Progress on Vision-Based Mobile Robot Target-Tracking Systems

Hua Chen

College of Mechanical and Electrical Engineering, Hohai University, Changzhou, 213200, China,
2261710106@hhu.edu.cn

Abstract. With the rapid advancement of mobile robotics, vision-based target tracking has shown great potential in intelligent warehousing, service robotics, and autonomous driving. As a key technology for autonomous navigation and environmental interaction, target tracking plays a crucial role in enhancing the intelligence of robotic systems. However, tracking accuracy, real-time performance, and robustness remain significant challenges, especially in complex environments with occlusions, lighting variations, and dynamic targets. Traditional tracking methods often struggle with feature extraction and adaptability, limiting their effectiveness in real-world applications. Therefore, this study explores deep learning-based target tracking methods, focusing on CNNs, RNNs, Transformers, and multimodal fusion techniques. These methods enable automatic feature extraction and improve tracking performance under challenging conditions. The research findings highlight significant advancements in tracking accuracy and robustness, demonstrating superior results in mobile robotic applications. Despite improvements, challenges in real-time efficiency and adaptability remain. Future research should focus on lightweight network architectures, multimodal data integration, long-term tracking with re-identification, and algorithm optimization. This study provides a comprehensive reference for researchers and contributes to the further development and application of target tracking in mobile robotics.

Keywords: Mobile Robot, Target Tracking, Deep Learning, Feature Fusion

1 Introduction

Target tracking is a critical technology for mobile robots, enabling them to autonomously navigate and interact with dynamic environments. With the continuous advancement of artificial intelligence and robotics technology, mobile robots have become increasingly prevalent in various domains, including industrial automation, healthcare, and service industries. One of the fundamental technologies enabling mobile robots to achieve autonomous navigation and efficient interaction with their environment is target tracking. An effective target-tracking system is essential for enhancing the intelligence and adaptability of mobile robots, enabling them to perceive, locate, and follow dynamic objects accurately. However, in complex real-world

J. Iqbal (ed.), *Proceedings of the 2025 2nd International Conference on Mechanics, Electronics Engineering and Automation (ICMEEA 2025)*, Advances in Engineering Research 272,
https://doi.org/10.2991/978-94-6463-821-9_53

scenarios, target tracking faces numerous challenges such as occlusions, illumination variations, background clutter, and fast-moving objects. These factors significantly impact the performance of traditional tracking methods, limiting their accuracy, real-time capability, and robustness.

Over the years, significant research efforts have been devoted to developing reliable target-tracking algorithms. Traditional methods, such as correlation filters and feature-based tracking techniques, have made notable progress in improving tracking accuracy. However, they often struggle to handle complex scenes with severe occlusions, abrupt target appearance variations, and background interference. With the rise of deep learning, target tracking has witnessed revolutionary advancements. Deep learning-based tracking methods leverage end-to-end learning to extract deep semantic features from target objects, demonstrating remarkable robustness and adaptability in dynamic environments.

Several notable deep learning-based approaches have been proposed to enhance tracking performance. Existing studies have shown that convolutional neural networks (CNNs) significantly enhance target representation capabilities through a hierarchical feature extraction mechanism [1].

Recurrent neural networks (RNNs) have been employed to model temporal dependencies in sequential data, allowing better motion prediction for target tracking [2]. More recently, Transformer-based architecture has been introduced, offering superior capability in capturing long-range dependencies and global contextual information, significantly improving tracking precision and robustness [3]. Additionally, multimodal fusion techniques have been explored to integrate visual, LiDAR, and infrared sensor data, further enhancing tracking reliability under challenging conditions such as low-light environments and occlusions.

Despite these advancements, current tracking models still face several critical challenges. One major limitation is the trade-off between model complexity and computational efficiency. While deep learning models achieve high accuracy, their large-scale architecture often requires significant computational resources, making real-time implementation difficult on resource-constrained mobile robots. Furthermore, existing methods struggle with generalization across different scenarios, as models trained on specific datasets often fail to adapt to unseen environments. Another pressing issue is long-term target re-identification, where the model needs to recognize and track a target after temporary disappearance due to occlusions or out-of-view situations. Addressing these challenges is crucial for developing a more robust and adaptable tracking system.

