A Rapidly Adjustable Object Recognition System through Language Based Prompt Engineering

Naoki Yamaguchi
Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu, Kitakyushu, 808-0196, Japan

Tomoya Shiba
Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu, Kitakyushu, 808-0196, Japan

Kosei Isomoto
Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu, Kitakyushu, 808-0196, Japan

Hakaru Tamukoh
Research Center for Neuromorphic AI Hardware, Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu, Kitakyushu, 808-0196, Japan
E-mail: yamaguchi.naoki892@mail.kyutech.jp
https://www.lsse.kyutech.ac.jp/

Abstract

We propose the use of language-based prompt engineering to achieve rapidly adjustable object recognition in RoboCup@Home. The proposed prompt engineering involves humans adding features, such as the color and material of an object, into the text prompts inputted into Language Segment Anything. In this research, we evaluated the effectiveness of our proposed method in three benchmark tests for object recognition at RoboCup@Home 2023. The results show that the highest scores were obtained in specific tasks, indicating that the proposed method applies to various recognition tasks.

Keywords: Home service robot, Object recognition, Prompt engineering, RoboCup@Home

1. Introduction

In recent years, service robots have attracted increasing attention against the aging population background [1]. This trend has spurred active research in this domain [2], [3]. These robots operate in dynamic environments, frequently encountering new and unfamiliar objects, making the learning process of recognition technologies and operational efficiency crucial. Ono et al.'s research [4] automates dataset generation and annotation, reducing the time and cost involved in the learning process while maintaining real-time performance using You Only Look Once v4 (YOLOv4) [5]. However, their experiments [6] showed that YOLO learning took two hours to prepare 500,000 training images and 18 hours. We aim to achieve a faster learning process and improve operational efficiency.

This study proposes Language Segment Anything (Lang SAM) [7] as a language-based object recognition system, offering quick adjustability. To validate its effectiveness, we conducted experiments at RoboCup@Home [8], an international competition focused on developing practical home service robots. Additionally, we evaluated operational efficiency by measuring inference time and power consumption, comparing it with the traditional method.

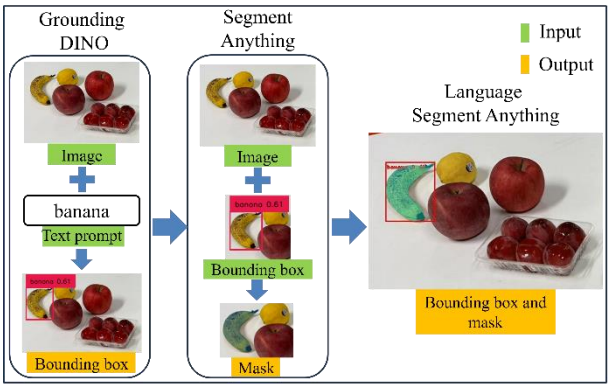

Fig. 1. Process flow of object recognition with language

2. Related Work

2.1. *Object recognition*

In object recognition, annotating datasets is an essential process [9]. Annotation involves labeling each object in an image with its name, location, shape, and other characteristics. Typically, manual processes accomplish this task. The performance of object recognition heavily depends on the quality and quantity of the dataset, as well as the precision of the annotations. Manually creating large volumes of high-quality annotations can incur substantial costs. Moreover, maintaining the accuracy of manual annotations is challenging.

2.2. *Dataset generation*

To address the issue of manual annotations in object recognition datasets, research on Sim2Real, which involves generating datasets using simulators, is progressing [10], [11]. Utilizing simulators eliminates human error and enables the rapid and high-quality generation of datasets. Additionally, this approach ensures consistency in annotations.

2.3. *Problems*

In addition to the challenges of dataset annotation, a significant issue in object recognition is the need to train recognition models with these datasets. Recognizing new objects requires generating datasets for these objects and conducting training, which can be very costly. This process involves substantial time and resources, mainly when introducing new objects to the system, thereby increasing the overall expense and complexity of developing effective object recognition models.

3. Proposed Method

3.1. *Object recognition with language*

For object recognition using language, our proposed method employs Lang SAM. Lang SAM combines Grounding DINO [12] and Segment Anything [13] into a single model, enabling language-based object recognition and segmentation. Fig. 1 shows a process flow of object recognition with language.

3.2. *Prompt engineering*

Adjust the text prompts that humans enter Lang SAM. Lang SAM's recognition accuracy depends on the text prompt. Below, we outline the procedure for adjusting text prompts:

(i) Take pictures of multiple scenes containing the object of interest.
(ii) The initial text prompt shall be the name of the recognition target.
(iii) Input both the images and text prompts into the recognition system to check if the object is correctly recognized and assess the accuracy.
(iv) Improve the text prompts by human addition of details like color and material, aiming to recognize objects not initially identified or to increase accuracy.

