Proceedings of CONF-MLA 2026 Symposium: Intelligent Systems and Automation: AI, IoT, Robotic Engineering & Algorithm
DOI: 10.54254/2755-2721/2026.LD35522

Research on Scene Perception and Task Semantic Understanding for Indoor Service Robots: Based on Literature Review and Case Analysis

Haoxiang Huang

Xiamen University Malaysia, Kuala Lumpur, Malaysia
huanghaoxiang17@outlook.com

Abstract. With the advancement of artificial intelligence and robotics, indoor service robots are gradually being deployed in complex environments including residences and hospitals to perform diverse tasks. However, relying solely on traditional localization and obstacle avoidance capabilities can no longer meet the complex demands of real-world scenarios. Robots must possess semantic-level understanding of environmental objects, spatial relationships, and service objectives. Currently, environmental representation for indoor robots is undergoing a transition from being geometry-dominated to semantic-enhanced. This study centers on scene perception and task semantic understanding for indoor service robots. It aims to explore how robots build a systematic cognition of scenes, objects, user commands and task workflows based on low-level visual and spatial data. Using literature review and case analysis methods, this paper synthesizes representative achievements in fields such as semantic mapping, semantic navigation, semantic SLAM, explicit knowledge representation, and task planning. The study argues that scene perception and task semantic understanding constitute a continuous intelligent chain from "environment recognition" to "task execution". In the future, indoor service robots require further improvements in multimodal information fusion, knowledge-driven modeling and task reasoning for open scenarios, so as to enhance their practical performance and operational reliability.

Keywords: Indoor service robots, scene perception, task semantic understanding, semantic mapping, semantic SLAM

1. Introduction

In recent years, service robots have gradually moved from laboratories to practical application scenarios such as home care and medical assistance. Compared with industrial robots, indoor service robots operate in more complex environments and perform a wider range of tasks, and need to interact frequently with humans. Therefore, their level of intelligence largely depends on their ability to understand environmental semantics and task semantics. From existing research, scene perception has evolved from geometric map building to semantic mapping, semantic navigation, and semantic SLAM, which integrate object recognition and spatial semantic annotation [1-3]. Meanwhile, task execution has also transitioned from predefined action sequences to high-level semantic reasoning

DOI: 10.54254/2755-2721/2026.LD35522

systems combining knowledge representation and task cognition [4-6]. Robot intelligence is no longer limited to environmental perception; instead, it focuses on converting perceptual data into task-oriented knowledge structures [7].

However, current indoor service robots still have obvious limitations in real environments. Indoor environments feature a rich variety of objects and frequent layout changes, making it difficult for robots to form a stable understanding of the environment relying solely on low-level visual recognition [3]. Dynamic environment adaptation and real-time processing remain key challenges for semantic SLAM [8]. At the same time, user instructions often feature omissions and ambiguity. If a robot lacks task semantic modeling capabilities, it is difficult to convert natural language expressions into executable task sequences [6, 7].

This paper adopts a combination of literature review and case analysis to sort out the conceptual foundations and technical paths of scene perception and task semantic understanding. It also analyzes semantic requirements and existing challenges in typical application scenarios, and puts forward prospective optimization strategies [4, 5]

2. Conceptual foundations of scene perception and task semantic understanding for indoor service robots

2.1. Functional positioning and application scope of indoor service robots

Indoor service robots are primarily deployed in human-centric complex indoor environments such as homes, hospitals, hotels, office buildings, and shopping malls, undertaking tasks such as delivery, guidance, inspection, companionship, interaction, and decision assistance. Unlike robots in structured industrial scenarios, these systems need to handle higher-frequency human-machine interactions, more complex spatial changes, and stronger contextual dependencies. Accordingly, their core competencies cover not only mobility and localization, but also comprehension and interpretation of service contexts. Studies conducted by Han et al. and Song et al. suggest that environmental representation ability has gradually become the cornerstone of high-level intelligence for service robots [1, 2].