To overcome these issues, researchers have explored various optimization strategies. Lightweight neural network designs focus on reducing model complexity through techniques such as network pruning, quantization, and knowledge distillation, enabling efficient inference on embedded systems. Real-time optimization algorithms employ efficient feature extraction and adaptive update mechanisms to enhance tracking speed without compromising accuracy. Additionally, sensor fusion methods integrate information from multiple sources, such as vision, depth sensors, and IMU data, to improve tracking robustness in diverse environments. These approaches have demonstrated promising results, but further research is needed to achieve an optimal balance between accuracy, efficiency, and adaptability.

This paper aims to provide a comprehensive review of vision-based mobile robot target tracking technologies, analyzing key techniques, application scenarios, and future research directions. By summarizing recent advancements and challenges in the field, this study seeks to offer valuable insights for researchers and practitioners, fostering further innovations in mobile robot target tracking. This study highlights the importance of integrating deep learning with efficient computational techniques, multimodal sensor fusion, and adaptive learning strategies to achieve a highly robust and real-time tracking system for mobile robots.

2 Deep Learning-based Object Tracking Methods

In recent years, the rapid development of deep learning technology has brought revolutionary breakthroughs to the field of object tracking. Deep learning-based object tracking methods employ an end-to-end learning approach, enabling automatic extraction of deep features and demonstrating outstanding performance in complex scenarios. This section systematically reviews deep learning-based object tracking techniques from four perspectives: CNNs, RNNs and Long Short-Term Memory (LSTM) networks, Transformer-based object tracking methods, and multimodal fusion approaches.

2.1 The Application of CNN in Object Tracking

CNNs are the most used feature extraction tools in deep learning, with their application in target tracking mainly reflected in feature representation and target matching. The basic principles of CNN are shown in Fig. 1. Early studies, such as MDNet (Multi-Domain Network), employed pretraining and online fine-tuning to adapt to target appearance variations. Subsequently, SiamFC (Fully Convolutional Siamese Network) introduced an efficient Siamese network structure that learns generic target features through offline training and achieves target localization via template matching during tracking. In recent years, CNN-based target tracking methods have further integrated attention mechanisms and multi-scale feature fusion techniques, such as ATOM (Attention-Based Tracking Model) and DiMP (Discriminative Model Prediction), significantly improving tracking accuracy and robustness.

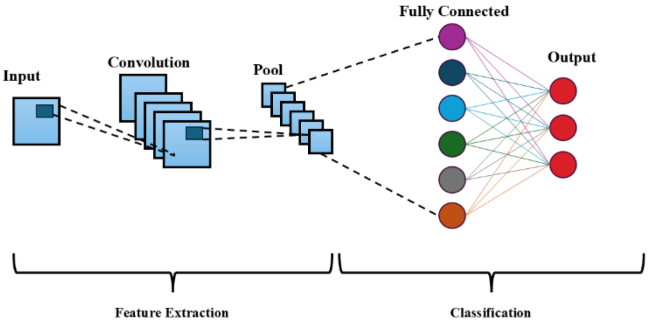

Fig. 1. Basic principles of CNN (Picture credit: Original)

Ding proposed a control method based on CNNs and Adaptive Radial Basis Function (RBF) Neural Networks to address the pose error problem in robotic arm trajectory tracking. By compensating for system uncertainties and optimizing joint velocity computation, the method achieves high-precision and fast-adaptive trajectory tracking control, with experiments validating its superior performance [4]. Wen introduced a novel video SAR moving target detection method using a faster dual-channel fast region-based convolutional neural network (R-CNN). By integrating shadow detection in SAR images with Doppler energy detection in the range-Doppler spectrum domain, this method effectively reduces false alarms and enhances detection accuracy. It outperforms traditional approaches on both simulated and real data, exhibiting fewer false alarms and an acceptable level of missed detection [5]. Lee proposed a single-object detection and tracking algorithm based on CNNs, efficiently combining advanced object detectors with visual trackers. It is particularly suitable for scenarios where only a few static images are available for training. The proposed algorithm was tested on drone detection tasks, and experimental results demonstrated its effectiveness [6].

