Proceedings of CONF-MLA 2026 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2026.LD36845

A Review on the Application of RGBD Human Motion Capture in Human Motion Analysis and Mobile Robot Following Control

Jinyu Li

Faculty of Engineering, University of Macau, Macau, China
UC32835@umac.mo

Abstract. Low-cost RGB-D sensorsenable practical human motion capture and robot following, key capabilities for service robotics and human-robot interaction. Unlike expensive, high-precision systems, RGB-D sensors offer affordability, easy deployment, and real-time performance, making them ideal for civil robotic applications. This paper reviews RGB-D-based human pose estimation, motion feature extraction, and mobile robot following control, centering on low-cost, deployable solutions. The research sorts out low-cost RGB-D sensor schemes, human pose estimation algorithms, motion feature extraction pipelines, and tracking control strategies for mobile robots. Existing bottlenecks are summarised, including deteriorated pose accuracy under occlusion, insufficient real-time performance of lightweight algorithms, and unstable following behaviours in complex surroundings. Based on a comprehensive survey of current literature, this study concludes that the combination of lightweight visual models and optimised control algorithms serves as a vital development direction to construct low-cost and highly reliable human-following robotic systems. Prospective research trends of integrated RGB-D motion capture and robot following platforms are further discussed, which can offer theoretical and engineering references for relevant follow-up investigations and practical deployments.

Keywords: RGB-D sensor, human motion capture, human pose estimation, mobile robot, human-following control

1. Introduction

The rise of consumer service robots and human-machine interaction (HMI) systems has increased demand for low-cost, real-time human motion perception in civilian scenarios. Conventional optical motion capture offers submillimeter accuracy but is prohibitively expensive, hard to calibrate, and requires controlled settings, making it impractical for small mobile robots [1]. In contrast, consumer RGB-D sensors combine color and depth imaging at low cost, enabling lightweight human skeleton estimation. They have become the dominant hardware for human-following robots over the past five years [2]. While RGB-D pose estimation and robot control have advanced separately, few studies integrate human motion analysis with closed-loop following control. Most reviews cover only isolated modules and overlook how motion features drive robot trajectory adjustments [3].

This paper surveys recent work on low-cost RGB-D motion capture, focusing on: (1) RGB-D data pipelines, (2) lightweight motion analysis algorithms, (3) hierarchical robot following controllers, (4) cross-scenario system adaptability. This study analyzes nearly 30 journal and conference papers, compares hardware and algorithm performance, and identifies recurring bottlenecks, including depth occlusion and real-time latency on embedded platforms. The review maps the full pipeline from raw sensor data to robot motion, highlights unresolved challenges, and provides actionable guidance for integrated algorithm design and hardware optimization. Furthermore, this paper also proposes co-optimizing lightweight visual models with adaptive controllers, a practical path toward affordable, reliable service robots [4].

2. Low-cost RGB-D technical motion capture systems

2.1. Overview of RGB-D sensors

RGB-D cameras capture synchronized RGB texture maps and per-pixel depth information through structured light, time-of-flight (ToF), or stereo matching principles, which directly eliminates the 2D-to-3D coordinate transformation limitation of monocular RGB cameras [5]. For civilian mobile robot applications, sensor performance indicators include depth measurement range, frame rate, power consumption, and anti-interference capacity under ambient light changes. Industrial high-end RGB-D devices achieve meter-level depth precision, while consumer-grade hardware prioritizes miniaturization and low power consumption to match embedded robot computing platforms [6]. This subsection classifies mainstream sensor working mechanisms and analyzes their respective applicable scene boundaries.

2.2. RGB-D data preprocessing technology

Raw RGB-D data contains inherent noise, including depth missing holes, depth distortion, and RGB-depth coordinate misalignment, which directly degrades the accuracy of subsequent human key point extraction [7]. Standard preprocessing pipelines consist of three core steps: noise reduction, spatial calibration, and depth completion. Bilateral filtering is widely adopted for depth map denoising to retain human body contour edges while removing random depth jitter, extrinsic parameter calibration corrects coordinate offsets between RGB and infrared depth lenses, depth completion algorithms based on lightweight inpainting networks repair occluded human body depth regions [8].

