Technologies of Indoor Cleaning Robots

Wenjing Wang *

No. 1188, Yangfan Road, Yinzhou District, Ningbo, China

* Corresponding Author Email: wwwj20071204@qq.com

Abstract. Indoor cleaning robots are with the profound connection of artificial intelligence and robotics, transforming automated tools into smart cleaning assistants. The paper offers a systematic discussion of the technological advances made in two fundamental capabilities of indoor cleaning robots: garbage recognition and handling, and navigation control. The paper is good in the field of navigation outlines the technological development between inertial navigation and SLAM-based path in real-time planning to a Vision-Language-Action (VLA) model-based multi-modal navigation, computation of the benefits, objectives, and difficulties of each strategy. It recognizes such intelligent navigation that merges learning via semantic understanding and reinforcement learning as one of the major future directions. Regarding garbage surveying, the paper discusses vision-based (e.g., YOLO, ViT) and multi-modal sensor fusion solid and liquid waste recognition and classification technologies and potential application of manipulation akin to humans. The review concludes that existing studies continue to have loopholes in the profound assimilation of cross-modal data and system-level execution. Future research will be on smart navigation, multi-mode garbage handling, and delicate anthropomorphic control, finally to achieve extremely autonomous cleaning robot systems based on a closed (perception-decision-action) loop.

Keywords: Indoor cleaning robots, intelligent navigation, multi-modal perception, garbage recognition.

1. Introduction

Indoor cleaning robots have become possible, with the advancement of smart home and service robot technologies becoming a hotspot in research and practice gradually. The classic cleaning machines are largely dependent on mechanical suction and simpler collision detection mechanisms, which are usually afflicted with such limitations as low cleaning ability, partiality or lack of full path coverage, and limited garbage classification capabilities. Over the past few years, various fields of study including artificial intelligence have stimulated these cleaning robots, sensor technology, and robotics, cleaning robots are slowly becoming more humanized intelligence. Through this process of development, two aspects have been investigated through academia and industry; the perception-decision -execution closed loop of cleaning robots. The initial efforts were directed at improving the control architecture underlying and achieving some basic navigation functionality, such as PID optimization of motion control and integration of multiple sensors to offer a path planning framework [1]. In the course of further study, the intelligent nature of robots became the subject of interest, including APP control and path planning. Operational frameworks at the system level have also over time received recognition, such as building ontology construction and cloud service systems which are now targeted to enhance the autonomy and maintainability of robot systems [2, 3]. Moreover, organizational optimization and the incorporation of indoor and outdoor technologies to improve the overall performance have also gained significance as areas of innovation.

However, current literature mostly concentrates on single navigation strategies or garbage handling methods, and still lacks in-depth comprehensive analysis regarding the comparison of different navigation modes, multi-modal perception fusion, and the systematic integration of human-like operation capabilities. This situation leaves researchers and engineering practitioners without a comprehensive reference basis during technological path selection and system design processes.

To address this, this paper systematically reviews and summarizes existing research findings around two core issues. First, this paper summarizes the development of indoor cleaning robot navigation technologies, including real-time path planning navigation, inertial navigation, and multi-

modal navigation methods based on VLA (Vision-Language-Action) models. Second, this paper systematically summarizes multi-modal garbage recognition and processing methods for cleaning robots, focusing on the implementation paths for human-like operation capabilities. Through systematically sorting existing literature, this paper aims to clarify the structure of technological development, assess its application potential, and look forward to future development trends, providing theoretical support and practical reference for further research in this field.

2. Cleaning Robot Navigation (Control) Strategies

Navigation control is the core function of indoor cleaning robots, directly determining their path coverage efficiency and autonomous obstacle avoidance capability. Current mainstream navigation strategies can mainly be divided into inertial measurement-based navigation, real-time path planning-based navigation, and the emerging multi-modal navigation based on Vision-Language-Action (VLA) models. The following sections will review the technical principles, advantages, challenges, and development trends of these strategies.