The application of CNNs in object tracking has been continuously evolving, with significant advancements in feature extraction and target matching. Early CNN-based tracking methods leveraged pre-training and online fine-tuning to adapt to target appearance variations. Subsequently, the introduction of Siamese network structures, attention mechanisms, and multi-scale feature fusion further improved tracking accuracy and robustness. Meanwhile, innovative approaches tailored to specific challenges, such as CNN-based trajectory control and SAR image processing, have demonstrated broad application potential in complex scenarios. Overall, CNN-based object tracking methods have progressed from basic models to diverse solutions capable of efficiently addressing various challenges. Notably, they have achieved remarkable results in enhancing accuracy, reducing false alarms, and handling limited training data. These studies not only drive the advancement of object-tracking technology but also provide more effective and reliable technical support for related applications.

2.2 Applications of RNNs and LSTM Networks in Object Tracking

RNNs and their variant, LSTM networks, are primarily applied in object tracking for temporal modeling and motion prediction. The basic principles of RNN are shown in Fig. 2 and the basic principles of LSTM are shown in Fig. 3. RNNs capture inter-frame temporal information, effectively handling motion blur and fast-moving objects. LSTMs, with their gating mechanisms, address the vanishing gradient problem in long-sequence training, enabling better modeling of long-term dependencies in object tracking. For example, some studies combine LSTMs with CNNs, where CNNs extract spatial features while LSTMs model temporal dependencies, leading to accurate motion trajectory prediction. However, the high computational complexity of RNNs and LSTMs limits their application in real-time tracking.

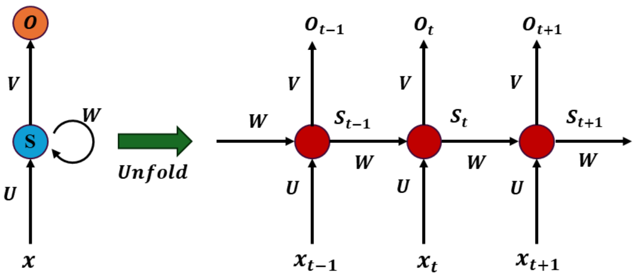

Fig. 2. Basic Principles of RNN (Picture credit: Original)

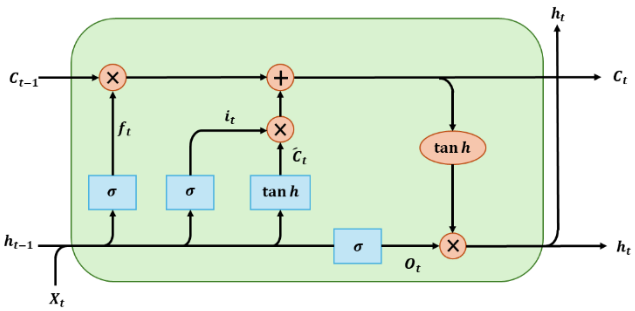

Fig. 3. Basic Principles of LSTM (Picture credit: Original)

Khan proposed a control method based on RNNs for tracking control in surgical robots while satisfying the Remote Center of Motion (RCM) constraint. By formulating the RCM constraint and tracking control as a unified optimization problem, an algorithm called "BAORNN" was introduced to solve the problem in real time. This approach enables the system to simultaneously track the surgeon's commands while ensuring that the RCM constraint is not violated [7]. Yaqi, after analyzing the computational structure of typical filters and Kalman filters, found that they share similarities with RNNs. Based on this insight, a unified filter was proposed by integrating feedforward neural networks, RNNs, and attention mechanisms. This filter unifies the target motion model, adaptive filtering method, and trajectory filtering framework, demonstrating overall superior performance compared to other benchmark filters [8]. Lu designed a multi-average LSTM prediction network to improve the accuracy and real-time performance of surgical instrument pose estimation. By narrowing the recognition range and accelerating algorithm execution, this method ensures effective prediction while achieving an estimation time of less than 1 millisecond [9].

