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 robotic autonomous operations, navigation and dynamic interaction in unstructured scenarios are key to achieving efficient operations. However, current research on visual perception-based robotic systems lacks adaptability to unstructured scenarios and integration with engineering applications. This article reviews visual perception-driven intelligent navigation and dynamic interaction for robotics, analyzing the technical modules of visual SLAM and navigation collaboration, visual object recognition optimization, and multi-node autonomous interaction. The research found that the collaboration between visual SLAM and navigation has mostly been demonstrated in ideal environments, visual recognition algorithms are less adaptable to complex interference, and the engineering integration of "navigation-recognition-interaction" needs to be strengthened. This review aims to establish a solid theoretical foundation for designing robotic systems in low- and medium-complexity scenarios, advance visual perception technology from laboratory research to industrial application, and contribute to breakthroughs and developments in related fields.

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].

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].

This article reviews the application of visual perception in robot navigation, focusing on visual SLAM and navigation collaboration, visual object recognition optimization, and multi-node autonomous interaction. It aims to provide a theoretical reference for the design of robotic systems for low—and medium-complexity scenarios and to promote the transition of visual perception technology from laboratory research to industrial practice.

2. Literature survey

Robots' intelligent navigation and dynamic obstacle avoidance are crucial for their wide application in home services, warehousing and logistics, and outdoor inspection. Traditional navigation methods have limitations in complex environments.

2.1. Collaborative design of visual SLAM and autonomous navigation strategies

Visual SLAM (Simultaneous Localization and Mapping) is a core technology for autonomous navigation of robots in unknown environments. Existing research focuses on the trade-off between accuracy and scene adaptability. Kim et al. proposed the PDN (Perception-Driven Navigation) algorithm, which combines visual saliency clustering with path planning to balance SLAM exploration and revisit in structured laboratory scenarios. However, this algorithm relies on the assumption of a static environment and carries the risk of misjudging dynamic obstacles [4]. To address dynamic positioning challenges in unstructured settings, Peng et al. [5] proposed a dynamic SLAM visual odometry based on instance segmentation, which improves positioning accuracy by separating dynamic targets from static backgrounds, offering a lightweight solution for dynamic scene adaptation.

Cong et al. expanded the application of 3D vision in SLAM, improving the navigation accuracy of dynamic scenes. However, this solution relies on dense point cloud data from a depth camera and has specific requirements for hardware configuration [8]. Complementing this, Kumar et al. [9,10] optimized the 3D SLAM algorithm. They combined it with a human-assisted movement strategy to enhance the robustness of robot positioning in low-texture scenes, demonstrating cost-effective adaptability to complex environmental features without heavy hardware dependencies.

Lightweight SLAM solutions have also attracted attention: some studies use the Gmapping algorithm that integrates 2D lidar and a monocular camera to adapt to conventional mobile robot platforms. However, the map construction integrity in narrow channels and multi-obstacle scenes must be further optimized [3]. Reference [11] uses a hierarchical cost map architecture (static-dynamic-expansion layer) to construct a dynamic obstacle priority coverage mechanism, providing a multi-dimensional hierarchical processing approach for visual dynamic target filtering logic. The method of coupling TSP path planning with dynamic prediction also provides a reference direction for task-level optimization of path replanning for multiple foreign object removal in narrow channel scenarios, which can further enhance the practicality of navigation tasks [12].

To further validate the practicality of lightweight SLAM in industrial scenarios, this study conducted experiments using the TurtleBot3 Waffle Pi platform, simulating a narrow-channel industrial environment with constrained passages and distributed static obstacles. Leveraging the Gmapping algorithm for integrated mapping, localization, and path planning, the experiment demonstrated that the approach accurately delineated passage boundaries and static obstacles while mitigating motion interference-induced map errors. This validates the feasibility of lightweight SLAM for industrial narrow-channel navigation, providing actionable insights for engineering applications.

Figure 1 illustrates the left half of a simulated industrial narrow channel map constructed using the Gmapping algorithm. It clearly shows the channel boundaries, the static obstacle ice cream cone locations, and the blue foreign object ball. The right half shows the real-time navigation image of the TurtleBot3 robot. It shows the robot moving along the planned path and avoiding motion interference through visual dynamic filtering logic. These two images demonstrate lightweight SLAM's mapping accuracy and navigation stability in narrow channel scenarios.

