Comprehensive Analysis of Mobile Robot Target Tracking Technology Based on Computer Vision

Hanchen Liu[1, *]

[1] Chongqing University-University of Cincinnati Joint Co-op Institute, Chongqing University, Chongqing, Chongqing, 400000, China
*20206151@cqu.edu.cn

Abstract. With the rapid development of sensor technology and machine learning, the performance and function of robots have been continuously improved, and then they have been applied in many fields. In the research of tracking robots, computer vision technology can be used to identify and track objects, helping robots achieve accurate positioning and tracking. This paper makes a comprehensive analysis of the target tracking technology of mobile robot based on computer vision. Firstly, this paper discusses the background, working principle, and research status of robot target tracking technology based on computer vision. Secondly, the principle and application of visual sensors applied to target tracking are summarized, including monocular vision, multi-vision, and RGB-D. Finally, the local path tracking method of robot is summarized, which mainly introduces the working principle and research status of artificial potential field algorithm, reinforcement learning algorithm and dynamic window algorithm. The authors believe that computer vision tracking is an evolving field that continues to advance artificial intelligence through different methods and technologies. Future research will continue to explore more accurate methods, multi-modal information fusion, and cross-innovation with other fields, opening up broader possibilities for the application of computer vision.

Keywords: mobile robot, target tracking, computer vision, environment perception, local path planning

1 Introduction

Since the last century, robotics has gradually penetrated into industrial production and daily life. With the rapid development of sensor technology, computer vision, and machine learning, the performance and functionality of robots have continuously improved, and their applications are no longer limited to industrial automation. Currently, service-oriented robots are gradually becoming a major trend in the robotics industry. Following robots are a type of service robot that is used to track specific moving targets and can automatically complete continuous tracking of the target. Today, this type of robot is mainly used to provide assistance to special groups

B. H. Ahmad (ed.), *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, Advances in Intelligent Systems Research 180,
https://doi.org/10.2991/978-94-6463-370-2_10

(such as elderly people and disabled people), unmanned logistics handling, and military transportation.

Tracking robots need to acquire and process information about object position and dynamics. Sensor technology is the key to achieving this goal. With the progress of sensor technology, such as cameras, infrared sensors, and laser radars, which can provide more accurate and high-resolution data, they provide a better foundation for the research of tracking robots. In the research of tracking robots, computer vision technology is used to recognize and track objects, helping robots achieve accurate positioning and tracking. In addition, the research on tracking robots is also driven by the demand for automation, intelligence, and personalized services. With social development and technological progress, people's demand for more efficient, precise, and convenient services is constantly increasing, and tracking robots, as one of the technical means to achieve this goal, have received widespread attention and research.

Therefore, this paper provides a comprehensive overview and analysis of mobile robot target tracking technology based on computer vision. The overall structure of this paper is as follows. Chapter 2 describes the applications and current status of computer vision. Chapter 3 classifies and compares robot tracking system controls. Chapter 4 points out the limitations and challenges of visual perception technology.

2 Target Tracking Based on Computer Vision

2.1 Technical Background

Computer vision, as an important branch of artificial intelligence, aims to enable computer systems to mimic human visual systems and achieve understanding and analysis of image and video data. The essence of numerous technologies, including image processing, pattern recognition, machine learning, and deep learning, has been distilled in the development of this field. This convergence has provided robust support for the implementation of practical applications across various domains. In computer vision, object detection, and object tracking techniques are key to achieving robot target tracking. Object detection can help robots identify different types of targets in images and determine their position. Object tracking enables robots to continuously track the motion and changes of target objects in a continuous image sequence.

The early stages of computer vision focused on the processing of digital images and the task of extracting meaningful features from images. Image processing aims to improve the quality of images, enhance target information in images, and reduce noise in images through various algorithms and technologies. It provides a clear image foundation for subsequent image analysis and pattern recognition tasks. Feature extraction is a key step in image analysis, which can extract key information from the image for higher-level analysis. The classic feature extraction methods include edge detection, corner detection, and texture feature extraction. With the development of image processing and feature extraction, object detection and object tracking have become important tasks in the field of computer vision, object detection aims to locate and identify different types of targets in the image, covering from region-based

methods to anchor-based methods, object tracking technology requires the system to continuously track the position and status of objects in continuous images or videos, and is applied in fields such as autonomous driving and video surveillance [1].

However, with the rise of deep learning, feature learning has become more automated and intelligent, Convolutional neural networks (CNN) and other deep learning models can automatically learn features in images through hierarchical convolution and pooling operations, which greatly improves the performance of image analysis. In robot target tracking, by using deep learning models such as CNN, the robot can automatically learn the features in the image, so as to achieve more accurate and robust target detection and object tracking. The deep learning model can be trained through a large number of image data to learn the feature representation of different target objects, so as to improve the performance of robot target tracking [2].