RNNs and LSTMs play a crucial role in object tracking and motion prediction, especially in handling temporal data and long-term dependencies. RNNs effectively capture object motion information, while LSTMs, with their gating mechanisms, overcome the vanishing gradient problem encountered in long-sequence training. Overall, RNNs and LSTMs provide powerful temporal modeling capabilities for object tracking and motion control, though challenges remain in terms of real-time performance and computational efficiency.

2.3 Multimodal Fusion-Based Object Tracking Methods

Multimodal fusion methods enhance the robustness and adaptability of object tracking by integrating data from different modalities, such as RGB images, depth information, and infrared images. The basic principles of Multimodal Fusion are shown in Fig. 4. For example, in low-light conditions or complex backgrounds, RGB images may not provide sufficient information, while depth or infrared images can compensate for these limitations. Multimodal fusion methods typically adopt either feature-level fusion or decision-level fusion strategies. Feature-level fusion combines or weights features from different modalities to generate more discriminative feature representations. Decision-level fusion, on the other hand, integrates the outputs of multiple unimodal trackers to improve tracking stability. In recent years, deep learning-based multimodal fusion methods, such as CMTNet and MMFT, have achieved significant success in complex scenarios, demonstrating improved tracking performance across diverse conditions.

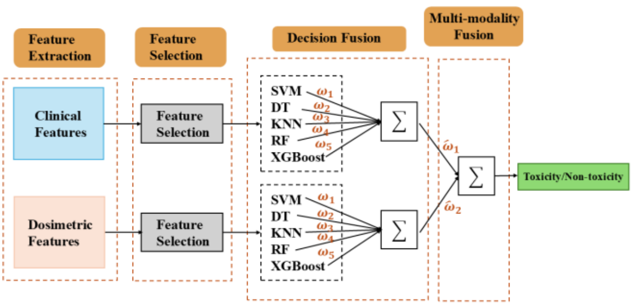

Fig. 4. Basic Principles of Multimodal Fusion (Picture credit: Original)

Zhang proposed a mobile robot target-following system that integrates visual tracking and autonomous navigation. The system achieves target tracking through Kalman filtering and pedestrian re-identification technology, while an autonomous navigation algorithm is employed to address re-following after target loss. This approach demonstrates strong robustness and real-time performance in complex environments, providing an effective solution for re-following target loss [10]. He addressed the challenges of low accuracy, poor real-time performance, and insufficient robustness in traditional object detection and tracking methods in complex driving scenarios by proposing an improved deep learning-based approach. This method features a two-stage detector that integrates an attention mechanism into the backbone network and employs Feature Pyramid Networks (FPN) for multi-scale feature fusion. Additionally, a Siamese network architecture and spatiotemporal attention modules are incorporated to enhance target feature learning. Experimental results show that the proposed algorithm significantly outperforms existing methods on the KITTI and OTB datasets, achieving high accuracy and real-time performance, thus offering new insights for the design of perception modules in autonomous driving [11].

In summary, deep learning-based object tracking methods, leveraging CNNs, RNNs, LSTMs, Transformers, and multimodal fusion techniques, have significantly improved tracking accuracy and robustness. However, key challenges remain in reducing computational complexity while maintaining performance and enhancing adaptability in complex environments, making these critical directions for future research.

3 Mobile Robot Target Tracking System Implementation and Application

3.1 System Architecture and Hardware Platform

The architecture of a mobile robot target tracking system typically consists of three layers: the perception layer, the decision layer, and the execution layer. The perception layer is mainly responsible for collecting environmental information, relying on visual sensors (such as monocular cameras, stereo cameras, and RGB-D cameras) as well as other auxiliary sensors (such as LiDAR, IMU, etc.). The decision layer, based on perception data, uses target detection and tracking algorithms to achieve target localization and motion prediction. The execution layer then controls the robot's movement based on the decision results to complete the tracking task.

The choice of hardware platform directly affects the system's performance and applicable scenarios. As shown in Fig. 5, the mobile robot hardware platform typically includes the perception system, control system, and execution system.