2.1. Research on Inertial Navigation Control Strategies

Inertial navigation is an autonomous navigation technology based on the Inertial Measurement Unit (IMU). A typical IMU integrates a three-axis gyroscope and a three-axis accelerometer, used to measure the carrier's angular velocity and linear acceleration respectively. By performing integral based on operations against these signals, the relative position and attitude changes of the robot may be determined [4]. The strategy is independent of outside references, has the ability to positively affect path predictability, and fall out of reliance upon random collision navigation modes. But the fundamental problem of inertial navigation is the cumulative error (i.e. drift problem). The small noise and bias of the integration process is steadily intensified by the process of integration itself gyroscope and accelerator, the error in position estimation is greatly growing over time [5]. The practical use of single inertial navigation more often than not needs regular in order to solve this issue calibration or data fusion to other cheaper external sensors (including infrared, ultrasonic) to enhance its path accuracy and system stability during little cleaning and medium-sized cleaning robots. As an example, Joon, A et al. obtained more accurate control of speed and obstacle avoidance through improving IMU and LiDAR sensor data, as well as sophisticated algorithms, such as PID controllers, and others. Artificial Potential Field (APF). The combination of these technologies helped cleaning robots to show stronger navigation skills in fluid and multifaceted indoor settings. [6]. Shan, Tixiao et al. proposed a tightly coupled framework for LiDAR and inertial navigation called LIO-SAM. This method pre-integrates IMU data to calibrate laser point clouds and provides initial values for LiDAR odometry. The obtained LiDAR odometry is then used to estimate IMU biases, further improving the real-time performance of the system [7].

Currently, the most widely used multi-sensor fusion technology is SLAM technology. The concept of SLAM was initially proposed by Professor Hugh Durrant-Whyte and Professor John Leonard from the University of Oxford in the UK in the 1990s [8]. The basic concept of SLAM is that robots acquire environmental information through sensors to complete the analysis and mapping of the surrounding environment. With the development of technology and computational methods, SLAM technology has also made progress in effectiveness and robustness. Based on the mathematical model of SLAM, Li proposed a SLAM observation system framework using LiDAR as the primary perception sensor. He detailed the basic principles and working mechanism of LiDAR ranging, and conducted in-depth analysis on the motion distortion phenomenon generated during movement. Based on the core principles of distortion correction, he researched a LiDAR motion distortion correction method, effectively solving the problem of laser point cloud motion distortion [9].

2.2. Path Planning Navigation Based on Vision

Real-time path planning navigation is mainly achieved based on LiDAR technology. The working mechanism of LiDAR is active scanning, achieving external environment perception through scanning. Its core principle is: projecting modulated light beams within the Field of View (FOV), based on the time-of-flight and phase shift principles, calculating the radial distance of the reflection point relative to the radar origin by resolving the time difference and angle deviation between the transmitted beam and the echo signal. There are two common ranging methods: triangulation ranging and time-of-flight ranging. The so-called triangulation ranging's key technology is to achieve high-precision calculation of the spatial distance between the target object and the radar device by measuring the time interval between laser pulse transmission and reception, combined with the horizontal inclination angle between the sensor and the target [10]. This method possesses high coverage, intelligent obstacle avoidance, and path optimization capabilities, making it a key technology for improving cleaning efficiency and user experience.

Although the hardware and algorithm costs for visual path planning navigation are relatively high, it has become mainstream in mid-to-high-end robots. Therefore, one future research trend is to utilize high-performance hardware combined with deep learning technology to achieve semantic-level map navigation, enabling cleaning robots to understand semantic information in the environment such as "living room" and "dining table", thereby further enhancing their autonomous decision-making capabilities. For example, by learning and understanding the content of the environment map, when a user issues the command "go to the living room", the robot can not only generate an optimal cleaning path but also avoid obstacles in the path in real-time. Through this intelligent integration of semantic understanding and path planning, cleaning robots will possess more flexible and efficient autonomous navigation capabilities, further promoting their application in complex home environments.