2.2 Working Principle

Computer vision is a technology that aims to realize computer system simulation of human visual ability. Its working principle covers a series of key steps, from data acquisition to final result interpretation and application. First, we need to use the target detection algorithm to locate the object of interest in the image. Traditional target detection methods include feature-based methods (such as Haar feature and hog feature) and machine learning methods (such as support vector machine and random forest), while modern deep learning methods(such as CNN). By using the trained target detection model, we can find the target object in the image and determine its position. Once the target object is detected, the next step is feature extraction. The purpose of feature extraction is to extract representative features from the local area of the target object or the whole image. Traditional feature extraction methods include edge detection, corner detection, color histogram, and so on. In deep learning, CNN can be used to automatically learn the feature representation in images. On the basis of target detection and feature extraction, the target tracking algorithm will track the position and state of the target object in the video sequence through the comparison and matching between successive frames. Common target tracking algorithms include correlation filter-based methods (such as traditional Kalman filter and particle filter) and deep learning-based methods (such as Siamese network and recurrent neural network). These algorithms update the position information of the target by calculating the similarity or feature matching degree between the target object and the surrounding area. Some challenges need to be considered in target tracking, such as robustness in the case of target occlusion, illumination change, angle of view change, etc. [3].

2.3 Research Status

2.3.1 Method Based on Optical Flow

Optical flow is an important concept in the field of computer vision, which is used to describe the motion mode of pixels in the image between consecutive frames. Optical flow-based methods are widely used in target tracking, motion estimation, behavior

analysis, and other fields, by capturing the motion information in the image, it provides key visual information for various applications. The basic principle of optical flow assumes of brightness invariance, that is, the brightness of pixels of the same object remains unchanged between different frames. The method of optical flow calculation covers the technology based on region, dense optical flow, and feature points, which is used to estimate the motion vector of pixels. In practical applications, the method based on optical flow plays an important role. In target tracking, optical flow helps the system identify and track the trajectory of moving targets, it provides support for video surveillance and automatic driving. It can be used to track feature points in an image, such as corner points or edge points. By finding the corresponding feature points in adjacent frames and calculating their motion vectors. It can track the position and motion state of the target object. In addition to tracking discrete feature points, the optical flow method can also be applied to the estimation of dense optical flow, that is, calculating the motion vector for each pixel in the image. By calculating the brightness change between adjacent frames, the motion information of each pixel in the whole image can be obtained, this can be used to track the trajectory and shape changes of the target object more accurately. The optical flow method can also be combined with other image segmentation techniques, such as the segmentation algorithm based on color or texture. By estimating the motion vector of pixels, the pixels belonging to the same target object can be clustered, to extract the contour or region of the target and realize the segmentation and tracking of the target [4].

2.3.2 Method Based on ROI

Region of interest(ROI) method is a vital concept in the area of computer vision. It can achieve more accurate analysis and information extraction by selecting a specific image or video region. This method has been widely used in target detection, image segmentation, face recognition, and other fields. The core of ROI based method is focusing. By selecting the region of interest, it can reduce the computational cost, improve the efficiency, and help to extract the key information about the task from the image. These areas can be determined by manual annotation, automatic detection, etc. ROI-based methods play an important role in many fields of computer vision. In target localization, ROI can help locate the position of the target object in the image. By specifying a rectangular bounding box containing the target as ROI in advance, the target area can be separated from the background area, and attention can be focused on the target. In this way, in the subsequent target tracking task, only motion estimation and position update are needed within the ROI, which reduces the amount of calculation and improves the tracking efficiency. When ROI is used for target recognition and classification, target detection or image segmentation techniques can be used to automatically determine ROI before target tracking begins. Then, the image in ROI is extracted, and the target is recognized and classified by classifier or deep learning model. This enables the target tracking system to adopt different tracking strategies according to different categories of targets and improve the accuracy and robustness of tracking. In some scenes where the shape of the target needs to be tracked, ROI can be used to limit the specific area of the target. By combining ROI with the shape model of the target, the shape change of the target can

be better captured, and the tracking algorithm can be optimized. For example, in face tracking, by defining ROI to constrain the face region, we can realize the tracking of facial expression and posture [5].