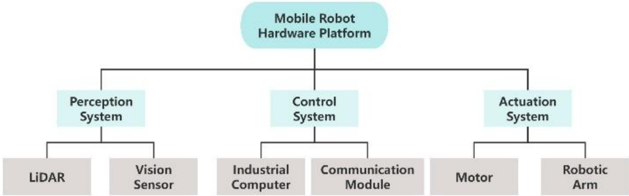

Fig. 5. Composition of Mobile Robot Hardware Platform (Picture credit: Original)

First, the perception system is a key component of a mobile robot for acquiring environmental information. The perception system typically includes devices such as LiDAR (Light Detection and Ranging) and visual sensors. LiDAR emits laser beams and receives reflected signals, enabling precise measurement of the distance and shape of objects around the robot, creating a high-precision environmental map. Visual sensors, on the other hand, capture image information through cameras and, combined with computer vision algorithms, achieve target recognition and tracking. These sensors provide the robot with rich environmental data, laying the foundation for subsequent decision-making and control.

Next, the control system serves as the "brain" of the mobile robot, responsible for processing data provided by the perception system and generating control commands. The core of the control system is typically an industrial computer, which possesses strong computational capabilities and can run complex algorithms such as path planning, target tracking, and obstacle avoidance. The communication module plays a crucial role in the control system, as it is responsible for data transmission between the robot's internal modules, as well as communication with external devices or systems. Through an efficient communication module, the robot can receive commands in real-time and upload status information, ensuring the system's timeliness and reliability.

Finally, the execution system is the part of the mobile robot that enables physical actions, primarily including motors and robotic arms. Motors drive the robot's movement by controlling the motor's speed and direction, allowing the robot to move flexibly in complex environments. The robotic arm is used to perform specific tasks, such as grasping and transporting. The performance of the execution system directly impacts the robot's task execution capabilities, so it must be highly precise and reliable. Table 1 lists of typical components of mobile robot hardware platforms and their key performance parameters:

Table 1. Typical Components of Mobile Robot Hardware Platforms

Typical Component	Key Performance Parameters	Reference
Stereo Camera	Resolution, frame rate, field of view, baseline length, depth measurement range, interface	Szeliski et al. [12]

Row 2	Detection range, accuracy, scanning frequency, angular resolution, laser wavelength, output data	Siegwart et al. [13]
Mecanum Wheel	Diameter, load capacity, material, roller angle, drive mode	Siciliano et al. [14]
Servo Motor	Torque, speed, control accuracy, interface, feedback type	Hughes et al. [15]
WiFi Module	Communication standard, transmission rate, frequency range, transmission distance, power consumption, interface	Rappaport et al. [16]

In summary, the three major systems of a mobile robot hardware platform—the perception system, control system, and execution system—work closely together to enable autonomous navigation and task execution. The perception system provides environmental information, the control system handles decision-making and planning, and the execution system carries out physical operations. This modular design not only enhances the robot's flexibility and adaptability but also lays a solid foundation for the further development of robotics technology.

3.2 Software Framework and Algorithm Implementation

The software framework of a mobile robot target tracking system typically includes modules for data preprocessing, target detection, tracking algorithms, and motion control. The data preprocessing module performs denoising, calibration, and fusion of sensor data to provide high-quality input for subsequent algorithms. The target detection module employs deep learning-based methods (such as YOLO and Faster R-CNN) or traditional methods (such as HOG+SVM) to achieve initial target localization. The tracking algorithm module then utilizes correlation filtering (such as KCF), deep learning-based approaches (such as SiamFC and Transformer-based trackers), or multimodal fusion methods to achieve continuous target tracking.

3.2.1 Transformer-based target tracking methods.

Due to its powerful global modeling capability and parallel computing advantages, the Transformer architecture has been widely applied in target tracking. Unlike CNNs and RNNs, Transformer utilizes the self-attention mechanism to capture global relationships between the target and background, effectively addressing issues such as target occlusion and deformation. Representative algorithms, such as TransT (Transformer-based Tracking Model) and STARK (Spatio-Temporal Attention Tracker), model target tracking as a sequence prediction problem, significantly improving tracking performance. Moreover, Transformer has demonstrated outstanding performance in multi-target tracking and long-term tracking tasks, making it a focal point of current research.

