Fully Automated Task Management for Generation, Execution, and Evaluation: A Framework for Fetch-and-Carry Tasks with Natural Language Instructions in Continuous Space

Motonari Kambara and Komei Sugiura
Keio University
{motonari.k714, komei.sugiura}@keio.jp

Abstract

This paper aims to develop a framework that enables a robot to execute tasks based on visual information, in response to natural language instructions for Fetch-and-Carry with Object Grounding (FCOG) tasks. Although there have been many frameworks, they usually rely on manually given instruction sentences. Therefore, evaluations have only been conducted with fixed tasks. Furthermore, many multimodal language understanding models for the benchmarks only consider discrete actions. To address the limitations, we propose a framework for the full automation of the generation, execution, and evaluation of FCOG tasks. In addition, we introduce an approach to solving the FCOG tasks by dividing them into four distinct subtasks.

1. Introduction

In our aging society, the need for daily care and support is increasing. As a result, the shortage of home care workers has become a social problem, and domestic service robots that can physically assist people with disabilities are attracting attention. However, the current ability of domestic support robots to understand instructions in natural language and execute daily tasks appropriately is insufficient.

The aim of this paper is to develop a framework for a robot to execute instructions based on visual information when given natural language instructions for fetch–and–carry tasks. Many frameworks have been proposed for the execution of everyday tasks by robots based on natural language instructions [1,8–11]. For example, ALFRED [11] is a large-scale benchmark containing 25K instructions. However, it is difficult to turn it into an on-the-fly simulation because the instructions are manually annotated. Therefore, it is also difficult to evaluate with diverse tasks, and evaluations have only been conducted with fixed tasks.

Furthermore, many methods are also proposed in the benchmarks. In the ALFRED benchmark [11], many multimodal language understanding models such as FILM [5] and HLSM-MAT [3] have been proposed. Many of these multimodal language understanding models consider only discrete actions. Therefore, it is challenging to apply them to the real world, where continuous actions are required. On the other hand, the RoboCup@Home competition [2] is a representative framework for daily tasks by domestic service robots in the real world. In this framework, tasks are not generated automatically.

In this paper, we propose a framework for fully automating fetch-and-carry tasks in a simulation environment. This framework allows for the generation, execution, and evaluation of tasks on–the–fly. In the proposed framework, the simulation environment is a continuous space.

2. Problem Statement

In this paper, we address the FCOG task. The task involves providing a robot with natural language instructions related to fetch–and–carry tasks, which the robot then executes utilizing visual information.

The inputs of the task are defined as natural language instruction and multiple images taken by a robot's camera. Given the input, the robot should grasp a target object and place it to a destination in accordance with the instruction.

3. Method

The proposed framework is closely related to tasks like the General Purpose Service Robot task in RoboCup@Home competition [2]. The framework fully automates the simulation of FCOG tasks using three systems: a task generation system, a task evaluation system, and a task execution system.

The Task Generator generates tasks through three main steps. First, the mechanism creates a simulated environment by placing static game objects. Next, it randomly places dynamic game objects in the environment and also positions the robot in a specified initial location for each environment. Finally, the task generation mechanism automatically generates tasks. In this step, the mechanism first randomly selects the objects to be used as target objects and destinations from the simulation environment. The task generation mechanism then uses Unity functionality to obtain the coordinates of these objects and takes pictures of the objects us-

Table 1. Quantitative comparison for FCOG tasks. The best scores are in bold.

Methods	Navigation Success rate[%]	OLR Accuracy[%]	Fetching Success rate[%]	Carrying Success rate[%]
Baseline (WRS-VS winning method)	**100** (40/40)	0 (0/0)	0 (0/0)	0 (0/0)
Ours	**100** (40/40)	**20** (8/40)	**100** (8/8)	**12.5** (1/8)

ing virtual cameras. The Task Generator generates instruction sentences using the CRT [4]. The model generates object manipulation instruction sentences based on the target region, destination region, and context regions. The model generates an instruction sentence like "Move the wooden toy car in front of the white bottle onto the table." The Task Generator obtains the name of the room where the target object is located from the simulator, and uses it to generate an instruction sentence like "Go to the bedroom." Finally, the Task Generator combines these instruction sentences.

In this study, we split the FCOG task into four subtasks: Navigation, Object Location Retrieval (OLR), Fetching, and Carrying. The OLR can be roughly split into three steps: Crawling & Image Collection, Object Detection and Multimodal Language Comprehension.

After identifying the target region and the destination region, the robot executes fetch–and–carry tasks. The Fetching task is carried out in the following specific steps. First, the robot moves to the point where it took the image. Then, it grasps the object using heuristic methods. Success in the Fetching task is defined as the ability to grasp the appropriate object. If the Fetching task is successful, the robot performs the Carrying task in similar steps.

The Task Evaluator determines that the session has ended when it reaches the following states:
- A certain amount of time has elapsed during the execution of the FCOG task.
- The robot successfully completes the FCOG task.
- The robot fails in executing any of the subtasks.

After a session termination decision is made, the Task Generator initializes the environment and starts the next session.

