



Research on People Following Technology Based on Multiple Sensors

Zhongyang Li

Department of Mechanical and Aerospace Engineering, Monash University, Wellington Road
Clayton, Victoria 3800, Australia
ryleli@asu.edu.pl

Abstract. With the continuous expansion of application scenarios for service robots, person-following technology has become a key capability for enabling human-robot interaction and intelligent control. This paper focuses on a multi-sensor fusion-based person-following system and systematically reviews its core technologies, including target recognition, feature extraction, path planning, and obstacle avoidance. First, the features of typical followable targets are categorized into five aspects: visual appearance, geometric structure, motion trajectory, multi-modal perception, and pose estimation. Representative methods such as deep learning, skeleton extraction, and trajectory prediction are discussed. Next, the mainstream path planning strategies and obstacle avoidance mechanisms are reviewed. Finally, an analysis of existing prototype systems reveals that current person-following technology still faces challenges in occlusion handling, real-time embedded computing, and ensuring user experience and comfort..

Keywords: Multi-sensor Fusion, Person-following Technology, Target Recognition, Path Planning, Obstacle Avoidance.

1 Introduction

With the continuous development of service robots and intelligent systems, person-following technology, as a key robotic capability, has shown great potential in diverse fields such as logistics handling, home companionship, exhibition guidance, and security patrol [1, 2].

Unlike traditional path planning methods, person-following systems must identify specific target individuals and be capable of real-time perception, intelligent decision-making, and flexible control in dynamic and complex environments. These challenges place higher demands on system recognition accuracy and environmental adaptability.

Over time, the recognition methods in person-following systems have evolved from traditional vision-based approaches relying on colour or templates to multi-modal perception systems that fuse data from multiple sensors. With the integration of technologies such as image recognition, deep learning, inertial navigation, and sound source localisation, these systems can maintain stable tracking under challenging conditions such as occlusion, lighting variations, and changes in target orientation.

© The Author(s) 2026

S. Zhang (ed.), *Proceedings of the 2025 International Conference on Electronics, Electrical and Grid Technology (ICEEGT 2025)*, Advances in Engineering Research 292,

https://doi.org/10.2991/978-94-6463-986-5_38

Additionally, integrating trajectory prediction and behaviour modelling further enhances the system's capability to handle dynamic environments [3, 4].

Notably, person-following is not solely a matter of perception and control—it also involves ensuring natural and comfortable human-robot interaction. In practical applications, ensuring that the robot follows at an appropriate distance, angle, and speed to minimise user discomfort has become a key focus of current research [5].

Although significant progress has been made, person-following technology continues to face challenges in target identity preservation, occlusion recovery in complex environments, and ensuring user comfort. This paper systematically reviews recent advances in person-following technology in terms of recognition features, recognition methods, path control strategies, and system implementation. It also highlights the key technical challenges and outlines future directions for research and development.

2 Automatically follow technological developments

In the development of person-following technology, visual recognition, as the most widely used approach in early-stage systems, achieves object localization and tracking by analyzing low-level features such as HSV (Hue, Saturation, Value) color space, image contours, and depth maps. It offers good real-time performance in environments with uniform lighting and simple backgrounds [6–8]. However, traditional vision methods exhibit poor stability under conditions such as occlusion, backlighting, or cluttered backgrounds. To address this, sensors such as LiDAR (Light Detection and Ranging) and IMU (Inertial Measurement Unit) have been gradually introduced to enhance perception capabilities. LiDAR supports obstacle boundary detection and environment mapping via SLAM (Simultaneous Localization and Mapping), while IMU estimates target motion trends through acceleration and direction changes, performing effectively in dynamic tracking scenarios [2, 9, 10].

The development of deep learning has significantly improved the accuracy of person recognition. Convolutional Neural Networks (CNNs) such as YOLO and ResNet are capable of extracting high-level image features and are widely applied in person re-identification tasks to distinguish between individuals with similar appearances [3, 4]. Additionally, skeleton extraction algorithms such as OpenPose and Kinect generate 17 to 25 key points, and gait features like stride and rhythm are used to enhance target identification under conditions involving partial occlusion or back-view perspectives [1, 4, 8]. For trajectory prediction, methods such as Kalman filtering and Long Short-Term Memory (LSTM) networks are employed to model trajectory continuity during short-term occlusions, improving system robustness in dynamic environments [5, 11, 12].