3 Robot Tracking Control

3.1 Vision-Based Perception Technology

Environment perception is one of the most important functions of autonomous robot control. Environmental perception performance, such as accuracy, ability to resist changes in light, shadow and noise, and adaptability to complex road environments and severe weather, directly affects the performance of the technology. The use of vision sensors helps with object detection and image processing by analyzing obstacles and traversable areas to ensure the mobile robot reaches its destination safely. Compared with other sensors, visual images contain a large amount of information, including not only information about the distance of objects, but also information about color, texture, and depth. According to the working principle of the camera, visual sensors can be divided into three categories: monocular, multi-camera and RGB-D cameras. Monocular cameras have only one camera, while multi-camera cameras have multiple cameras. A more sophisticated vision sensor is RGB-D, which not only captures color images but also reads the distance between each pixel and the camera for many types of cameras. In addition, by combining vision sensors with machine learning, deep learning and other artificial intelligence, better detection results can be achieved [6].

3.1.1 Monocular Vision

As for the depth estimation of monocular vision, Firstly, the input image is preprocessed, and then the object features in the image are matched to identify. Finally, the distance is estimated according to the size of the target in the database. The monocular camera has the advantages of low computing resource requirements, relatively simple system structure, fast detection speed, and low cost. However, in the recognition and estimation stages, it is necessary to compare with the established sample database. Therefore, monocular vision does not have a self-learning function, so the perception results are limited by the database. Therefore, unlabeled targets are usually ignored, giving rise to the problem that unusual targets cannot be identified. For monocular depth estimation applied to autonomous driving, the main targets are vehicles and pedestrians, so geometric relation method, data regression modeling method and inverse perspective mapping can be used. In addition, the Structure Form Motion (SFM) ranging can be realized by combining the motion information of the vehicle. At present, the road condition judgment applied to automatic driving is mainly based on monocular visual perception.

3.1.2Multi-Vision

Binocular depth estimation depends on the parallax generated by two cameras arranged in parallel. Parallax refers to the horizontal distance between the projection points of the same object in two images, which is inversely proportional to the distance from the object to the camera. According to the formula $Z=(b \cdot f)/d$, the depth information is obtained by finding the point of the same object and performing accurate triangulation, where Z is the depth information, b is the baseline (i.e. the distance between two optical centers), f is the focal length, and d is the parallax. Binocular vision perception can reconstruct the 3D information of the environment in the case of a public view, which is less dependent on pattern recognition and does not require a lot of data learning. As long as a stable key point on the target is obtained, depth estimation can be performed. However, binocular vision perception also has the following disadvantages. First, if the visual sensor is difficult to obtain the key points in a specific case, the distance measurement will fail. Therefore, there are certain requirements for the texture of the target. Second, the binocular vision system has very high requirements for the calibration between cameras. It requires a very accurate online calibration function.

The principle of operation of the three-eye camera is equivalent to the use of two stereos placed in the same direction and distance. three cameras capture their own images from different angles, and then use the stereo vision matching algorithm to get depth information. By combining three monocular cameras with different focal lengths, the field of view for each camera is different. The three cameras are front-view narrow camera, front-view main camera, and front-view wide field camera. For the camera, the sensitivity will either lose the field of view or lose the distance. The three-eye camera can solve the perception problem better by ordering the three cameras do their own jobs, so they are widely used in the industry. The advantage of the three-eye vision system is to make full use of the information of the third camera, reduce the error matching, solve the ambiguity of the binocular vision system matching, and improve the positioning accuracy. However, the trinocular vision system needs to reasonably place the relative positions of the three cameras, and its structural configuration is more cumbersome than the binocular vision system. Secondly, it is necessary to calibrate and correlate the data of the three cameras at the same time, which is higher, more time-consuming and less real-time.

3.1.3 RGB-D

RGB-D cameras usually contains a color camera, an infrared emitter camera, and an infrared receiver camera so that RGB-D can actively measure the depth of each pixel. In addition, 3D reconstruction based on RGB-D sensors is cost-effective and accurate. This makes up for the computational complexity and lack of guaranteed accuracy of monocular and multi-view vision.

RGB-D cameras are divided into two groups. One is lighting systems, such as Kinect v1 and iphone X. The lighting system is designed to solve the problem of binocular contrast and sensitivity to ambient lighting. It is infrared light, so it is independent of lighting and texture. The principle is to emit a pattern with characteristics onto the surface of an object using an infrared laser, and record the pattern changes due to depth differences. Rather than relying on specific elements on

target objects, the lighting structure allows light to diffuse. Therefore, the feature content will not change as the scene changes, greatly reducing the difficulty of matching. Another is the time-of-flight (TOF) method, such as Kinect v2 and Phab 2 Pro. The time-of-flight method emits a pulse of light (usually visible light), reflects it back from an object, and then receives the light pulse. By analyzing the flight time, the distance between the measured object and the camera is calculated. TOF measurement accuracy will not decrease as the measurement distance increases, so it is suitable for situations where the measurement distance is longer. However, it has disadvantages such as high-power consumption, low resolution, and poor quality of depth map.