2.3. Mainstream low-cost RGB-D hardware solutions

Three mainstream low-cost sensor product lines dominate academic and small robot development: Microsoft Azure Kinect, Intel RealSense series, and open-source stereo RGB-D assemblies [2]. Each solution differs in unit price, maximum detection distance, and embedded compatibility. Azure Kinect provides the most stable human skeleton output yet carries higher power consumption, Intel RealSense achieves lower power consumption for miniature robots, self-assembled stereo matching RGB-D equipment minimizes hardware costs but requires complex manual calibration.

2.4. Performance comparison between RGB-D and traditional motion capture

Traditional marker-based optical capture systems rely on dozens of infrared cameras and reflective markers, with tracking error less than 1 mm, but hardware investment exceeds tens of thousands of

DOI: 10.54254/2755-2721/2026.LD36845

US dollars and requires fixed laboratory deployment [9]. RGB-D capture abandons wearable markers and extra camera arrays, reducing overall deployment cost by over 90%, while tracking error rises to centimeter level and faces severe accuracy attenuation under human body occlusion [1]. This subsection quantifies trade-offs of two technical routes from cost, precision, deployment difficulty, and real-time performance dimensions.

3. RGB-D based human pose estimation and motion understanding methods

3.1. 2D and 3D human keypoint pose estimation algorithms

Pose estimation serves as the core module converting raw RGB-D data into structured human motion information. Early 2D pose models only extract human joint coordinates on color images, lacking depth scale information and failing to calculate real-world human body motion trajectories [10].

Modern RGB-D 3D pose algorithms fuse depth constraints into network loss functions and can be divided into two categories: top-down detection-first frameworks and bottom-up keypoint aggregation frameworks (Figure 1). Top-down methods detect human bounding boxes before regressing joint coordinates, delivering higher single-person pose accuracy, bottom-up models simultaneously extract all joint points in the field of view and match human instances, showing advantages in crowded multi-human scenes.

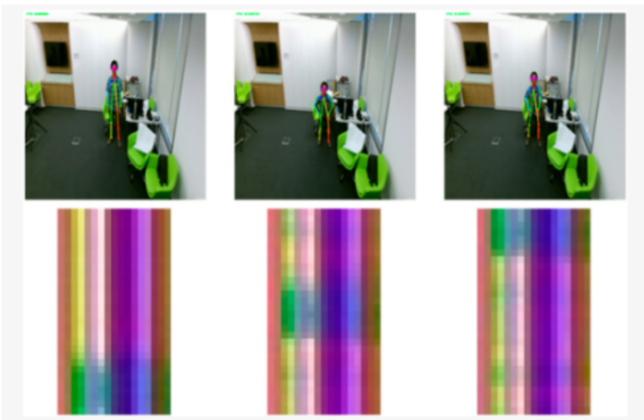

Figure 1. Structural differences between top-down and bottom-up RGB-D 3D human pose estimation networks [10]

3.2. Lightweight human motion feature extraction and behavior recognition

Deploying deep pose networks on low-power embedded robot boards requires lightweight model compression, including network pruning, quantization, and knowledge distillation [4]. After obtaining 3D human joint sequences, temporal motion features such as joint velocity, joint angle change rate, and human body center-of-mass displacement are extracted to recognize human walking, stopping, turning, and obstacle avoidance behaviors. Lightweight feature classifiers such as tiny LSTM and depthwise separable CNN achieve real-time behavior recognition at less than 10 ms per frame on ARM embedded chips, where depthwise separable convolution decouples spatial filtering and channel-wise feature mapping to cut redundant floating-point operations, while tiny LSTM shrinks hidden state dimensions and discards redundant fully connected layers to reduce temporal computation overhead, combined with 8-bit uniform quantization proposed in Wang's

DOI: 10.54254/2755-2721/2026.LD36845

research, the hybrid lightweight architecture drastically lowers memory footprint and hardware computing burden for resource-limited robot edge devices [11].