2.2. Connotations of scene perception, semantic mapping, and task semantic understanding

Scene perception refers to a robot's ability to acquire and organize information about objects, regions, geometric structures, and contexts using various sensors such as cameras, RGB-D, and LiDAR. Semantic mapping extends geometric maps by integrating additional information including object categories, regional functions and spatial relationships, allowing robots to model indoor environments in a manner consistent with human cognition. Han et al. argue that semantic maps serve as a crucial intermediate layer connecting low-level perception and high-level interaction [1]. Task semantic understanding emphasizes the robot's ability to link environmental information, user intentions, task objectives, and action constraints. Cui et al. contend that service robots in open worlds must possess semantic task representation and reasoning capabilities [6]; Cui et al. further point out that task cognition is a key step from "knowing the environment" to "understanding the task" [7].

DOI: 10.54254/2755-2721/2026.LD35522

2.3. Technical evolution logic from geometric perception to semantic understanding

The shift from geometric perception to semantic understanding essentially reflects the evolutionary trend of robots from spatial localization toward task-oriented cognition. Traditional geometric maps can support navigation and obstacle avoidance but struggle to support complex tasks with object, location, and action constraints, such as "go to the pharmacy to get medicine and deliver it to ward 302". Alqobali et al. point out that semantic navigation enhances task adaptability in indoor navigation by fusing target objects, regional functions, and task contexts [4]. Kanso et al. illustrate that semantic SLAM enables robots to maintain higher-level scene expressions in more dynamic environments by integrating object recognition, scene relationships, and map updates [8]. Furthermore, Aguado et al. stress that ontology and knowledge representation approaches can organize environmental elements, task rules and abnormal scenarios into knowledge structures supporting logical reasoning [5].

3. Current status analysis of related research on indoor service robots

3.1. Scene perception research based on visual recognition and multimodal understanding

In recent years, robotic scene perception has increasingly leveraged deep learning and multimodal sensing to boost the accuracy and integrity of environmental semantic understanding. Ni et al. summarize that modern autonomous robot scene understanding systems typically combine visual recognition, semantic segmentation, context modeling, and cross-modal information fusion to achieve more comprehensive environmental cognition [3]. This means that scene perception is no longer limited to "detecting a certain object" but has further developed into a holistic interpretation of environmental structures, object functions, and spatial contexts. For indoor service robots, this paradigm shift is particularly critical, as service tasks rely heavily on inter-object functional correlations, instead of mere category recognition of individual targets.

3.2. Environment modeling research based on semantic mapping, semantic navigation, and semantic SLAM

Semantic mapping represents one of the core research directions in robotic perception, as it integrates object semantics, regional semantics and spatial structures within a unified representation framework. Han et al. provided an early and systematic review of semantic mapping for mobile robots in indoor scenes, highlighting its foundational role in natural human-machine interaction and task execution [1]. Subsequently, Song et al. reviewed the research progress of semantic mapping techniques for indoor mobile robots from a more recent perspective, emphasizing the extended value of semantic maps in environmental understanding and decision support [2]. Alqobali et al. further note that semantic navigation systems can boost path planning performance and human-robot collaboration capabilities via the fusion of geometric and semantic information [4]. Kanso et al. conducted a systematic review of semantic SLAM methods and applications, arguing that semantic information can enhance map accuracy, depth of environmental understanding, and system robustness, though real-time processing, dynamic scene adaptation, and scalability remain core challenges [8].

DOI: 10.54254/2755-2721/2026.LD35522

3.3. Task semantics research based on knowledge representation, task planning, and task cognition

Task semantic understanding requires robots not only to recognize scenes but also to interpret instructions, infer task constraints, and organize actions accordingly. Aguado et al. point out that ontology-driven and knowledge-enhanced methods help improve the adaptability, interpretability, and reliability of autonomous robot systems, as these methods can explicitly represent object categories, spatial functions, task rules, and anomalous situations [5]. Furthermore, the same research team further demonstrates from the perspective of task cognition and planning that service robots should not only passively parse user commands but also proactively mine task clues according to real-time environmental states, supplement missing task constraints, and formulate executable execution workflows [7]. It is thus evident that task understanding is not a simple extension of scene perception but a high-level semantic process that transforms environmental information into executable goals and plans.