In recent years, as artificial intelligence and robotic perception technologies continue to evolve, person-following has emerged as a comprehensive research direction that integrates multimodal perception, multi-target recognition, and intelligent control.

To overcome the limitations of single-modality perception, researchers have proposed the design of multimodal fusion systems. By fusing RGB images, depth maps,

IMU signals, and sound source localization results at either the feature level or decision level, these systems collaboratively identify targets across heterogeneous data sources, significantly enhancing system adaptability and robustness under environmental disturbances [2, 9, 10].

Person-following systems have achieved initial success in several practical applications. For instance, in intelligent logistics, robots assist workers with material handling in warehouse environments [1]. In home service and companion robots, the robot autonomously adjusts its position in response to user movements to achieve natural human-robot interaction [2]. Furthermore, some wearable systems implement identity binding through Bluetooth or inertial data, enabling commercial deployment in scenarios such as exhibition guidance, supermarket assistance, and airport navigation.

In summary, person-following technology is evolving from “single-modal visual detection” toward a comprehensive stage of “cross-modal perception and intelligent control.” However, further improvement is needed, especially in dense multi-target environments and in re-identifying targets under dynamic occlusion.

In actual deployments, path planning must consider not only obstacle avoidance and efficiency but also social acceptability, ensuring that the robot’s behavior does not cause psychological discomfort to the user or nearby individuals. To this end, recent studies have introduced social behavior modeling, incorporating concepts such as personal space from interpersonal communication into the control framework. By defining minimum safety distances and preferring following angles, robots can dynamically adjust their trajectories to conform to social norms. For example, demonstrates experimentally that when a robot follows at an offset angle from the user’s rear and maintains a buffer zone of 1.2–1.8 meters, users report greater comfort and acceptance during subjective evaluations [5].

In addition, a path planning algorithm based on a minimum social pressure cost function has been proposed, enabling the robot to proactively avoid other pedestrians while maintaining tracking accuracy. This design aligns better with the principle of non-invasive interaction in public spaces.

This paradigm—where path planning is guided by the principle of human-sociality—marks the transition of person-following technology from technical feasibility to behaviorally natural interaction. In the future, further research can incorporate social psychology models and reinforcement learning techniques to enable stable and comfortable following in complex crowd environments.

3 Feature recognition

Humans typically recognize one another based on features such as clothing color, facial appearance, height, body shape, and movement patterns. This recognition process is primarily visual and draws heavily on memory and prior experience. When visual information is limited—for instance, when following someone from behind or at a distance—people tend to rely more on physical features such as clothing and body shape to infer identity. Similarly, in person-following technology, robotic systems must

perceive and recognize human-specific features to achieve continuous and stable tracking.

With the advancement of sensing and computing technologies, the features leveraged by person-following robots have evolved from simple visual cues to multi-modal fused characteristics. Based on current research and representative systems, human recognition features can generally be categorized into five types: visual appearance features, geometric structural features, motion features, multi-modal sensing features, and pose and angle features.

Visual appearance features are the most fundamental and intuitive, relying primarily on image-based color and shape information. Color features—such as clothing color, color histograms, and color distribution in HSV (Hue, Saturation, Value) space, as illustrated in Figure 1—are computationally efficient and well-suited for fast matching. However, these features are often sensitive to changes in lighting conditions and susceptible to background interference [7].