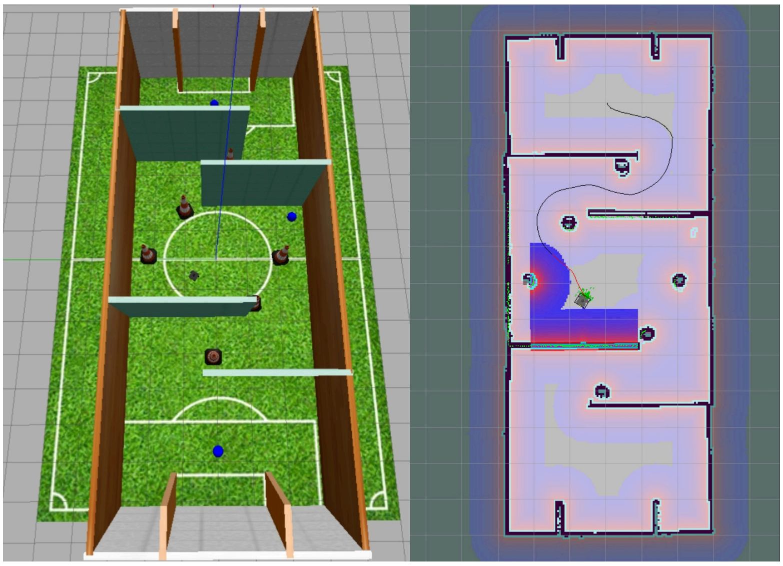

Figure 1. Narrow channel SLAM mapping and navigation results

2.2. Visual feature extraction and target recognition optimization

Visual target recognition is key for robots to perceive the environment and locate targets/obstacles. The core challenge lies in robustness under complex interference (illumination changes, target occlusion, etc.). Among traditional algorithms, Sun et al. proposed the KTBER_AORB model, which improves the stability of feature matching in low-texture scenes through adaptive ORB feature extraction and a multi-stage matching strategy. However, this model uses a fixed HSV color threshold, easily leading to missed targets in bright or shadowy environments [6]. Mousavi et al. designed a random sampling feature selection algorithm to meet real-time requirements, shortening feature screening time and adapting it to mobile robot platforms. However, the recognition effect is limited when the target is partially occluded [13].

Deep learning technology provides a new path for this field. Some studies use convolutional neural networks (CNNs) to achieve end-to-end target detection and improve environmental adaptability through pre-training with large amounts of image data [14]; Misir et al. [7] further demonstrated the efficacy of CNNs by achieving end-to-end detection of blue spherical targets in dynamic scenes, with an obstacle avoidance success rate exceeding 90%, illustrating the potential for real-world task integration. Other studies combine attention mechanisms to optimize CNN models and reduce the impact of background interference on recognition results [9].

It is worth noting that existing research has also explored breakthroughs from the perspective of multimodal feature decoupling and cross-spectral fusion: the dual-branch multi-scale spatiotemporal network proposed in the literature [15] enhances the robustness of weak texture targets through spatiotemporal feature decoupling; the thermal-visible light fusion scheme effectively alleviates the recognition blind spot problem under firm light/shadow through cross-modal complementarity, providing a reference for the expansion of visual perception dimensions [16].

To systematically synthesize these methods, a critical trade-off emerges across robustness, computational efficiency, and environmental adaptability: Traditional feature-based techniques (e.g., KTBER_AORB [6], random sampling [13]) offer lightweight computation for resource-constrained platforms but suffer from brittle performance under dynamic interference (e.g., illumination shifts, partial occlusion) due to fixed parameters and handcrafted features. Deep learning models (e.g., CNNs [7], attention-augmented networks [9]) achieve strong generalization and robustness via large-scale pre-training, yet their high computational demands limit deployment on conventional robotic hardware. Multimodal fusion strategies (e.g., spatiotemporal decoupling [15], thermal-visible fusion [16]) strike a balance: cross-sensor complementarity enhances robustness to extreme conditions (e.g., darkness, full occlusion), with moderate computational overhead, while expanding adaptability to diverse scenarios.

This trade-off highlights the need for scenario-specific optimization — a gap addressed by the subsequent experimental validation of dynamic parameter adjustment, which balances traditional efficiency with deep learning-like robustness on conventional 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.