3.2 Local Path Planning

In complex environments, how to plan a path that can avoid obstacles is the main challenge for autonomous following research. As a path generation method for mobile robots from a certain starting point to the target end point, path planning algorithm is one of the hotspots of robot research and an important part of robot motion control. Local path planning is to plan an optimal or sub-optimal path without obstacles from the starting position to the target position in real-time. During the process, data information is perceived by the sensor in an unknown scene, so that the mobile following robot can effectively cope with the dynamic complex environment. This chapter mainly introduces the artificial potential field algorithm, reinforcement learning algorithm, and dynamic window algorithm.

3.3 Artificial Potential Field

Artificial potential field (APF) refers to the artificial establishment of a virtual force field, so that the robot is simultaneously repulsive to obstacles and attractive to target points in the environment. Among them, the gravity is proportional to the distance from the robot to the target point, and the repulsion is inversely proportional to the distance from the robot to the obstacle. Under the action of the resultant force, the robot will go to the next target point. The APF algorithm has the advantages of simple structure, fast response, smooth planning path, and so on. However, APF still has many problems. For example, the target point near the obstacle has a small resultant force, which makes it difficult for the robot to reach and fall into local optimum. Moreover, it is easy to produce the problem of left and right oscillation when passing through a narrow path.

Sheng et al. proposed to improve the APF method by considering deformation factors, so as to achieve effective protection in the complex UAV flight environment with high speed and large inertia. By obtaining the velocity vector field around the problem, the direction angle control instructions can be obtained, thereby realizing the avoidance of rapid disturbances. Priyanka Sudhakara et al. proposed Enhanced Artificial Potential Field (E-APF) to solve problems that classical APF cannot adapt to complex trajectory planning and is easy to fall into local optimal solution[7, 8]. This approach does not take into account the traditional effects of attraction and

repulsion. The repulsive potential is determined by discretizing the contours of various types of obstacles and their boundary points using the repulsive force function.

3.2.2 Reinforcement Learning

Reinforcement learning is a trial-and-error learning of inform, ation interaction between sensors and the environment, which makes rewards and punishments for the behavior of mobile robots in different environments, so as to achieve the best by constantly optimizing the action plan. Reinforcement learning can better cope with complex dynamic environments, but its learning ability is difficult to improve. When facing complicated environments, the calculation convergence speed is often slow, which limits the actual performance of real-time local path planning.

Maw et al. proposed a hybrid path planning algorithm, introducing deep reinforcement learning for local path planning of autonomous drones. By using reinforcement learning method for local planning between way-points, it enhanced the ability of adapting to the environment composed of static and moving obstacles and realize real-time collision avoidance for the UAV mission planning system. Wang et al. proposed a learning-based technique using environmental spatiotemporal data and introduced global guided reinforcement learning (G2RL) [9, 10]. This approach introduces a new reward model that can be extended anywhere. Therefore, inefficiencies caused by repeating the path process in the presence of dynamic obstacles can be avoided.

3.2.3 Dynamic Window Approaches

Dynamic Window Approaches (DWA) is a sub-optimal method based on predictive control theory, because it can avoid obstacles safely and effectively in unknown environments. Moreover, it has the characteristics of small amount of calculation, rapid response, and strong operability. The DWA algorithm mainly includes velocity sampling, trajectory prediction (estimation), and trajectory evaluation. Firstly, the robot speed samples are collected by the mathematical model of the robot, and the motion trajectory generated within a period of time at the sample speed is predicted and simulated. Then, the standard evaluation of these motion trajectories is performed to select a set of optimal trajectories. Finally, the robot will move according to the optimal trajectory. DWA method transforms the position change into linear velocity and angular velocity control, and transforms the obstacle avoidance problem into a motion constraint problem in space. In this way, the local optimal path can be selected through motion constraints. At present, although some achievements have been made in the research of DWA algorithm, there are still the following problems: In complex environment, it is difficult for robots to plan the optimal path in dense obstacle areas by using traditional DWA algorithm. The obtained trajectory is also not smooth, resulting in too long obstacle avoidance time.

Zhang et al. proposed a local path planning method based on improved DWA, aiming to deal with the problems that the robot cannot adapt to complicated environment quickly and the path is sometimes unstable using traditional DWA [11].