4. Experimental Condition

First, we compared the preparation costs for recognizing new objects using the recognition models YOLOv8 [14] and Lang SAM. We generated the dataset for training YOLOv8 using conventional methods. Next, we implemented the proposed method on the Human Support Robot developed by TOYOTA MOTOR CORPORATION [15] to reveal its impact on enhancing object recognition capabilities. The effectiveness of this implementation was validated based on the scores obtained at RoboCup@Home. Finally, we compared the operational efficiency of YOLOv8 and Lang SAM, focusing on their inference time and power consumption, to determine which method was more efficient in real-world scenarios.

Table 1. Specifications of the Computational System Used in Experiments

CPU	11th Gen Intel® Core™ i7-11700 @ 2.50GHz × 16
GPU	NVIDIA GeForce RTX 3080
Memory	32GB

Table 2. Time to prepare recognition model

Method	Dataset [images]	Prepare time [hours]
Dataset generator + YOLOv8	500,000	18
Proposed method	10	2

4.1. *Time to prepare recognition model*

We created a dataset to train YOLOv8 by generating 500,000 images using the method developed by the Ono et al. system. Following this, we conducted manual prompt engineering on these images. We then trained this dataset using YOLOv8. Next, images from 10 scenes were captured in real environments to serve as inputs for Lang SAM's prompt engineering. We then manually conducted prompt engineering on these images. The time taken to prepare these recognition models was measured.

4.2. *Evaluate the effectiveness of the proposed method*

The conditions of the second task performed at the RoboCup@Home 2023 are summarized:

(1) Storing Groceries (SG): The task involved organizing and shelving items scattered on a table, including categorizing objects and handling unknown items.
(2) Stickler for the Rules (SR): In this task, the robot acts as a party host, identifying guests not following house rules and enforcing compliance by explaining the rules to them. The rules are as follows:
    a. Remove shoes inside the house.
    b. Do not enter prohibited areas.
    c. Do not throw garbage on the floor.
    d. Always hold a drink in hand.

    This task primarily requires object recognition and human interaction technologies.

4.3. *Inference time and power consumption*

In section 4.1, we measured the inference time and power consumption for each of the prepared recognition models. The inference time is the time recorded as the output of each recognition system, and this experiment uses a maximum wattage recorded by the NVIDIA System Management Interface [16] as a power consumption. Table 1 shows the specifications of the PC used in the experiments.

Table 3. Results of RoboCup@Home2023 (proposed method: HMA)

		1st	2nd	3rd
SG	Team	**HMA**	Tech United Eindhoven	eR@sers +Pumas
	Score	**420**	410	220
SR	Team	**HMA**	TRAIL	Tech United Eindhoven
	Score	**1000**	1000	800

Table 4. Inference time and power consumption.

	Inference time [msec]	Power consumption [W]
YOLOv8	71.8	95.7
Lang SAM	577.7	233.0

5. Experimental Result

5.1. *Time to prepare recognition model*

Table 2 shows the time taken for dataset generation, training with YOLOv8, and the preparation of the proposed method. YOLOv8 required 8 hours to create the training dataset and 10 hours for training, for a total of 18 hours. In contrast, the proposed method required approximately 2 hours for prompt tuning with 10 scene images and text prompts.

5.2. *The score of RoboCup@Home 2023*

Table 3 shows the top three teams and their scores in SG and SR categories at RoboCup@Home 2023. The results show that HMA achieved the highest SG and SR scores.

5.3. *Inference time and power consumption*

Table 4 shows the inference time and power consumption between the traditional YOLOv8 method and the proposed method. The results indicated that the proposed method had approximately 8 times later inference time and about 2.4 times higher power consumption than YOLOv8.

6. Conclusion and Discussion

In this study, we proposed using language-based prompt engineering to achieve rapidly adjustable object recognition in RoboCup@Home. We validated the effectiveness of this approach through participation in RoboCup@Home2023. We can adjust the proposed method in just 2 hours, which is only one-ninth of the time required by conventional methods. Furthermore, The results from RoboCup@Home 2023 support the feasibility and effectiveness of our approach in various tasks in terms of scoring. However, compared to traditional methods, it became clear that there are issues with operational efficiency, such as inference time and power consumption. We consider the need to select recognition models appropriately, considering the preparation cost and operational efficiency.