2.3. Autonomous Navigation Control Based on VLA Technology - Generalization, Semantic Reasoning

Multi-modal recognition autonomous navigation on the basis of VLA (Vision-Language-Action) model dwells upon the combination of three most significant aspects: vision, language, and action. It aims to enable robots to comprehend natural language instructions, be able to perceive the surrounding environment accurately, and make corresponding behavior in response to visual information and language response instructions, thus developing independent functioning of complex activities. This is one of the major advances that provide robots with better adaptation to complex surroundings, and allows them to become flexible with tasks and smarter language instructions, which have better manifesting efficiency of execution and less manual effort intervention. It greatly extends the application situation variability of robots in their practical application, making their use much more useful [11]. The intelligence is further improved with the multi-modal navigation approach using the VLA model level of cleaning robots. The VLA model combines visual information, language instructions and action planning, which allows the robot not only to comprehend a complicated set of natural language tasks, but also to correctly perform suitable navigation movements. As an illustration, in Hi Robot [12], one model is accountable for converting intricate instructions into simpler instructions, and another model then converts these instructions to actions that are done by the machine. This enables the VLA model to support the management of complex tasks of long sequence and early recognize and trouble-shoot successive malfunctions in the model. The language module gives a vision module, which gives a spatial and obstacle information action module, multi-modal inputs are converted to a particular navigation through task constraints behaviors, realizing the enhancement of environmental adaptability and the ability of the generalization of tasks. Besides, The VLA model increases the robustness of the robot due to multi-modal perception fusion in changing conditions, letting it react to human operating or environmental conditions in real-time changes. Experimental verification of multi-modal navigation in high-end service robots in recent years has shown that it significantly enhances autonomy, semantic understanding capability, and task execution efficiency in complex indoor environments, providing

strong technical support for the future development of humanoid cleaning robots. For example, AutoEval [13] designed an automated scene reset and success detection system, replacing manual operations, and created a round-the-clock autonomous evaluation system, improving the reproducibility and standardization of VLA model evaluation systems. 1XWM [14] proposed the first world model capable of predicting full-body humanoid robot contact and full-body manipulation, achieving precise motion control capabilities.

3. Garbage Recognition Technology for Cleaning Robots

Beyond basic cleaning capabilities, the development of cleaning robots towards garbage processing abilities closer to humans has become a research focus. Among these, multi-modal perception and processing strategies based on sensor data are key to achieving efficient garbage classification and cleaning.

3.1. Recognition Technology for Solid and Liquid Garbage

In the current context of deep integration of artificial intelligence and robot technology, vision-based intelligent garbage classification and processing systems have become an important research direction for environmental service robots. Currently, deep learning technologies such as You Only Look Once (YOLO) and Vision Transformers (ViTs) provide robots with high-precision garbage recognition and rapid environmental detection capabilities. The YOLO algorithm is capable of target completion detecting both speed and accuracy by performing localization and classification in one forward pass. In the meantime, ViTs are effective at transforming global contextual information into pictures, which is done with the aid of the selfattention mechanism, which contributes to increasing the robustness of garbage items classification in complicated scenarios. This integration of the two creates a system of multi-modal visual perception that has the ability to cope with practical needs, that is, change of lights, change of bodies, and varying poses, thus accomplishing effective and precise garbage sorting. As an example, the model of Gao Yan et al. was the YOLOv8 model, which is found to be well-detected since the obstacle detection baseline in indoor cleaning is good robots. It is on this basis that they did targeted model optimization depending on what problems were experienced several times when the pre-trained original model was detecting obstacles. Inertial navigation was also utilized by them merge laser radar with the cooperation to enter the information of garbage and classify into multiple levels operations [15]. In the real world situation that the modern indoor cleaning robots are operating in, types of garbage will show extremely heterogeneous nature; and this primarily consists of solid particles (paper scraps, plastic etc fragments), liquid stains (e.g. beverages, oil stains), and viscous semi-solid substances (e.g. food residues). To identify and categorize such complicated pieces of garbage, A single sensing mode hardly addresses any practical requirements, whereas the conception of multi-modal is forthcoming. This is achievable through perception systems. Through the combination of information, the robot can receive a comprehensive judgment through different sensors like vision, touch and humidity the physical features and morphological peculiarities of the garbage, so that the choice they make is dynamic proper cleaning/processing strategies. For example, liquid garbage requires water absorption or mopping, while solid particles can be handled using suction or grasping modules. Jiang Peng [16] proposed a front-fusion multi-sensor information fusion framework with LiDAR as the main sensor and vision sensor as the auxiliary. First, image preprocessing is performed to get the region for subsequent processing. Then, the segmentation network processes the color image to obtain the segmented image. Further, it processes the registered depth image accordingly, reduces the dimensionality of the depth image, corrects and outputs the LiDAR data, and conducts comparative experiments in real scenes to verify the performance of the information fusion framework from qualitative and quantitative perspectives.