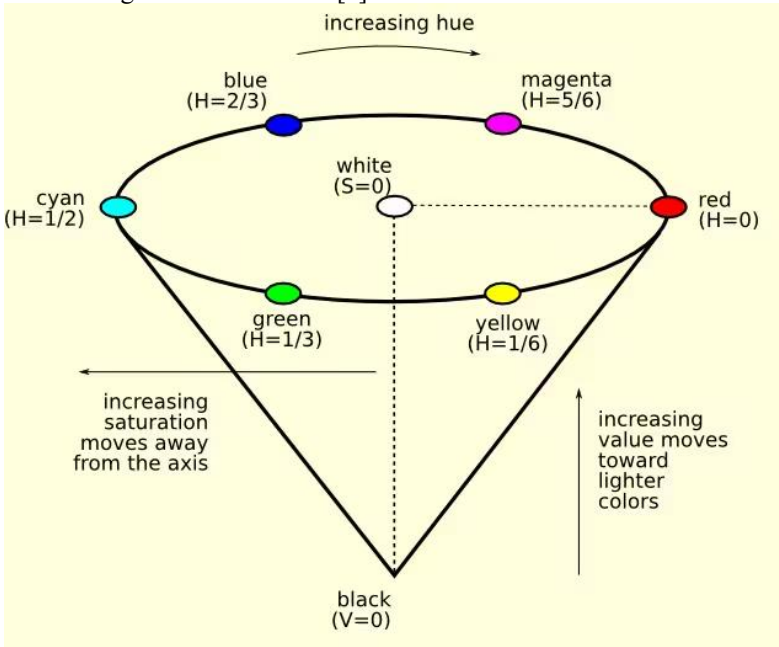


Fig. 1. Hue Saturation Value Color Model

Appearance embedding vectors extracted by Convolutional Neural Networks (CNNs)—such as YOLO, ResNet, and related models—can achieve high-precision identity re-identification, offering strong discriminative power and robustness [3]. In addition, local image texture features, including SIFT (Scale-Invariant Feature Transform) and ORB (Oriented FAST and Rotated BRIEF), are widely used for keypoint matching under conditions involving rotation and scale variation [4, 8].

While facial features can achieve high recognition accuracy, they typically require unobstructed frontal views to function reliably [7, 10].

Geometric structure features describe the physical dimensions and structure of the human body. For instance, a person's height can be estimated using stereo vision or depth cameras and compared against predefined values. As shown in Figure 2, skeleton keypoints extracted by tools such as Kinect, OpenPose, and PoseNet provide 17 or 25 anatomical landmarks across the body, enabling effective action recognition or target re-identification following occlusion [4, 6].

Additionally, upper body contour templates constructed from depth maps can be used to build target shape models. While this approach offers high robustness, it often involves greater algorithmic complexity.

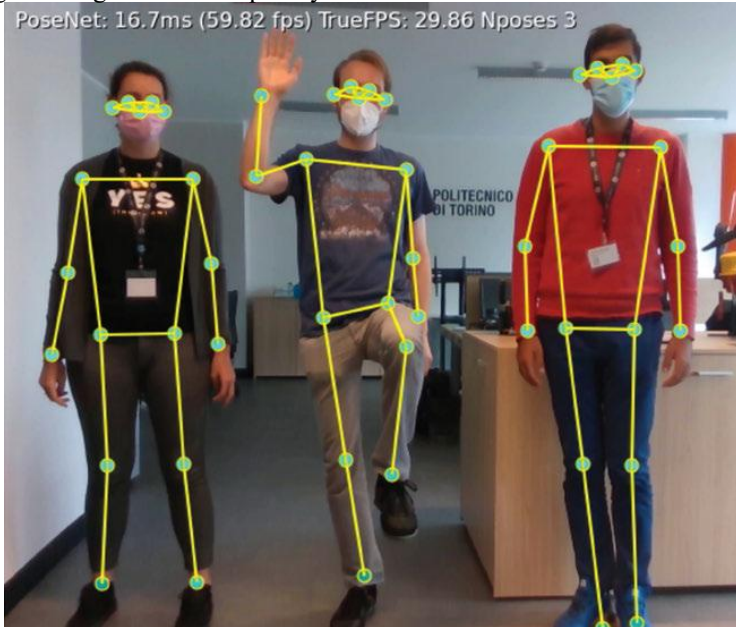


Fig. 2. Example of skeletal structure of multiple individuals predicted by PoseNet [6].

When motion features are utilized, the system leverages the target's dynamic information over time for recognition. Gait recognition identifies individuals based on parameters such as stride length and rhythm, making it particularly effective for rear-view tracking or situations involving occlusion [4].

Trajectory consistency analysis, using algorithms such as the Kalman filter or Long Short-Term Memory (LSTM) networks, can predict the direction of the target, thereby supporting occlusion recovery and path planning [12]. The offset of the person's center point in the image frame can also serve as a real-time control reference. Although it cannot distinguish identity, it is effective for positional correction.

Multi-modal sensing features enhance system adaptability to complex environments by incorporating non-visual data sources. Acoustic features, obtained via microphone arrays, are suitable for low-light environments or voice-interaction scenarios. Inertial features, derived from IMU data, enable identification of motion trends and are well-

suit for identity-bound wearable applications. Ultrasonic data can be applied for boundary detection and short-range error correction.