They converted the evaluation stuff form the angle difference to the distance to the target so that unsuitable path due to vibration can be avoided. It allows the USV to find an efficient and stable path in complicated environments. Zeng, D et al. proposed a multi-module enhanced DWA (MEDWA) algorithm for complex environments with dense obstacles [12]. It is based on the multi-obstacle coverage model, combined with Mahalanobis distance, Frobenius norm and covariance matrix in order to improve the ability of judging obstacles in obstacle-intensive areas.

4 Conclusion

Firstly, this paper explores the technical background, working principle, and research status of computer vision. In the technical background, it includes the origin, development, and application fields of computer vision, from image acquisition, pre-processing, to feature extraction, model training, and result interpretation. In the research status based on optical flow, the basic principle, application fields, challenges, and the influence of deep learning are discussed.

Secondly, this paper describes the principles and applications of visual sensors, including monocular vision, multi-view vision, and RGB-D. Monocular vision has low cost and simple structure, but it may encounter problems such as the inability to obtain depth information and uncertain size. Multi-view vision can achieve more realistic 3D visual effects, but it requires high performance of the computing unit and is computationally slow. RGB-D can obtain high-resolution depth maps with little influence from the object's own color, but it is not suitable for dynamic scenes and has high computational and cost requirements.

Finally, this paper mainly lists the working principles and research status of APF algorithm, reinforcement learning algorithm, and DWA algorithm for path planning. The APF algorithm is simple, responsive, and smooth in path planning, but it is prone to get stuck in local optimization and may cause oscillation in narrow environments or unreachable targets near obstacles. The reinforcement learning algorithm has strong planning ability in complex environments, but its learning ability is difficult to improve, adaptive ability is poor, and convergence speed is slow. The DWA algorithm is sensitive to the environment, has high real-time performance, and low computational complexity, but it produces longer planned paths and low path-motion matching.

Mobile robot target tracking is a comprehensive technology that combines graphics and image processing, pattern recognition, mechanical electronics, multi-sensor fusion, machine learning, kinematics, dynamics, and other disciplines. Scholarly research on target detection and tracking technology has been ongoing for decades and remains a hot topic in both academic and practical applications. Currently, there is still significant room for development in autonomous following robot technology to meet commercial needs, especially for following robot products based on visual guidance. In summary, computer vision tracking is a constantly developing field that continues to advance artificial intelligence through different methods and technologies. Future research will continue to explore more precise methods,

multi-modal information fusion, as well as cross-disciplinary innovations with other fields, opening up broader possibilities for the application of computer vision.

References

1. Voulodimos, Athanasios, et al. "Deep learning for computer vision: A brief review." Computational intelligence and neuroscience 2018 (2018).
2. Jarvis, Ray A. "A perspective on range finding techniques for computer vision." IEEE Transactions on Pattern Analysis and Machine Intelligence 2 (1983): 122-139.
3. Lu Hongtao, and Zhang Qinchuan A Review of the Application Research of Deep Convolutional Neural Networks in Computer Vision. "Data Collection and Processing 31.1 (2016): 1-17.
4. van der Eijk, Jerine AJ, et al. "Individuality of a group: detailed walking ability analysis of broiler flocks using optical flow approach." Smart Agricultural Technology (2023): 100298.
5. Chityala, Ravishankar N., et al. "Region of interest (ROI) computed tomography." Medical imaging 2004: Physics of medical imaging. Vol. 5368. SPIE, 2004.
6. Liu, Fei, Zihao Lu, and Xianke Lin. "Vision-Based Environmental Perception for Autonomous Driving." arXiv preprint arXiv:2212.11453 (2022).
7. SHENG, Hanlin, et al. "New multi-UAV formation keeping method based on improved artificial potential field." Chinese Journal of Aeronautics (2023).
8. Sudhakara, Priyanka, et al. "Obstacle avoidance and navigation planning of a wheeled mobile robot using amended artificial potential field method." Procedia computer science 133 (2018): 998-1004.
9. Maw, Aye Aye, et al. "iADA*-RL: Anytime graph-based path planning with deep reinforcement learning for an autonomous UAV." Applied Sciences 11.9 (2021): 3948.
10. Wang, Binyu, et al. "Mobile robot path planning in dynamic environments through globally guided reinforcement learning." IEEE Robotics and Automation Letters 5.4 (2020): 6932-6939.
11. Zhang, Lanyong, Yu Han, and Bo Jiang. "Research on Path Planning Method of Unmanned Boat Based on Improved Artificial Potential Field Method." 2022 6th Asian Conference on Artificial Intelligence Technology (ACAIT). IEEE, 2022.
12. Zeng, Dequan, et al. "Microrobot Path Planning Based on the Multi-Module DWA Method in Crossing Dense Obstacle Scenario." Micromachines 14.6 (2023): 1181.