3.3. Difficulties and optimization strategies for RGB-D motion analysis under complex scenes

Occlusion (human-human mutual occlusion, human-object occlusion) and dense crowd scenes form the primary bottleneck limiting RGB-D motion analysis robustness [8]. When human joints are blocked by obstacles, depth data loss leads to keypoint drift or complete loss. Existing optimization solutions fall into three categories: temporal motion prediction models that fill missing joint trajectories, multi-sensor fusion that supplements laser ranging data, and segmentation-aided pose estimation that separates human foreground from background clutter [7]. Crowded multi-human scenarios face instance-matching confusion, improved bottom-up algorithms add human re-identification branches to maintain stable single-target tracking for robot following tasks.

4. Intelligent following control technology of mobile robots

4.1. Overall architecture of robot human-following system

Integrated human-following robot systems generally adopt a hierarchical control architecture consisting of perception, decision, and control layers. As illustrated in Figure 1, the perception layer acquires RGB-D human pose sequences and extracts real-time human motion states, the decision layer generates safe following strategies based on human states and obstacle information, the control layer outputs wheel speed commands to realize stable tracking [12]. Information interaction exists across all layers: human pose prediction results from the perception layer compensate trajectory planning delay in the decision layer, and obstacle feedback from the control layer adjusts human target tracking range in the perception layer.

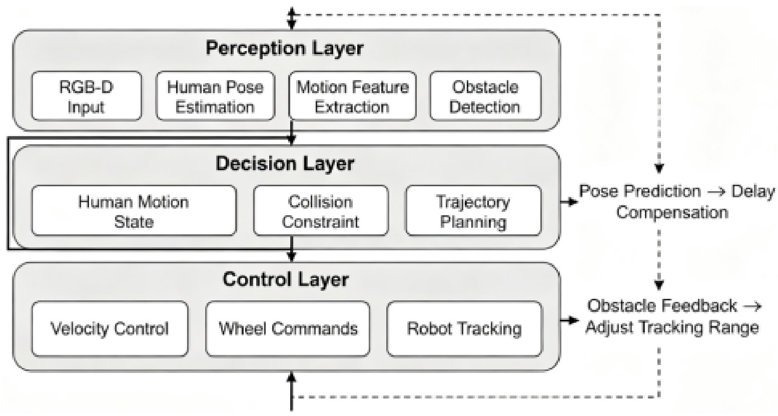

Figure 2. Overall architecture of an intelligent human-following robot system

4.2. Classic human-following control algorithms

Three mainstream control frameworks are widely applied in existing research: PID closed-loop position control, Kalman filter trajectory smoothing and artificial potential field (APF) obstacle avoidance control [3]. Table 1 summarizes the characteristics, advantages, limitations, and typical applications of these representative algorithms.

DOI: 10.54254/2755-2721/2026.LD36845

Table 1. Comparison of classical human-following control algorithms

Algorithm	Advantages	Limitations	Application
PID Closed-Loop Position Control	Simple structure, easy parameter tuning, reliable position tracking	Poor adaptability to abrupt human direction changes	Maintaining target distance and basic following
Kalman Filter Traject-Tajectory Smoothing	Supresses high- frequency jitter from RGB-D pose data,outputs smooth reference trajectories	Lag may occur during fast human motion	Pose estimation and trajectory preprocesssing
Artificial Potential Field Obstacle Avoidance	Unifies attractive following and repulsive obstacle avoidance	Risk of local minimum trapping in narrow corridors	Safe following in cluttered indoor environments

4.3. Human-aware following intelligent motion posture stragy design combined with human

Traditional control algorithms only utilize human centroid distance and velocity, ignoring human body posture characteristics such as turning intention, arm swing and body orientation [5]. Recent research integrates 3D joint angle features into trajectory planners: if pose estimation detects human left/right turning motions, the robot pre-adjusts its heading angle in advance, if the human body leans backward to stop walking, the control layer triggers deceleration logic to maintain comfortable following distance. This posture-coupled control strategy significantly reduces sudden acceleration/deceleration of mobile robots and improves human following comfort in indoor service scenarios.