Pose and angle features are primarily used to assess the target’s orientation and behavioral state. For instance, head orientation and body rotation information can assist in determining whether the target remains the intended person, especially in scenarios requiring human-robot interaction feedback.

In summary, in modern person-following systems, relying on a single feature is insufficient for dynamic and complex environments. As a result, multi-feature fusion has become the prevailing trend. For example, combining color information with skeletal and gait data can significantly enhance recognition accuracy and robustness while maintaining real-time performance.

Deep learning-based appearance embeddings perform effectively in multi-target scenes, while skeletal and gait features provide stable support in complex behavioral contexts, such as occlusion and human turning. Through the integrated use of multimodal features, person-following systems are increasingly capable of achieving accurate perception and continuous tracking of human behavior in real-world environments.

4 Key methods and function implementation

To achieve stable and intelligent person-following functionality, modern systems typically adopt a modular architecture. In this structure, sensors are responsible for perceiving the external environment and target characteristics. Recognition and control algorithms then determine the target’s position and formulate an appropriate following strategy. Finally, motion actuators are driven to perform path tracking and obstacle avoidance control.

Based on the typical workflow illustrated in Figure 3, this section introduces the functional components and commonly used implementation methods of each key subsystem.

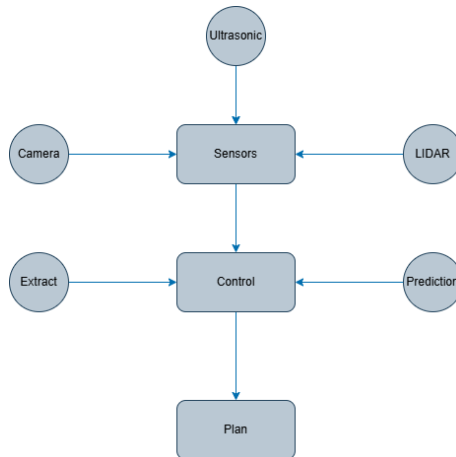


Fig. 3. Work Model of People following Technology

4.1 Perception: groups of sensors

Automatic person-following robots typically adopt a multi-sensor configuration to improve the accuracy and robustness of target detection in complex environments. Common perception hardware includes RGB cameras, which are used to capture image features such as color, appearance, texture, and skeletal structure, providing essential input for visual recognition and person re-identification (Re-ID) tasks [3, 4]. Depth cameras (RGB-D), such as Intel RealSense and Kinect, can output image and depth information simultaneously, enabling height estimation, skeletal key point extraction, and depth template matching [1, 4]. LiDAR and ultrasonic sensors are mainly applied for obstacle detection and short-range boundary identification [8, 11]. IMU-based wearable devices assist with identity binding and movement trend estimation, making them well-suited for individual-following scenarios [2]. Microphone arrays are primarily used for sound source localization, enabling directional perception or voice-interactive following under low-light conditions [2].

4.2 Algorithm: Control module

The core function of the control module in automatic person-following systems is to extract effective distinguishing features from sensor data and to perform identity recognition, position estimation, and trajectory prediction based on these features, thereby enabling continuous and robust following control. Current mainstream target recognition approaches can be categorized as follows.

First, appearance-based recognition utilizes convolutional neural networks (CNNs) such as YOLO and ResNet to extract deep feature vectors from target images for person re-identification (Re-ID). This method demonstrates strong discriminative power and stability in multi-target environments [3]. Second, skeleton extraction algorithms such as OpenPose and Kinect are employed to detect 17–25 keypoints of the human body, which are useful for pose estimation and occlusion recovery [1, 4]. Gait recognition further extracts behavioral features, such as stride length, rhythm, and movement cycles by analyzing skeletal trajectories or contour variations, making it particularly suitable for long-range or rear-view target identification [4, 8]. In addition, trajectory prediction algorithms, including the Kalman filter and Long Short-Term Memory (LSTM) networks, can simulate human prediction behavior to estimate future target positions, thereby reducing tracking failure caused by short-term occlusion or misrecognition [5, 12].