4. Semantic needs and realistic challenges in typical indoor application scenarios

4.1. Service task characteristics in home, hospital, and hotel scenarios

Although indoor service scenarios vary in specific task categories, all of them impose stringent contextual understanding requirements on service robots. In home scenarios, robots often need to recognize personalized items, understand user preferences, and adapt to highly unstructured layouts. In hospital scenarios, robots must not only recognize room functions but also meet high safety and reliability requirements in environments shared by medical staff, patients, and equipment. In hotel scenarios, robots are generally tasked with delivery, guest guidance and standard human-robot polite interaction, and are capable of real-time responses to dynamic environmental variations.

These scenarios collectively illustrate that robots must understand "the roles of objects and regions in specific tasks," rather than just judging "what they are."

4.2. Impact of dynamic scene changes on perception accuracy and task continuity

Indoor environments are inherently dynamic. Furniture can be rearranged, pedestrians may pass by unexpectedly, and task-related objects may be occluded, displaced or even vanish during task execution. These changes can undermine the effectiveness of fixed environmental representations and create new uncertainties between perception results and action decisions. Kanso et al. point out that dynamic scene adaptation is one of the core challenges facing semantic SLAM [8]. Alqobali et al. further contend that indoor semantic navigation systems will fail to sustain stable task continuity in complex scenarios without timely updates to environmental semantic information [4]. Therefore, for service robots, scene perception must be closely linked with task reasoning, enabling the system to continuously revise decisions based on environmental changes.

4.3. Semantic difficulties brought by human-machine interaction, ambiguous instructions, and safety constraints

In real-world usage contexts, human instructions are often incomplete, ambiguous, and even highly context-dependent. For example, a user might just say, "Help me get the medicine," without specifying the type of medicine, its location, or the priority order. In such cases, the robot must integrate linguistic information with real-time scene data, task knowledge and user status to derive

DOI: 10.54254/2755-2721/2026.LD35522

feasible actions. Cui et al. argue that the completion of open-world service tasks depends on the semantic modeling of tasks and environments [6]; Cui et al. also point out that task cognition mechanisms can help robots complete task constraints under incomplete information conditions [7]. Furthermore, in elderly care, medical delivery and crowded public spaces, safety constraints bring in new semantic dimensions, since robots need to balance task execution, social compliance and operational safety simultaneously.

5. Discussion: paths for synergistic optimization of perception and task semantics for indoor service robots

5.1. Direction of combining multimodal information fusion and high-level semantic representation

Future indoor service robots need to fuse multi-modal information including vision, depth, natural language, pose and task context within a unified semantic representation framework. Ni et al. point out that multimodal understanding is an important direction for enhancing perception capabilities in complex environments [3]. Alqobali et al. demonstrate that semantic navigation requires simultaneous utilization of spatial and semantic information [4]. Kanso et al. further prove that the performance improvement of semantic SLAM is inseparable from the synergy between perception information, object relationships, and map representation [8]. It can thus be seen that multimodal information fusion not only helps improve perception robustness but also helps connect scene understanding with task decision-making.

5.2. Knowledge modeling and task reasoning mechanisms for open scenarios

Open-environment service scenarios require robots to not only conduct object recognition, but also master environmental rules and infer follow-up behaviors. Aguado et al. argue that ontology-driven knowledge modeling provides interpretable structural support for robot autonomy [5]. Cui et al. note that semantic task planning enables robots to overcome the constraints of fixed preset workflows in closed scenarios [6]. Cui et al. emphasize that task cognition mechanisms enable robots to actively discover task clues, complete task conditions, and dynamically adjust execution processes in complex home or service environments [7]. Therefore, knowledge modeling and task reasoning mechanisms will be an important bridge connecting environmental semantics and behavior generation.