5. Comparison of existing technical schemes and application scenario analysis

5.1. Performance comparison of integrated low-cost RGB-D motion capture + following control combinations

Different collocations of RGB-D hardware, pose estimation networks, and control algorithms produce distinct overall system performance. Combinations adopting high-frame-rate Azure Kinect + medium-size pose networks + model predictive control achieve the highest tracking stability but exceed embedded computing power limits of miniature robots, collocations with Intel RealSense + quantized lightweight pose models + PID control meet low-power deployment requirements yet degrade accuracy under partial occlusion [4]. This subsection establishes a multi-dimensional evaluation index system, including tracking latency, maximum occlusion tolerance, embedded power consumption and following error, and horizontally scores mainstream integrated schemes.

5.2. Adaptability analysis of schemes under indoor and semi-outdoor scenarios

Indoor office and home environments feature stable ambient light, limited movement space, and frequent static obstacles such as tables and chairs, where RGB-D sensors maintain stable depth output and occlusion becomes the primary interference factor [5].

Semi-outdoor scenarios face strong sunlight infrared interference, which generates invalid depth noise and drastically reduces the effective detection distance of consumer-grade RGB-D cameras.

DOI: 10.54254/2755-2721/2026.LD36845

For semi-outdoor deployment, supplementary fusion with millimeter-wave radar is required to compensate for depth data loss under strong illumination.

5.3. Deployment bottlenecks on embedded platforms

Most state-of-the-art 3D pose networks require desktop GPU computing resources, when transplanted to ARM single-board computers equipped on mobile robots, frame rates drop below 5 FPS and fail real-time following demands [4]. The core embedded deployment bottlenecks include unoptimized network inference pipelines, insufficient memory bandwidth for high-resolution RGB-D streams, and lack of hardware acceleration for depth data preprocessing. Corresponding mitigation measures include model quantization, frame downsampling, and dedicated depth processing peripheral chips.

6. Challenges and future directions

Three universal defects restrict large-scale industrial promotion: first, depth measurement failure under strong outdoor light and heavy occlusion cannot be thoroughly eliminated by single RGB-D sensors, second, the contradiction between pose network accuracy and embedded real-time performance remains unsolved, with high-precision models being unable to run on low-cost robot controllers, third, most existing following control strategies only adapt to single human targets and lack robust multi-human target screening logic in crowded scenes [1, 4, 7]. In addition, current research rarely considers human subjective comfort thresholds, fixed following distance and speed easily generate uncomfortable robot motion shocks.

Dual optimization of hardware and software forms the core solution path for existing limitations. On the hardware side, new miniaturized ToF RGB-D chips with anti-infrared-interference filters are under development to enhance semi-outdoor adaptability [6]. On the algorithm side, two research branches show great potential: multimodal sensor fusion (RGB-D + laser radar + millimeter wave radar) to complement single depth sensor defects and end-to-end lightweight perception-control joint networks that directly output robot motion commands from raw RGB-D frames without separate pose estimation modules. Future research can also introduce human comfort reinforcement learning to adjust robot following parameters dynamically according to extracted human body posture states [5].