3.2. Appearance-Based Garbage Recognition Technology

Important bases used in classification are the appearance characteristics of garbage in terms of size, shape and texture. Computer vision building deep learning models (CNN or Transformer) can do common classification between garbage objects. In the meantime, the point cloud data can be used to determine the type and shape of garbage. In comparison with object detection task in 2D, 3D object detection is capable of additional determination of the 3D spatial coordinates of the target, its 3D dimensions, and orientation founded on identifying which category the target belongs to and has considerable strengths in spatial perception and understanding of the environment. The technique of using LiDAR point cloud allows to detect objects with high accuracy in three dimensions. This technique may attain the sorting and recycling of various forms of wastes including paper balls, plastic bottles, and metal cans giving a foundation on how future activity can be conducted.

As an illustration, Chen Xiaoxuan [17] developed a point cloud object detection network called PMFE-Point RCNN which was built on the basis of the Point RCNN network. The network can be split into two phases: 3D candidate box generation and 3D bounding box optimization. The backbone network was designed to have a parallel multi-branch feature enhancement module, which includes a channel attention module, spatial reconstruction unit, channel reconstruction unit, and residual structure. This enhanced the quality and strength of the machine detection. He Kaixuan offered an algorithm of Point Att-RCNN. He proposed a non-parametric geometric encoding module (Point-NN) to learn geometric features based on trigonometric functions and multi-level framework aggregation. At the same time, a triple attention network was proposed to dynamically increase the weights of features of regions of interest, with substantial impacts on the detection of targets that had geometric structure [18].

Garbage can be measured by other dimensions such as color and material information. Combining RGB images with spectral sensors, robots are able to distinguish between recyclables, kitchen garbage or dangerous garbage. An illustration of this is that a blue bottle can be identified as recyclable plastic whereas brown food residue will be reference to kitchen waste. Moreover, it is also possible to optimize garbage management methods with material identification (plastic, metal, paper, etc.). The next stage of progress is to have a profound visual-tactile-spectral integration of information which would allow the robots to have human-like capabilities of perception of garbage and operation in complex environments.

4. Conclusion

This paper, This paper concentrating on the two fundamental functions of the indoor cleaning robots which are navigation and garbage collection systematically reviews and summarizes the technological progress, current state of research and technology development trends. Technologically, the route of navigation has been developed through some primitive clashful and inertial based navigation to SLAM based real-time path planning and then has been extended further to include multi-modal form of navigation that incorporates the vision-language-action framework which has greatly increased the sense and autonomy of the decisions made by the robot in any given complex setting. The research interest in the area of garbage processing has moved away from single suction operating modes and instead to recognition and classifying under the term of multi-modal perception, which has incrementally enhanced the level of human-like operation of the robot, which synthesizes the multi-source data, including touch and vision.