Ma proposed the Unified Transformer Tracker (UTT) to address tracking challenges in both single-object tracking (SOT) and multi-object tracking (MOT). By developing a trajectory transformer, UTT effectively associates target features with tracking framework features, enabling accurate localization in both tasks. The model alternately optimizes SOT and MOT objectives and is trained end-to-end on both datasets. Experimental results indicate that the proposed model achieves excellent performance in both SOT and MOT tasks [17]. Shan introduced the Point-Track-Transformer (PTT) module based on the Transformer architecture to tackle the challenges of sparse LiDAR point clouds and partial occlusions, enhancing the accuracy of 3D single-object tracking. By embedding PTT into mainstream methods, the PTT-Net tracker was developed, applying the PTT module in both the voting and proposal generation stages. Results show that PTT-Net improves performance by approximately 10% over baseline methods on the KITTI and NuScenes datasets while achieving real-time processing at 40 FPS [18]. Zhang designed a Transformer-based target tracking system for mobile robots by improving the MobileNetV2 network and introducing an attention mechanism. A highly efficient feature extraction and fusion method was developed, along with a dual-head prediction module to enhance tracking accuracy. The algorithm achieved a success rate of 0.681 and a precision rate of 0.925 on the OTB100 dataset, outperforming seven mainstream algorithms, and demonstrated accurate and stable target tracking on mobile robot platforms [19].

Transformer-based target tracking methods, with their superior global modeling capability and self-attention mechanism, significantly enhance performance in target tracking tasks, particularly in handling occlusion, deformation, long-term tracking, and multi-object tracking in complex scenarios. Researchers have combined Transformer architecture with traditional tracking approaches, leading to innovative solutions such as UTT and PTT, which improve both tracking accuracy and computational efficiency. As Transformer continues to be widely applied in target tracking, its superior performance has made it a key research direction, showing great potential in real-world applications such as autonomous driving and robotic navigation.

3.3 Typical Application Scenarios

3.3.1 Target tracking in industrial environments.

In industrial environments, target tracking technology is widely used in automated warehousing, logistics sorting, and intelligent manufacturing. For example, in automated warehousing, mobile robots need to track the location of goods or pallets in real time to achieve precise grasping and transportation tasks. These scenarios typically require high tracking accuracy and real-time performance while addressing challenges such as complex backgrounds, occlusions, and lighting variations. The combination of multi-sensor fusion technology and robust tracking algorithms can effectively enhance system performance in industrial settings.

Sun proposed an improved RRT path planning algorithm for the UR3 robotic arm, combining it with the artificial potential field method to improve path search efficiency. By calculating pose errors in real time, a dynamic target-tracking control method for robotic arms was developed and simulated on the ROS platform. Using a visual algorithm optimized with SURF and RANSAC and an Intel RealSense D435i camera, 3D localization of target objects was achieved. Finally, an experimental platform for robotic arms was built, successfully achieving both static and dynamic target tracking [20]. Zhang, addressing the issue of insufficient dynamic performance in robotic vision servo systems, designed a robotic arm vision servo system based on high-speed image processing. By using a high-speed camera to improve image acquisition speed and combining BFS-Canny and Harris algorithms for optimized feature extraction, GPU parallel computing was implemented, increasing edge detection and corner point extraction speeds by 9 times and 4.8 times, respectively. A velocity compensation model was further established to optimize the dynamic tracking performance of the robotic arm. Experimental results showed that the system achieved 110 FPS on 4K images, significantly enhancing robot responsiveness in dynamic scenarios [21].

In summary, target tracking technology in industrial environments has demonstrated significant potential for improvement, particularly in scenarios requiring high precision and real-time performance. Research has made important progress in optimizing path planning, enhancing dynamic target tracking control, and improving image processing capabilities. For example, improved RRT path planning algorithms and high-speed image processing-based vision servo systems have significantly enhanced robotic arm performance in dynamic environments. Additionally, the integration of multi-sensor fusion and robust algorithms provides effective solutions to challenges such as complex backgrounds, occlusions, and lighting variations. These research advancements lay the foundation for further developments in industrial automation and intelligent manufacturing.

3.3.2 Target tracking in service robots.

Target tracking technology in service robots is primarily used for human-robot interaction, object delivery, and scene understanding. For example, in-home service scenarios, robots need to track users or specific objects to provide personalized services. These scenarios demand higher adaptability and generalization capabilities from the algorithms while also considering user privacy and safety. Deep learning-based multimodal tracking methods have shown great application potential in service robots.