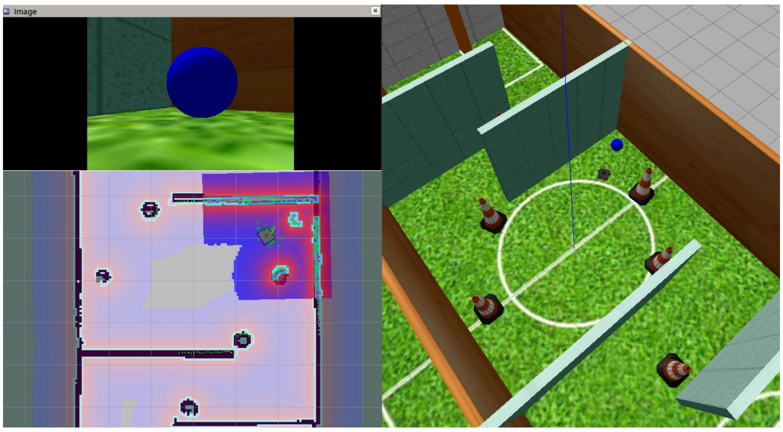

Figure 2. Visual recognition effect of the sphere

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 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].

Regarding the collaborative logic and practical optimization of multimodal fusion, a technical framework for coordinating multi-source sensor fusion and trajectory planning is proposed for wheeled robots navigating unstructured environments. The core is to build an environmental perception model by fusing visual and lidar data to achieve a closed loop of "real-time perception-dynamic trajectory adjustment". The core conclusion that the perception accuracy determines the rationality of the trajectory provides theoretical support for the relationship between sensor data quality and obstacle avoidance efficiency in multimodal solutions. At the same time, its lightweight optimization ideas also provide a feasible direction for adapting conventional platforms [18]. Soualhi et al. [19,17] further expand this collaborative framework by integrating motion perception with deep reinforcement learning, devising a visual control strategy to mitigate response delays in dynamic environments — a method that strengthens the link between real-time environmental perception and agile decision-making in fast-evolving scenarios. The LQR control right dynamic allocation framework transfers the task priority logic in human-machine collaboration to the robot scene, providing an engineering decision reference for the connection between "obstacle avoidance action and target processing task" (such as avoiding obstacles first and then executing target-related operations), which is particularly suitable for multi-task parallel scenarios [20].

In addition, the “vision + infrared” fusion solution has also attracted attention. This solution uses infrared sensors to detect low obstacles that are difficult to identify visually, further enriching the perception dimension. However, in current research, the triggering logic of multi-sensor collaboration (such as when to enable infrared and when to rely on vision) is still mainly customized, and a standardized process adapted to conventional robots has not yet been formed. It can be further optimized based on specific scenario requirements [21]. From the perspective of existing research, multimodal perception fusion technology has formed various exploration paths in the field of robot obstacle avoidance: dual-modal fusion (such as vision + ultrasound, vision + LiDAR) has proven its effectiveness in specific scenarios, and literature also provides theoretical and engineering references for multimodal collaborative logic and lightweight adaptation. These explorations cover the technical characteristics of different sensor combinations and form preliminary optimization strategies tailored to the adaptation requirements of conventional robotic platforms, laying the foundation for the practical application of multimodal 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 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 the core link for robots interacting with complex environments. Its collaboration with SLAM navigation, target recognition, and multimodal fusion has become a key path for robots to achieve autonomous navigation and dynamic obstacle avoidance. This article systematically sorts out the research progress in this field, focusing on the three core technology modules to summarize the current status: In the field of visual SLAM and autonomous navigation collaboration, existing research has formed a mature pattern of "high-precision 3D solution" and "lightweight 2D solution" in parallel. The high-precision solution effectively improves the positioning accuracy of dynamic scenes, and the lightweight solution is fully adapted to conventional robot platforms. Based on experiments conducted with low-cost robots, this study further verified the practicality of lightweight SLAM in simulating industrial narrow channel scenes, providing a practical reference for applications in medium and low complexity scenes. In terms of visual target recognition optimization, traditional algorithms ensure the recognition stability of low-texture scenes through feature matching technology, and deep learning technology has significantly enhanced environmental adaptability and robustness. Explorations such as cross-spectral fusion and dynamic parameter adjustment provide multiple feasible paths for optimizing 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

[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.