7. Conclusion

This review systematically organizes the full technical workflow of low-cost RGB-D human motion capture and its deployment for intelligent human-following mobile robot control, summarizing four core technical modules, including RGB-D sensor hardware and raw data preprocessing, 2D and 3D human pose estimation alongside lightweight motion feature extraction, the three-layer perception-decision-control hierarchical tracking architecture, and performance contrasts of integrated sensing-control solutions across different operating scenarios. Cross-cutting analysis of existing technical research demonstrates that RGB-D motion capture lowers the deployment barrier of human-tracking robots relative to conventional marker-based optical motion capture setups, and adaptive control algorithms with human pose awareness also deliver more stable, comfortable tracking performance for indoor service robot tasks, categorized sorting of available technical solutions further confirms that the combination of lightweight depth vision networks and optimized trajectory planning control represents the most viable technical path for low-power embedded robot hardware.

DOI: 10.54254/2755-2721/2026.LD36845

This review still has notable limitations: the analysis focuses excessively on indoor service robot research with insufficient discussion of semi-outdoor and all-terrain tracking cases due to limited supporting technical data, and cross-scheme performance comparisons depend on scattered experimental metrics from inconsistent test environments without unified hardware benchmark tests, which reduces the impartiality of performance evaluation. Looking ahead, three mainstream research trends are identified based on the above analysis: developing multi-sensor fusion perception frameworks to address depth sensing failures under occlusion and strong sunlight, refining end-to-end lightweight perception-control integrated models adapted to low-power embedded chips, and constructing human comfort-focused reinforcement learning control strategies to dynamically adjust robot motion via real-time human posture features. Advances in these technologies will facilitate broader popularization of low-cost human-following service robots in household, commercial shopping, and elderly care environments.

References

[1] Eirale, A. (2025). Human following and guidance by autonomous mobile robots: A comprehensive review. *IEEE Access, 13*, 4218-4235. https://doi.org/10.1109/ACCESS.2025.3548134
[2] Jiang, Q. Y., & MCO Online. (2026). Budget motion capture hardware guide. *MCO Online*. https://mocaponline.com/blogs/mocap-news/budget-motion-capture-hardware-guide.
[3] Teuliere, C., & Lugo, J. (2018). Person following by autonomous robots: A categorical overview. *arXiv*. https://arxiv.org/abs/1803.08202
[4] Li, H., & Wang, Y. (2024). Lightweight two-layer control architecture for human-following robot. *Sensors, 24*(23), 7796. https://doi.org/10.3390/s24237796
[5] Zhang, L., & Liu, X. (2018). RGB-D sensor based human comfortable following behavior for service robots. *Robot, 40*(6), 721-728. https://doi.org/10.13973/j.cnki.robot.180717
[6] 36Kr. (2026). Flashcap: Millisecond-level low-cost motion capture system. *36Kr*. https://36kr.com/p/3746464082428417
[7] Jiang, L. (2023). Tracking by segmentation with future motion estimation applied to person-following robots. *Frontiers in Neurorobotics, 11*, 1255085. https://doi.org/10.3389/fnbot.2023.1255085.
[8] Map-aware human pose prediction for robot follow-ahead. (2024). arXiv preprint arXiv:2403.13294.
[9] Future Tech. (2025). Simple AI motion capture – no suit needed for realistic animations. *AICourser*. https://aicourser.com/blog/simple-ai-motion-capture-no-suit-needed-for-realistic-animations/.
[10] Lee, J., & Ahn, B. (2020). Real-time human action recognition with a low-cost RGB camera and mobile robot platform. *Computers & Electrical Engineering, 88*, 106872. https://doi.org/10.1016/j.compeleceng.2020.106872.
[11] Wang, Z., & Chen, L. (2024). Quantized lightweight pose estimation network for edge robot computing. *IEEE Internet of Things Journal, 11*(18), 32109-32118. https://doi.org/10.1109/JIOT.2024.3305267.
[12] Kim, J., & Park, S. (2019). Control design for robotic human-following and obstacle avoidance using an RGB-D camera. In *Proceedings of the 19th International Conference on Control, Automation and Systems* (pp. 897-902).