With the advancement of sensor fusion technologies, multimodal fusion recognition has emerged as a critical approach for enhancing system robustness when combined with predictive algorithms. By integrating diverse features—such as color, depth, skeletal structure, and IMU signals—at either the feature level or decision level, the recognition accuracy and stability of the system can be significantly improved in dynamic and complex environments [2, 9].

4.3 Following control strategy: path planning

Once the target is identified and localized, the control module must plan an appropriate motion path based on the following rules and the target's trajectory information. It then adjusts the robot's motion behavior in real time through feedback mechanisms to achieve stable and expected person-following performance. A simple path planning strategy typically involves calculating the robot's linear and angular velocities based on the positional offset of the target in the image. This is achieved by detecting the deviation of the target from the image center, which is the most used method for basic visual following—especially in planar environments with simple structures and no occlusion.

In scenarios involving environmental obstacles or interference from other targets, a composite strategy that combines PID control with Kalman filtering is more effective. By estimating the target's velocity and direction, and integrating prediction and feedback mechanisms, this approach enhances the smoothness and robustness of the control process and improves adaptation to uncertain target movements [12].

For more complex environments, path optimization algorithms such as Dijkstra's algorithm, or the Dynamic Window Approach (DWA) can be applied for real-time trajectory generation and obstacle avoidance [13]. In recent years, some studies have also introduced the concept of social acceptability to develop control strategies based on social behavior modeling. These strategies allow the robot to dynamically adjust its following distance and avoid obstacles in a socially appropriate manner—thereby preventing discomfort caused by proximity that is too close or too distant and improving the perceived naturalness and acceptability of robot behavior in human-robot coexisting environments [5].

4.4 Obstacle avoidance system

To achieve stable and reliable person-following in dynamic environments, automatic following robots must not only accurately track the target trajectory but also account for environmental safety and obstacle avoidance capabilities. Common environmental perception and obstacle avoidance methods primarily rely on RGB-D cameras to detect ground protrusions or obstacle edges and to acquire real-time 3D structural information. For instance, literature [9] employs RGB-D cameras to extract the 3D shape of obstacles ahead, combined with a depth contour template of the human body, which effectively enhances obstacle avoidance performance in unstructured indoor environments.

Traditional obstacle avoidance systems also adopt low-cost approaches such as ultrasonic sensing. As shown in literature [2], ultrasonic data is integrated with visual inputs to build redundant safety mechanisms that compensate for low-light conditions and visual blind spots. Literature [1] demonstrates that combining laser point cloud mapping with path re-planning algorithms, such as the Dynamic Window Approach (DWA), enables stable person-following with obstacle tracking in crowded or multi-target scenarios.

Occlusion remains one of the key challenges in person-following tasks. Literature [12] proposes a trajectory prediction and Re-ID mechanism, in which the Kalman filter predicts the likely position of the target when temporarily occluded, and visual re-identification is subsequently used to relocate the target. This approach effectively mitigates the risk of tracking failure caused by occlusion.

Furthermore, several studies have integrated multiple perception modules to construct multi-sensor fusion obstacle avoidance systems. For example, the synchronous processing of LiDAR and RGB-D data provides richer spatial information and improves the response speed and accuracy of obstacle avoidance. This multi-modal fusion strategy offers a promising direction for enhancing robustness in complex and dynamic environments.

4.5 Hardware implementation

Based on the target recognition algorithms and control strategies, numerous research groups have developed various representative prototype systems for automatic person-following. For instance, literature [1] presents a two-wheel differential robot built on the Kinect platform, which achieves stable human-following control by extracting the upper-body depth template and skeletal key points. Literature [2] introduces a tri-modal recognition system that integrates IMU, microphone, and camera data. In this system, the target individual is required to wear a recognition device, enabling the robot to perform identity binding and directional judgment, thereby improving the specificity and robustness of the recognition process.

In terms of hardware platforms, most current person-following systems employ embedded computing environments such as Raspberry Pi, NVIDIA Jetson, or the Robot Operating System (ROS) framework. These platforms offer a favorable balance between flexibility, scalability, and real-time algorithm deployment.

5 Conclusion

This paper provides a comprehensive summary of the key technologies, system architecture, and primary feature recognition methods employed in multi-sensor fusion-based automatic person-following systems. Based on a thorough review of existing research, the commonly used perceptual features for human target recognition are classified into five categories: appearance, structural and geometric features, motion dynamics, multi-modal perception, and pose-related attributes. Representative algorithms and implementation methods for each feature type are analyzed in detail. Key modules—such as skeletal keypoint recognition using OpenPose, appearance-based Re-ID feature extraction using CNNs, and trajectory prediction via Kalman filtering—are also discussed in depth.