References

1. "Fuji Keizai Group," https://www.fuji-keizai.co.jp/report/detail.html?code=162208813, (Accessed 12/13/2023).
2. T. Shiba, T. Ono and H. Tamukoh, Object Search and Empty Space Detection System for Home Service Robot, Proceedings of the 2023 International Conference on Artificial Life and Robotics (ICAROB2023), 2023.
3. K. Isomoto, Y. Yano, Y. Tanaka, and H. Tamukoh, Robust trash can lid opening system, Proceedings of the 2023 International Workshop on Smart Info-Media Systems in Asia (SISA), 2023.
4. T. Ono, D. Kanaoka, T. Shiba, S. Tokuno, Y. Yano, A. Mizutani, I. Matsumoto, H. Amano and H. Tamukoh, Solution of World Robot Challenge 2020 Partner Robot Challenge (Real Space), Advanced Robotics, 2022, 36, pp.870-889
5. A. Bochkovskiy, C. Wang, H. M. Liao, YOLOv4: optimal speed and accuracy of object detection. 2020, arXiv:2004.10934.
6. T. Shiba, A. Mizutani, Y. Yano, T. Ono, S. Tokuno, D. Kanaoka, Y. Fukuda, H. Amano, M. Koresawa, Y. Sakai, R. Takemoto, K. Tamai, K. Nakahara, H. Hayashi, S. Fujimatsu, Y. Mizoguchi, M. Anraku, M. Suzuka, L. Shen, K. Maeda, F. Matsuzaki, I. Matsumoto, K. Murai, K. Isomoto, K. Minje, Y. Tanaka, T. Morie, and H. Tamukoh, Hibikino-Musashi@Home 2023 Team Description Paper, 2023, arXiv:2310.12650.
7. L. Medeiros, Lang Segment Anything, https://lightning.ai/pages/community/lang-segmentanything-object-detection-and-segmentation-with-text-prompt (Accessed 12/13/2023).
8. RoboCup@Home, https://athome.robocup.org/, (Accessed 12/13/2023).
9. Shu Yang, JingWang, Sheeraz Arif, Minli Jia, Shunan Zhong, SAL-Net: Self-Supervised Attribute Learning for Object Recognition and Segmentation, Wireless Communications and Mobile Computing, 2021.
10. M. Denninger, M. Sundermeyer, D. Winkelbauer, D. Olefir, T. Hodan, Y. Zidan, M. Elbadrawy, M. Knauer, H. T. Katam and A. Lodhi, BlenderProc: Reducing the Reality Gap with Photorealistic Rendering, Proceedings of International Conference on Robotics: Science and Systems (RSS), 2020.
11. S. Max and B. Sven, Stillleben: Realistic Scene Synthesis for Deep Learning in Robotics, Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2020.
12. S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu and L. Zhang, Grounding DINO: Marrying DINO with Grounded PreTraining for Open-Set Object Detection, 2023, arXiv:2303.05499
13. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P, Dollár and R. Girshick, Segment Anything, 2023, arXiv:2304.02643
14. Ultralytics, Ultralytics YOLOv8 Docs, https://docs.ultralytics.com/ (Accessed on 12/13/2023).
15. T. Yamamoto, K. Terada, A. Ochiai, F. Saito, Y. Asahara and K. Murase. Development of human support robot as the research platform of a domestic mobile manipulator, ROBOMECH journal, Vol. 6(1), 2019, pp. 1-15.
16. NVIDIA.DEVELOPER, System Management Interface SMI, https://developer.nvidia.com/nvidia-system-management-interface (Accessed 12/13/2023).

Authors Introduction

Mr. Naoki Yamaguchi

He received the B.Eng. degree from the National Institute of Technology, Ube College, Japan, in 2023. He is a Master's degree student at the Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology. His research interests include dataset creation and Visualization.

Mr. Tomoya Shiba

He received the B.Eng. degree from National Institute of Technology, Kagoshima College, Japan, in 2021. He received the M.Eng. from Kyushu Institute of Technology, Japan, in 2023. He is currently in a Ph.D. student in the graduate school of Life Science and Systems Engineering, Kyushu Institute of Technology. His research interest includes image processing, motion planning, and domestic service robots.

Mr. Kosei Isomoto

He received the B.Eng. degree from Kyushu Institute of Technology, Japan, in 2023, respectively. He is a Master's degree student at the Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology. His research interests include brain-inspired artificial intelligence models and home service robots.

Prof. Hakaru Tamukoh

He received the B.Eng. degree from Miyazaki University, Japan, in 2001. He received the M.Eng and the Ph.D. degree from Kyushu Institute of Technology, Japan, in 2003 and 2006, respectively. He was a postdoctoral research fellow of 21st century center of excellent program at Kyushu Institute of Technology, from April 2006 to September 2007. He was an assistant professor of Tokyo University of Agriculture and Technology, from October 2007 to January 2013. He had been an associate professor from February 2013 to March 2021 and is currrectly a professor in the graduate school of Life Science and System Engineering, Kyushu Institute of Technology, Japan. His research interest includes hardware/software complex system, digital hardware design, neural networks, soft-computing and home service robots. He is a member of IEICE, SOFT, JNNS, IEEE, JSAI and RSJ.