4. Experimental Results

We used an extended version of the standard simulator used in the World Robot Summit 2018 Partner Robot Challenge/Virtual Space Competition (WRS-VS, [7]). Table 1 shows the quantitative results for each subtask. The table displays the number of executions and successful attempts for each task. We used the method [6] that won WRS–VS as the baseline method. In this method, trajectory generation for a robotic arm, path planning, and navigation methods were the same as those of the proposed method.

In the OLR task, the baseline method had an accuracy of 0%. This is likely because it could not understand the diverse reference expressions contained in the instruction sentences generated by the Task Generator. On the other hand, the proposed method had an accuracy of 20%. The baseline method did not execute the Fetching and Carrying tasks,
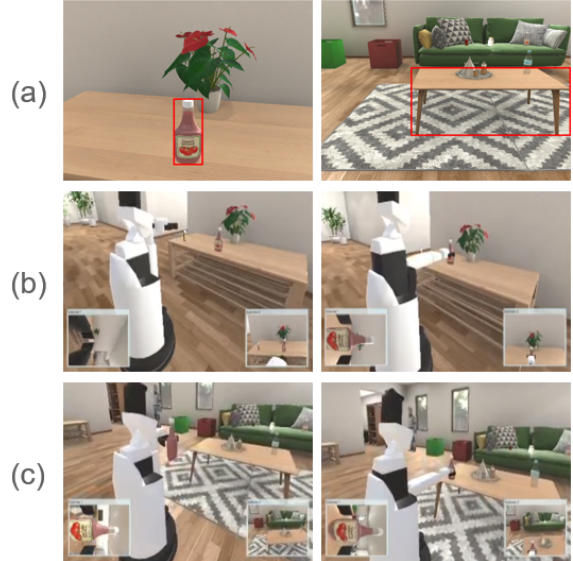

Figure 1. A successful session. In the figure, (a) shows the images of the target object and the destination acquired by the task generation system.

as these tasks were only performed upon successful completion of the OLR task. In contrast, the proposed method achieved success rates of 20% and 12.5% for the Fetching and Carrying tasks, respectively.

Fig. 1 shows the qualitative results for the FCOG tasks. In this case, as shown in Fig. 1(a), the Task Generator selected a red bottle and a table in front of a sofa as the target object and the destination. The instruction generated by the Task Generator was "Go to the living room, move a plastic bottle from the shelf to the table." In the Navigation task, the robot successfully navigated to the living room. In the OLR task, a red bottle and a table in front of the desk were correctly identified as the target object and the destination, respectively. The robot then successfully grasped the red bottle in the Fetching task, as shown in Fig. 1(b). Finally, as shown in Fig. 1(c), the robot placed the red bottle on the table in the Carrying task.

5. Conclusions

In this paper, we focused on building a framework for FCOG tasks. The main contributions are as follows:
- We proposed a framework for fully automating the generation, execution, and evaluation of FCOG tasks.
- We proposed an approach to solving the FCOG tasks by dividing them into four subtasks.

References

[1] Ankit Goyal, Arsalan Mousavian, Chris Paxton, et al. IFOR: Iterative Flow Minimization for Robotic Object Rearrangement. In *CVPR*, pages 14787–14797, 2022. 1

[2] Luca Iocchi, Dirk Holz, et al. RoboCup@Home: Analysis and results of evolving competitions for domestic and service robots. *Artificial Intelligence*, 229:258–281, 2015. 1

[3] Shintaro Ishikawa and Komei Sugiura. Moment-based Adversarial Training for Embodied Language Comprehension. In *ICPR*, pages 4139–4145, 2022. 1

[4] Motonari Kambara et al. Case Relation Transformer: A Crossmodal Language Generation Model for Fetching Instructions. *IEEE RA-L*, 6(4):8371–8378, 2021. 2

[5] So Min, Devendra Chaplot, Pradeep Ravikumar, Yonatan Bisk, et al. FILM: Following Instructions in Language with Modular Methods. In *ICLR*, 2022. 1

[6] Yoshiaki Mizuchi and Tetsunari Inamura. Optimization of Criterion for Objective Evaluation of HRI Performance that Approximates Subjective Evaluation: A Case Study in Robot Competition. *AR*, 34(3-4):142–156, 2020. 2

[7] Hiroyuki Okada, Tetsunari Inamura, and Kazuyoshi Wada. What competitions were conducted in the service categories of the World Robot Summit? *AR*, 33(17):900–910, 2019. 2

[8] Aishwarya Padmakumar, Jesse Thomason, Ayush hrivastava, Patrick Lange, et al. TEACh: Task-driven Embodied Agents that Chat. In *AAAI*, pages 2017–2025, 2022. 1

[9] Yuankai Qi, Qi Wu, Peter Anderson, et al. REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. In *CVPR*, pages 9982–9991, 2020. 1

[10] Bokui Shen, Fei Xia, Chengshu Li, et al. iGibson 1.0: A Simulation Environment for Interactive Tasks in Large Realistic Scenes. In *IROS*, pages 7520–7527, 2021. 1

[11] Mohit Shridhar, Jesse Thomason, Daniel Gordon, et al. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *CVPR*, pages 10740–10749, 2020. 1