc.2023.041900

[6] Sun, C., Wu, X., Sun, J., Qiao, N., & Sun, C. "Multi-Stage Refinement Feature Matching Using Adaptive ORB Features for Robotic Vision Navigation". IEEE Sensors Journal, 2022, 22 (3), 2603 - 2617. https: //ieeexplore.ieee.org/document/9663302

[7] Misir, O., & Celik, M. "Visual-based obstacle avoidance method using advanced CNN for mobile robots". Internet of Things, 2025, 31, 101538.https: //www.sciencedirect.com/science/article/abs/pii/S2542660525000514

[8] Cong, Y., Chen, R., Ma, B., Liu, H., Hou, D., & Yang, C. "A Comprehensive Study of 3-D Vision-Based Robot Manipulation". IEEE Transactions on Cybernetics, 2023, 53 (3): 1682 - 1695. https: //ieeexplore.ieee.org/abstract/document/9541299

[9] A. Kumar, K. U. Singh, P. Dadheech, A. Sharma, A. I. Alutaibi, A. Abugabah, A. M. Alawajy. Enhanced Route navigation control system for turtlebot using human-assisted mobility and 3-D SLAM optimization [J]. Heliyon, 2024, 10: e26828. https: //doi.org/10.1016/j.heliyon.2024.e26828

[10] P. Li, D. Chen, Y. Wang, L. Zhang, and S. Zhao, "Path planning of mobile robot based on improved TD3 algorithm in dynamic environment, " Heliyon, vol. 10, 2024, Art. no. e32167. doi: https: //doi.org/10.1016/j.heliyon.2024.e32167.

[11] R. Ospina and K. Itakura, "Obstacle detection and avoidance system based on layered costmaps for robot tractors, " Smart Agric. Technol., vol. 11, p. 100973, 2025, doi: 10.1016/j.atech.2025.100973.

[12] L. Zhang, X. Shi, L. Tang, Y. Wang, J. Peng, J. Zou. "RRT Autonomous Detection Algorithm Based on Multiple Pilot Point Bias Strategy and Karto SLAM Algorithm". Computers, Materials & Continua, 2024, 78(2): 2112-2136. https: //doi.org/10.32604/cmc.2024.047235

[13] Mousavi, H. K., & Motee, N. "Estimation With Fast Feature Selection in Robot Visual Navigation". IEEE Robotics and Automation Letters, 2023, 5 (2), 3572 - 3579. https: //ieeexplore.ieee.org/document/9001183

[14] X. Chi, Z. Guo, F. Cheng. A probabilistic neural network-based bimanual control method with multimodal haptic perception fusion [J]. Alexandria Engineering Journal, 2025, 127: 892-919. https: //doi.org/10.1016/j.aej.2025.06.024

[15] N. Li, X. Yang, H. Zhao. DBMSTN: A Dual Branch Multiscale Spatio-Temporal Network for dim-small target detection in infrared image [J]. Pattern Recognition, 2025, 162: 111372. https: //doi.org/10.1016/j.patcog.2025.111372

[16] T. Gaber, M. Nicho, E. Ahmed, A. Hamed. "Robust thermal face recognition for law enforcement using optimized deep features with new rough sets-based optimizer". Journal of Information Security and Applications, 2024, 85: 103838. https: //doi.org/10.1016/j.jisa.2024.103838

[17] Y. Liao et al., "Refining multi-modal remote sensing image matching with repetitive feature optimization, " International Journal of Applied Earth Observation and Geoinformation, vol. 134, p. 104186, 2024, doi: 10.1016/j.jag.2024.104186.

[18] H. Xu, G. Zhang, H. Zhao. Energy-Efficient Human-Like Trajectory Planning for Wheeled Robots in Unstructured Environments Based on the RCSM-PL Network [J]. iScience, 2025. https: //doi.org/10.1016/j.isci.2025.113296

[19] T. Soualhi, N. Crombez, A. Lombard, Y. Ruichek, S. Galland. Leveraging motion perceptibility and deep reinforcement learning for visual control of nonholonomic mobile robots [J]. Robotics and Autonomous Systems, 2025, 189: 104920. https: //doi.org/10.1016/j.robot.2025.104920

[20] Z. Su, H. Yao, J. Peng, Z. Liao, Z. Wang, H. Yu, H. Dai, and T. C. Lueth, "LQR-based control strategy for improving human–robot companionship and natural obstacle avoidance, " Biomimetic Intell. Robot., vol. 4, p. 100185, 2024, doi: 10.1016/j.birob.2024.100185.

[21] A. Bhuiyan, A. An, J. X. Huang, and J. Shen, "Optimizing domain-generalizable ReID through non-parametric normalization, " Pattern Recognition, vol. 162, p. 111356, 2025, doi: https: //doi.org/10.1016/j.patcog.2025.111356.