Despite considerable advancements recorded in areas like perception architecture, path planning and classifications of existing research, the majority of the modern research is yet to cover the various areas of independent consideration of the technological path. The systematic comparison of the various navigation modes, the deep integration of the multi-modal perception information, and the system integration to real-world scenarios still have research gaps. According to the synthesis of the previous success, this paper demystifies the intelligent development plan whose dominating point is the perception-decision-execution closed loop framework, but the key point is system-level function

and maintenance as well as multi-scenario integration leads to the overall enhancement of robots' performance. The upcoming trends in the development of indoor cleaning robots may take place in the following directions: Intelligent Navigation: The synthesis of semantic maps with reinforcement learning is used to provide environmental understanding and autonomous decision-making; Multi-modal Garbage Handling: Developing an in-depth integration of the visual, tactile, and chemical awareness will help handle the complex types of garbage; Human-like Operation: It entails integrating dexterous robotic arms or cleaning modular units to generate human-like picking, sorting, and disposal capabilities.

In conclusion, cleaning robots are slowly evolving out of the category of automated tools to intelligent cleaning assistants and they have a wide application potential in domestic, offices and outdoor environs.

References

[1] Zhou Shengrong. Research on Intelligent Household Cleaning Robots [D]. Harbin Institute of Technology, 2006.

[2] Guans, Si. Design Innovation of Intelligent Cleaning Robot [D]. Nanchang University, 2018.

[3] Liu Zongkun. Design of Remote Intelligent Operation and Maintenance System for Cleaning Robots [D]. University of Jinan, 2023.

[4] G. Patrizi, M. Carratù, L. Ciani, P. Sommella, M. Catelani and A. Pietrosanto, "Analysis of Inertial Measurement Units Performances Under Dynamic Conditions," in IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1-13, 2023, Art no. 3520713

[5] Woodman, Oliver J. An introduction to inertial navigation. No. UCAM-CL-TR-696. University of Cambridge, Computer Laboratory, 2007.

[6] Joon, A. Kowalczyk, W. Design of Autonomous Mobile Robot for Cleaning in the Environment with Obstacles. Appl. Sci. 2021, 11, 8076.

[7] Debeunne, C., & Vivet, D. A review of visual-LiDAR fusion based simultaneous localization and mapping. Sensors, 2020, 20(7), 2068.

[8] Xu Yu. Research on Autonomous Navigation of Indoor Mobile Robots Based on the Fusion of Vision and Lidar [D]. Wuhan Textile University, 2025.

[9] Ma Chenlong. Autonomous Operation Algorithm for Manipulator Based on Visual Language Model and Imitation Learning [D]. Hangzhou Dianzi University, 2025.

[10] Du Xin. Research and Implementation of Robot Complex Indoor Environment Mapping Based on Multi-sensor Fusion [D]. Xi'an University of Technology, 2024.

[11] Ma Xingzhe. Design of Multi-sensor Fusion-Based Mobile Robot Motion Control and Navigation System [D]. Dalian University of Technology, 2024.

[12] Li Yinqi. Research on Laser SLAM Algorithm for Indoor Mobile Robots [D]. Changchun University of Technology, 2025.

[13] ZHOU Z, ATREYA P, TAN Y L, et al. Autoeval: autonomous evaluation of generalist robot manipulation policies in the real world [A], 2025.

[14] HO D, MONAS J, REN J, et al. 1X World Model: Evaluating bits, not atoms[EB/OL], 2025

[15] SHI L X, ICHTER B, EQUI M, et al. Hi Robot: open-ended instruction following with hierarchical vision-language-action models[J]. CoRR, abs/2502.19417.

[16] Gao Yan. Research on Obstacle Detection for Indoor Cleaning Robots Based on Deep Learning [D]. University of Electronic Science and Technology of China, 2025.

[17] Chen Xiaoxuan. Real-time Detection and Tracking of Orbital Obstacles Based on Point Cloud Analysis [D]. Dalian Jiaotong University, 2025.

[18] He Kaixuan. Research on 3D Point Cloud Object Recognition Algorithm Based on Deep Learning [D]. Xi'an Polytechnic University, 2025.

[19] Jiang Peng. Research on Obstacle Avoidance Strategies for Indoor Mobile Robots Based on Reinforcement Learning.