OntoScene: Ontology Guided Indoor Scene Understanding for Cognitive Robotic Tasks

Snehasis Banerjee *, Pradip Pramanick, Chayan Sarkar, Balamuralidhar P

TCS Research, Tata Consultancy Services, India

Abstract. We have developed a robotics ontology, OntoScene, that extends IEEE CORA [5] and SemNav [1] ontologies. Contrary to the prior work that lacked usage of ontology in scene understanding, the proposed system uses OntoScene to figure out objects and their relations in a scene and create a scene graph for aid in various cognitive robotic tasks where object localization, scene graph generation is important. This work positions semantic web technology as a key enabler in robotic tasks.

Keywords: Ontology, Cognitive Robotics, Semantic Scene Processing

1 Background

Scene understanding and scene graph generation are critical components in most robotic tasks. It is difficult for a robot to semantically understand the world (sequence of scenes) based only on sensor inputs, typically an RGB-D camera. To this extent, approaches presented in [1] and [2] are enhanced to take the help of querying the OWL ontology to handle various tasks like navigation, description based object localization, visual question answering, etc. In contrast to the end-to-end learning approaches, symbolic approaches are more reliable, interpretable, and safe from the perspective of task actuation in human co-occupied spaces.

2 OntoScene Ontology

OntoScene was created using 'Competency Questions' specific to concepts relevant to cognitive robotic tasks in indoor environments. The initial ontology was built on top of CORA, which is an abstract robotics ontology that cannot be used without extension to more granular domain knowledge of types of robot tasks and the indoor world. SemNav was extended by adding granular relations and restrictions some of which are enlisted below:
(a) Relative object positions 'left-of', 'right-of', 'top-of', 'below' and reverse.
(b) Surface location - object located at the center of a table or near the edge.
(c) Mobility - if an object can be moved, e.g., a 'cup' has a multi-zone affinity.
(d) Algorithm linking - the ontology also stores call links to apt algorithms to detect a relation. For example, if object 'cup' is detected in a scene, then the color detection algorithm is invoked if 'cup' has property 'hasColor' in the ontology.

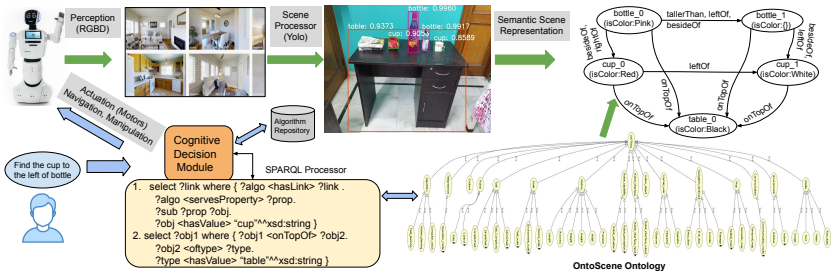

Fig. 1. System using OntoScene for scene understanding to aid in robotic tasks

3 Scene Understanding and Scene Graph Generation

As shown in Fig. 1, suppose a user issues an instruction of finding an object (cup) with specific criteria (to the left of the bottle). The input to the Cognitive Module is perception from the robot's ego view (RGB-D camera image scene sequences), and output is actuation (navigation or manipulation). We use Yolo [6] and DenseCap [4] as the object detection algorithms to locate the objects within the current scene. When a 'cup' is detected, the Cognitive Module queries the ontology to get a list of properties corresponding to that object entity. The corresponding algorithms are selected from the Algorithm Repository to extract the relevant attributes and relations from the scene. Finally, a semantic scene graph is generated to aid in complex tasks needing semantic object localization.

The standard way to summarize scenes is using deep image captioning [3] and object detection. While machine learning models may infer the presence of a pair of objects, it is difficult to ground an accurate relationship due to the large space of possible relationships. For example, it is difficult to disambiguate between 'table-beside-cup' and 'cup-on-table'. Also, such models can not predict transitive relationships directly. Thereby utilizing the ontology, commonsense disambiguation becomes possible, by querying whether such an (un)directed edge predicate exists or not between entities (objects connected by a property).

References

1. Banerjee, S., Purushothaman, B.: Semnav: How rich semantic knowledge can guide robot navigation in indoor spaces. In: ISWC (Industry). pp. 398–400 (2020)
2. Bruno, B., et. al.: Knowledge representation for culturally competent personal robots. International Journal of Social Robotics, Springer pp. 1–24. (2019)
3. Hossain, M.Z., et. al.: A comprehensive survey of deep learning for image captioning. ACM Computing Surveys **51**(6), 1–36 (2019)
4. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: IEEE CVPR, pages=4565–4574, year=2016
5. Prestes, E., et. al.: Core ontology for robotics and automation. In: Workshop on Knowledge Representation and Ontologies for Robotics and Automation. p. 7 (2014)
6. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)