The literature indicates that person-following technology has evolved significantly, transitioning from early-stage color-based tracking to intelligent systems that integrate RGB-D cameras, IMUs, microphone arrays, and deep learning techniques. By incorporating multi-source information fusion and predictive algorithms, the robustness

of these systems has been greatly enhanced, particularly in challenging scenarios involving occlusion, lighting variation, and dynamic environments.

Despite these advancements, several challenges remain. The system may still fail under extreme conditions, such as complete occlusion or poor lighting, particularly when relying on single-modal features. Additionally, while deep learning-based approaches offer high recognition accuracy, their heavy reliance on computational resources limits real-time deployment on embedded platforms. Moreover, deep models often require large-scale training datasets and struggle to adapt quickly to individual user characteristics.

Future research should focus on integrating continuous learning mechanisms, enabling person-following robots to self-adapt to changes in user appearance, behavior, and environmental context. Such advancements will be essential for achieving truly robust, personalized, and reliable human-robot interaction in real-world applications.

References

1. J. Miura, J. Satake, M. Chiba, Y. Ishikawa, K. Kitajima, H. Masuzawa. Development of a person following robot and its experimental evaluation. In Proceedings of the 11th International Conference on Intelligent Autonomous Systems, Ottawa, Canada, (2010), pp. 89–98.
2. A. K. Sharma, A. Pandey, M. A. Khan, A. Tripathi, A. Saxena, P. K. Yadav. Human following robot. In Proceedings of the 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, March 2021, pp. 440–446. IEEE.
3. M. J. Islam, J. Hong, J. Sattar. Person-following by autonomous robots: A categorical overview. *Int. J. Robot. Res.*, 38(14), 1581–1618 (2019). <https://doi.org/10.1177/0278364919881684>
4. M. N. A. Bakar, M. F. M. Amran, A study on techniques of person following robot. *Int. J. Comput. Appl.*, 125(13), 27–30. (2015). <https://doi.org/10.5120/ijca2015906165>
5. Montesdeoca, J, J. M. Toibero, J. Jordan, A. Zell, R. Carelli. Person-following controller with socially acceptable robot motion. *Robotics and Autonomous Systems*, 153, 104075. (2022)
6. A. Eirale, M. Martini, M. Chiaberge. Human following and guidance by autonomous mobile robots: A comprehensive review. *IEEE Access*, 13, 60694–60713. (2025). <https://doi.org/10.1109/ACCESS.2025.3387293>
7. Z. Chen, S. T. Birchfield. Person following with a mobile robot using binocular feature-based tracking. Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, USA, 2007, pp. 815–820.
8. K. Koide, J. Miura, Identification of a specific person using color, height, and gait features for a person following robot. *Robotics and Autonomous Systems*, 84, 76–87. (2016)
9. M. Kristou, A. Ohya, Target person identification and following based on omnidirectional camera and LRF data fusion. In 2011 RO-MAN: The 20th IEEE International Symposium on Robot and Human Interactive Communication (pp. 419–424). IEEE. (2011). <https://doi.org/10.1109/ROMAN.2011.6005281>
10. T. Sonoura, H. Nakamoto, M. Nishiyama, N. Matsuhira, S. Tokura, T. Yoshimi, *Mobile Robots Navigation* (INTECH Open Access Publisher, Rijeka, 2008)

11. L. Pang, Z. Cao, J. Yu, P. Guan, X. Chen, W. Zhang, A robust visual person-followZhang, (ach for mobile robots in disturbing environments. *IEEE Syst. J.*, 14(2), 2965–2968, (2020), <https://doi.org/10.1109/JSYST.2019.2942953>
12. F. Rafi, S. Khan, K. Shafiq, M. Shah, Autonomous target following by unmanned aerial vehicles. In *Unmanned Systems Technology VIII*. Orlando, USA, 2006, pp. 325–332.
13. Y. Cao, K. Ni, T. Kawaguchi, S. Hashimoto, Path following for autonomous mobile robots with deep reinforcement learning. *Sensors*, 24(2), (2024). <https://doi.org/10.3390/s24020527>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