5.3. Development trends in future applications of indoor service robots

Future indoor service robots will develop towards stronger adaptability, interpretability, and collaboration. From a research perspective, semantic mapping, semantic navigation, semantic SLAM, knowledge representation, and task cognition are gradually converging into a more complete embodied intelligence framework. From an application perspective, the key issue is no longer just whether a robot can move in an indoor environment, but whether it can understand the "service meaning" in the environment and act accordingly. Synthesizing existing studies, we conclude that future service robots need to integrate scene perception, knowledge reasoning, task interpretation and human-robot collaboration capabilities to achieve stable and reliable deployment in real-world environments [1, 4, 5, 7, 8].

DOI: 10.54254/2755-2721/2026.LD35522

6. Conclusion

In summary, the core capabilities of indoor service robots are no longer confined to localization, navigation and basic obstacle avoidance, but are gradually evolving toward comprehensive comprehension of complex correlations among environments, objects, robot behaviors and service tasks. Semantic mapping, semantic navigation, and semantic SLAM provide a transitional framework from geometric space to conceptual space for robots, enabling them to understand indoor environments in a way closer to human cognition. Explicit knowledge representation, open-world task planning, and task cognition further propel robots from "recognizing the world" to "completing tasks." Accordingly, scene perception and task semantic understanding should be regarded as two interconnected hierarchical modules within the intelligent system of indoor service robots: the former aims to collect and organize environmental perceptual data, while the latter converts such data into task-oriented semantic knowledge and effective decision-making evidence.

From an application perspective, scenarios including home companionship, hospital delivery and hotel guidance require robots to not only identify object positions, but also comprehend the functional roles of objects in specific tasks, and adjust behaviors promptly according to user intentions and scene variations. However, current research still faces challenges such as incomplete environmental semantic information, insufficient connection between perception and planning, limited universality of knowledge representation, and inadequate real-time adaptation capabilities in open scenarios. In the future, it is necessary to strengthen multimodal perception fusion, ontology modeling, and open-world task reasoning to realize more stable and trustworthy intelligent service systems.

The main limitation of this paper lies in that all analyses are primarily based on published literature and typical application cases, without experimental validation on physical robot platforms or simulation environments. Therefore, the conclusions are more inclined towards theoretical induction and deduction. Subsequent research could introduce simulation experiments or scenario prototype designs where conditions permit, to enhance the empirical and operational nature of the research.

References

[1] Han, X., Li, S., Wang, X., & Zhou, W. (2021). Semantic mapping for mobile robots in indoor scenes: A survey. *Information, 12*(2), 92.

[2] Song, X., Liang, X., & Zhou, H. (2024). Semantic mapping techniques for indoor mobile robots: Review and prospect. *Measurement and Control*.

[3] Ni, J., Chen, Y., Tang, G., et al. (2023). Deep learning-based scene understanding for autonomous robots: A survey. *Intelligence & Robotics*.

[4] Alqobali, R., Alshmrani, M., Alnasser, R., et al. (2024). A survey on robot semantic navigation systems for indoor environments. *Applied Sciences, 14*(1), 89.

[5] Aguado, E., Gomez, V., Hernando, M., et al. (2024). A survey of ontology-enabled processes for dependable robot autonomy. *Frontiers in Robotics and AI, 11*, 1377897.

[6] Cui, G., Shuai, W., & Chen, X. (2021). Semantic task planning for service robots in open worlds. *Future Internet, 13*(2), 49.

[7] Cui, Y., Zhang, Y., Yang, S. X., et al. (2025). Task cognition and planning for service robots. *Intelligence & Robotics, 5*(1), 119–142.

[8] Kanso, H., Singh, A., & El Zarif, E., et al. (2025). Semantic SLAM: A comprehensive survey of methods and applications. *Intelligent Systems with Applications, 28*, 200591.