Zhang proposed a local multi-scale feature-based Transformer tracking model (LoTPM) to address tracking failures and control precision issues in visual-following robots in complex scenarios. Combined with a depth camera distance measurement algorithm and a sliding mode control method, the model significantly improved target tracking accuracy and system stability [22]. Qian developed a mobile robot system for smart airports, utilizing deep learning and binocular vision technology to detect, track, and estimate the relative speed of target pedestrians. By replacing the traditional VGG network with MobileNet V2, pedestrian detection accuracy was improved. A target tracking method based on KCF and Kalman filtering was proposed, achieving an accuracy of 0.853 and a success rate of 0.731 in experiments. Additionally, binocular vision was used for target distance measurement and relative speed estimation, with errors controlled within 3%-5%. The overall experiment validated the system's high accuracy and real-time performance [23].

Overall, target tracking technology in service robots has significant application value in enhancing human-robot interaction, object delivery, and scene understanding. With continuous advancements in deep learning and multimodal tracking methods, related technologies have demonstrated strong adaptability and generalization capabilities in complex scenarios. For example, by improving Transformer models and integrating depth camera distance measurement algorithms, the accuracy and stability of visual-following robots have been significantly enhanced. Additionally, pedestrian tracking and relative speed estimation systems based on binocular vision have exhibited high precision and real-time performance. As these technologies continue to improve, service robots will provide more precise and efficient services in smart homes, smart airports, and other fields.

3.3.3 Target tracking in autonomous driving.

In the field of autonomous driving, target-tracking technology is crucial for environmental perception and decision planning. Autonomous vehicles must track pedestrians, vehicles, and traffic signs in real time to predict their trajectories and plan safe paths. These scenarios impose extremely high requirements on the real-time performance, robustness, and long-term tracking capabilities of the algorithms. In recent years, Transformer-based multi-object tracking methods (such as TrackFormer) and end-to-end tracking frameworks (such as FairMOT) have achieved significant progress in autonomous driving. However, further optimization is needed to adapt to complex and dynamic traffic environments.

Liu proposed a perception method that fuses radar and camera information. By using Mahalanobis distance for target matching and a joint probability function for data fusion, experimental results demonstrated that radar-camera fusion effectively reduced the missed detection rate under adverse weather conditions, enhanced environmental perception robustness, and provided accurate perception information for autonomous driving systems [24]. You introduced a method that fine-tunes detectors using automatically generated high-quality pseudo-labels. The pseudo-labels were generated by replaying previous driving sequences and were refined using temporal smoothing and extrapolation techniques, effectively addressing missed detections caused by occlusions or long distances. Experimental results showed that fine-tuning the detector with pseudo-labels significantly improved detection accuracy and reliability in new environments [25].

The development and application of mobile robot target tracking systems have made significant progress, yet challenges remain in real-time performance, adaptability, and generalization capabilities. In the future, with the further development of lightweight network design, multimodal data fusion, and long-term tracking technology, target tracking systems will play an even greater role in industrial, service, and autonomous driving applications, driving the intelligence and popularization of mobile robot technology.

4 Conclusion

This paper systematically reviews vision-based target-tracking technologies for mobile robots, analyzing key aspects such as feature extraction, target detection, tracking algorithms, and performance evaluation. By leveraging deep learning, particularly

CNNs, RNNs, Transformer architectures, and multimodal fusion techniques, significant advancements have been made in tracking accuracy and robustness. These technologies have demonstrated considerable potential in various applications, including industrial automation, service robotics, and autonomous driving.

Despite these advancements, several challenges remain. Real-time performance, adaptability to diverse environments, and generalization across different scenarios continue to be major research bottlenecks. Traditional deep learning models, while powerful, often suffer from high computational costs, making their deployment on mobile robotic platforms challenging. Additionally, long-term tracking and re-identification remain open issues, particularly in dynamic and occlusion-heavy environments.

