Hierarchical Understanding in Robotic Manipulation

Subjects: Automation & Control Systems

Contributor: Runqing Miao, Qingxuan Jia, Fuchun Sun, Gang Chen, Haiming Huang

In the quest for intelligent robots, it is essential to enable them to understand tasks beyond mere manipulation. Achieving this requires a robust parsing mode that can be used to understand human cognition and semantics.

Keywords: robotic manipulation ; knowledge reasoning

1. Introduction

Task understanding and skill refinement are crucial abilities for service robots. The decomposition of tasks in a manner similar to human cognition can be transformed into a planning problem within a symbolic space. Task and motion planning (TAMP) is the mainstream method for handling long-term tasks in robotic manipulation, relying on predefined planning domains, symbolic rules, and complex strategy searches. The limitation of such methods lies in the requirement for robots to possess a comprehensive model prior to task execution, thereby impeding their ability to achieve skill transfer and generalization, as well as adaptability to dynamically the evolving task scenes.

In order to address this issue, the knowledge-based approaches have been considered for the representation and task planning of robotic manipulation. Semantic knowledge serves as a medium for skill transfer among humans, providing concise explanations of the world. Knowledge bases effectively express and store the experiences generated in human or robotic manipulation, enabling reasoning and reuse. However, the discrete nature of knowledge poses challenges in directly describing the continuously manipulated data. The existing knowledge-based methods for robotic manipulation focus on characterizing static objects without achieving reasonable decoupling between the different factors. In querying and reasoning, they solely rely on rule-based symbolic computation. To properly represent the complex manipulation knowledge, it is essential to simultaneously consider both the continuous and discrete data as well as the static and dynamic factors. Additionally, robots need to acquire human knowledge and record the existing experiences in order to achieve real-time responses to new tasks and continuous updates.

2. Hierarchical Understanding in Robotic Manipulation

2.1. Knowledge Representation in Robotic Manipulation

Robotic manipulation involves three levels, ranging from high-level understanding and task planning to mid-level strategy and behavior planning and down to low-level execution. Knowledge, as a representation of the condensed data and information, primarily corresponds to the top level. The knowledge base provides the semantic context for the robots' input and output in their tasks, including defining the meaning or function of the manipulated objects. The early robot knowledge bases primarily focused on static objects, such as RoboEarth [1,2], KnowRob [3,4], RoboBrain [5], and the recently developed articulated object knowledge base AKB-48 [6]. However, all the aforementioned knowledge bases for robots are large-scale static repositories primarily focused on describing stationary objects in robotic manipulation tasks. The graph composition is excessively intricate and convoluted, leading to a heightened level of complexity when querying. Several knowledge representation methods have been proposed, specifically targeting the behaviors of robots. Action tree bank [7] generates a symbolic high-level representation in the form of a tree, encompassing the knowledge derived from demonstrations. FOON [8] is a structured method for representing the knowledge the models' objects and their movements in manipulation tasks, constructed through the manual annotation of instructional videos. Instead of a comprehensive symbolic representation system, several works apply knowledge to different robotics tasks, such as vision [9], grasping [10,11], assembly [12], and path planning [13,14], to describe robots' behavior processes. Furthermore, the utilization of knowledge graph embedding enables inference and finds application in the field of robotics [15]. However, its efficacy is constrained by the scale of the knowledge graph.

Currently, there are several key challenges facing robot knowledge graphs. Firstly, the discretization of continuous data results in a lack of proper decoupling between the high-level semantics and low-level data. Secondly, the dynamic

changes in relations overlook the impact of robot actions on object relations. Thirdly, traditional symbolic computation is the only consideration when querying knowledge graphs. Researchers propose a hierarchical architecture for knowledge-based framework, which achieves a layered decoupling of knowledge manipulation and the data, while also considering the dynamic factors in robotic manipulation. Moreover, the utilization of a state-of-the-art graph database as the foundation for knowledge graph enhances the query speed.

2.2. Knowledge Sources of Robotic Manipulation

The hierarchical organization of robotic manipulation skills determines its complexity. In the framework, manipulation knowledge is decoupled into two categories: static and dynamic. Static knowledge describes the stable common-sense knowledge obtained from resources, such as the internet, general knowledge databases, and vertical domain databases. Dynamic knowledge describes the continuously changing entity state process associated with actions generated based on existing experiences or real-time observations. Knowledge can be derived from semantically annotated descriptions by humans or different types of sensors, such as cameras, force sensors, and light-sensitive tactile sensors.

Manipulation knowledge originates from various sources due to its diverse types, primarily encompassing the following three categories: (1) Human-constructed manipulated datasets, which involve focused, single tasks, such as Push [16] and Bigs [17], and those encompassing multiple tasks like RoboTurk [18], MIME [19], and RoboNet [20]. (2) Common knowledge in the general knowledge base. In addition to domain-specific knowledge bases, there are publicly available cross-domain knowledge bases that serve as encyclopedias, such as the language knowledge base WordNet [21]; the concept knowledge base ConceptNet [22,23]; the world knowledge bases Freebase [24], Wikidata [25], DBpedia [26], and YAGO [27]; and others. These common knowledge bases encompass a vast amount of general information, including object definitions, classifications, and functionalities. (3) LLMs refers to transformer language models with trillions or more parameters, which are trained on extensive text data, such as GPT-3 [28], PaLM [29], and LLaMA [30]. These models exhibit exceptional proficiency in comprehending natural language and tackling intricate problems [31]. Because LLMs have extensive prior knowledge and reasoning capabilities, they facilitate knowledge extraction and update. LLMs have already found applications in robotic manipulation tasks. Text2Motion [32] accomplishes the end-to-end planning of robotic manipulation tasks using LLMs. Wu et al., on the other hand, integrated language-based planning and perception with the limited summarization capability of LLMs to facilitate robots in understanding the users' preferences [33].

2.3. Knowledge Reasoning with Representation Learning

Knowledge reasoning involves deriving new knowledge or conclusions from the existing knowledge using various methods. Knowledge reasoning methods via representation learning are primarily implemented through translating models and graph embedding. Translating models is based on word2vec [34]. TransE [35] utilizes the concept of translating invariance within a word vector space, considering the relations in knowledge bases as translating vectors between entities. Graph networks are primarily used for tasks in non-Euclidean spaces, which align well with the topological graph structure of knowledge graphs. Graph embedding, also known as graph representation learning, expresses the nodes in a graph as low-dimensional dense vectors. It necessitates that the nodes with similar characteristics in the original graph are also close to each other in the low-dimensional representation space. The output expression vector can be used for downstream tasks, such as entity alignment [36] and knowledge fusion [37]. The most classic graph embedding method is DeepWalk [38], which utilizes random walks to sample nodes within the graph and acquire co-occurrence relations among them. Furthermore, there are other methods such as node2vec [39] and LINE [40]. Researchers assess the advantages and disadvantages of the above methods, and combined with the hierarchical structure of the framework, researchers design a suitable embedding method to accomplish knowledge reasoning.