Future research directions should focus on lightweight neural network design, multimodal data fusion, and the integration of long-term tracking with re-identification techniques. Further optimization of tracking algorithms for real-world applications will be crucial in enhancing efficiency and robustness. As these technologies continue to evolve, target tracking systems will play an increasingly vital role in mobile robots, facilitating their adoption in intelligent warehousing, service robotics, and autonomous navigation. This study provides a comprehensive reference for researchers in the field, helping to identify key challenges and opportunities for future innovation.

References

1. LeCun Y, Bottou L, Bengio Y, et al.: ‘Gradient-based learning applied to document recognition’. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
2. Rumelhart D E, Hinton G E, Williams R J.: ‘Learning representations by back-propagating errors’. Nature, 1986, 323(6088): 533-536.
3. Vaswani A, Shazeer N, Parmar N, et al.: ‘Attention is all you need’. Advances in Neural Information Processing Systems, 2017, 30.
4. Ding G, Li S, Fu G P, et al.: ‘Research on robotic arm target trajectory tracking control method based on convolutional neural network’. Machinery & Electronics, 2024, 42(11): 53-57.
5. Wen L, Ding J, Loffeld O: ‘Video SAR moving target detection using dual faster R-CNN’. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 2984-2994.
6. Lee D H: ‘CNN-based single object detection and tracking in videos and its application to drone detection’. Multimedia Tools and Applications, 2021, 80(26): 34237-34248.
7. Khan A H, Li S, Cao X: ‘Tracking control of redundant manipulator under active remote center-of-motion constraints: An RNN-based metaheuristic approach’. Science China Information Sciences, 2021, 64: 1-18.
8. Yaqi C, You H E, Tiantian T, et al.: ‘A new target tracking filter based on deep learning’. Chinese Journal of Aeronautics, 2022, 35(5): 11-24.
9. Lu S, Yang J, Yang B, et al.: ‘Surgical instrument posture estimation and tracking based on LSTM’. ICT Express, 2024.
10. Zhang R, Jiang W Y: ‘Mobile robot target-following system based on visual tracking and autonomous navigation’. Chinese Journal of Engineering Design, 2023, 30(06): 687-696.

11. He H B: ‘Research on target detection and tracking technology for autonomous vehicles based on deep learning’. Special Purpose Vehicle, 2024, (07): 61-65.
12. Szeliski R: ‘Computer vision: algorithms and applications’. Springer Nature, 2022.
13. Siegwart R, Nourbakhsh I R, Scaramuzza D: ‘Introduction to autonomous mobile robots’. MIT Press, 2011.
14. Siciliano B, Sciavicco L, Villani L, et al.: ‘Force control’. Springer London, 2009.
15. Hughes A, Drury B: ‘Electric motors and drives: fundamentals, types and applications’. Newnes, 2019.
16. Rappaport T S: ‘Wireless communications: principles and practice’. Cambridge University Press, 2024.
17. Ma F, Shou M Z, Zhu L, et al.: ‘Unified transformer tracker for object tracking’. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 8781-8790.
18. Shan J, Zhou S, Cui Y, et al.: ‘Real-time 3D single object tracking with transformer’. IEEE Transactions on Multimedia, 2022, 25: 2339-2353.
19. Zhang P: ‘Research on the design of a mobile robot target tracking system based on vision’. Xi'an University of Architecture and Technology, 2023.
20. Sun Y: ‘Research on path planning and tracking control of a robotic arm based on vision’. Henan University, 2024.
21. Zhang Z Q: ‘Visual servo control system of robotic arm based on high-speed imaging’. Jimei University, 2024.
22. Zhang X H: ‘Research on intelligent control system of a following service robot based on vision’. Shenyang University of Technology, 2023.
23. Qian W Q: ‘Research on target detection and following algorithm of airport service robot’. Suzhou University of Science and Technology, 2023.
24. Liu Z, Cai Y, Wang H, et al.: ‘Robust target recognition and tracking of self-driving cars with radar and camera information fusion under severe weather conditions’. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(7): 6640-6653
25. You Y, Diaz-Ruiz C A, Wang Y, et al.: ‘Exploiting playbacks in unsupervised domain adaptation for 3d object detection in self-driving cars’//2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022: 5070-5077