Efficient Object Detection with Color-Based Point Prompts for Densely Packed Scenarios in WRS FCSC 2024

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

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

This paper introduces a novel object detection method designed for densely packed environments, such as those encountered in the World Robot Summit Future Convenience Store Contest (FCSC) 2024. Our system leverages color-based point prompts in conjunction with Segment Anything (SAM) 2 to achieve precise object segmentation and grasp point estimation, specifically targeting scenarios where objects like rice ball cluster tightly within containers. Unlike traditional methods that depend on pre-defined grasping strategies susceptible to positional drift in mobile robots, our approach dynamically identifies and isolates objects without requiring extensive retraining inherent to CNN or Transformer models. We conducted comprehensive experiments comparing our method against SAM 2 and Grounding DINO using a dataset of 10 test images containing 202 rice balls. Additionally, we deployed the system on a Toyota Human Support Robot during the FCSC Stock Task to assess real-world performance metrics. Results demonstrate that our method achieves higher detection accuracy and operational efficiency, validating its potential for autonomous retail applications.

Keywords: Object Detection, Human Support Robot, Future Convenience Store Contest

1. Introduction

Automating retail environments, especially convenience stores, demands advanced robotic systems that accurately detect, identify, and manipulate a diverse range of products in compact and cluttered settings [1]. The World Robot Summit (WRS) Future Convenience Store Contest (FCSC) exemplifies this need by presenting tasks that mimic real-world retail scenarios. One such task is the Stock Task, which involves handling densely packed rice balls within containers [2]. A major challenge in these tasks is grasping tightly clustered and visually similar objects. Traditional robotic grasping strategies often rely on fixed grasp points, which become unreliable for mobile robots due to inevitable discrepancies between planned and actual positions. This positional drift reduces the effectiveness of static grasping approaches, highlighting the need for more dynamic and adaptable object detection and segmentation methods.

Convolutional Neural Networks (CNNs) and Transformer-based architecture offer high accuracy but demand extensive training datasets and computational resources to adapt to new objects. This retraining process increases preparation costs and limits system flexibility in dynamic environments where product assortments frequently change [3].

Prompt-based segmentation models, such as the Segment Anything Model (SAM) [4] and its successor SAM2 [5], present promising alternatives. These models use prompts—points, boxes, or language instructions—to guide segmentation without exhaustive retraining. Building on this paradigm, we propose a color-based point prompt methodology integrated with SAM2 to achieve efficient and robust object detection in densely packed scenarios. By leveraging color cues, our system dynamically generates prompts that facilitate the accurate isolation and identification of visually similar objects, enhancing the robot's ability to perform reliable grasping tasks.

This paper details the design and implementation of our proposed method, evaluates its performance through controlled experiments and real-world deployments, and discusses its implications for future autonomous retail systems.

2. Related Work

2.1. *Object Detection in Dense Environments*

Detecting objects in densely packed environments presents significant challenges due to occlusions, similar appearances, and spatial constraints. Traditional object detection frameworks, including CNN-based models like YOLO [6] and Transformer-based architectures, have achieved substantial progress in accuracy and speed. However, these models typically require extensive training on large, object-specific datasets, limiting their adaptability in dynamic retail settings where product assortments frequently change.

2.2. *Prompt-Based Segmentation and Grounding DINO*

Prompt-based segmentation methods have gained traction for their ability to guide segmentation models with minimal input prompts. The Segment Anything (SAM) and its advanced variant SAM2 exemplify this approach by enabling the segmentation of objects based on various prompt types, including points, bounding boxes, and textual descriptions. These models leverage extensive pre-training on diverse datasets, allowing them to generalize across various objects and environments without retraining. While language-based prompts, as utilized in Grounding DINO [7], offer flexible object specification, they may need more precision in environments with visually similar objects. Our approach focuses on color-based point prompts to enhance segmentation accuracy in challenging scenarios.

2.3. *Grasping Strategies for Mobile Robots*

Robust perception systems enable effective grasping in dynamic environments by accurately identifying and localizing objects for manipulation. Fixed grasping strategies are insufficient for mobile robots due to potential positional discrepancies and variations in object arrangements. Recent advancements emphasize integrating real-time object detection with adaptive grasp point estimation to ensure reliable manipulation in cluttered environments.

3. Proposed Method

Our method targets the segmentation of tightly packed objects, such as rice balls, using a combination of human-guided color range extraction and automated processing steps integrated with SAM2. Fig. 1 shows a process flow of proposed method. The workflow is as follows:

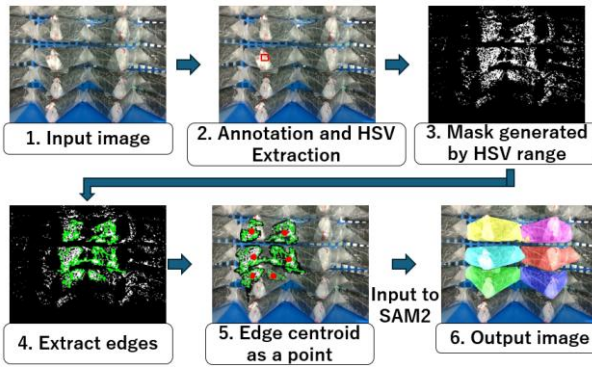


Fig. 1 Process flow of proposed method

3.1. *Overview*

The proposed method captures real-world images, extracts HSV color ranges from annotated regions, and generates precise point prompts based on edges detected from binary masks. These prompts guide SAM2 for accurate segmentation. This process minimizes the need for extensive training or pre-existing object models, making it adaptable for dynamic retail environments like the WRS FCSC Stock Task.

3.2. *Detailed Steps*

i. Capturing Input Images
Capture images of the target environment using a standard RGB camera.

ii. Human annotation and HSV Extraction
A human operator manually annotates a portion of the target object by drawing bounding boxes around regions of interest. Extract the HSV color range characteristic of the object from these annotated areas.

iii. Mask Creation
Generate a binary mask using the extracted HSV ranges to isolate regions matching the target object's color.

iv. Edge Detection and Filtering
Detect edges from the binary mask and filter out small regions below a predefined area threshold to reduce noise.

v. Point Prompt Generation
Calculate the centroids of valid edge regions and use them as point prompts to indicate object locations.

vi. Segmentation Using SAM2
Feed the input image and generated point prompts into SAM2 to produce detailed segmentation masks for each detected object.

Table. 1 Result of the Stock sub task

Team	Happy Robot	TAK	TMU Mecha	NaRIPa	HAR Chuo	Team Meijo	**HMA (Our Team)**	eR@sers	Optimator	HSRL -CoR
Score	59	59	55	40	16.5	14	7	5	5	0

Table. 2 Comparison of Zero-Shot Object

Image ID	Proposed Method	Grounding DINO	Ground Truth
1	2	2	3
2	5	3	5
3	3	3	5
4	4	5	15
5	6	3	15
6	6	3	15
7	6	5	16
8	8	6	20
9	46	1	54
10	47	1	54
ALL	133	32	202

4. Experiments

This section outlines the experimental setups used to evaluate the proposed method. We conducted two primary experiments: implementing the WRS FCSC method and assessing its zero-shot object detection performance against Grounding DINO.

4.1. *Evaluation in the WRS FCSC Stock Task*

To test the method's real-world applicability, we integrated it into the Human Support Robot developed by TOYOTA MOTOR CORPORATION [8] and participated in the WRS FCSC Stock Sub Task.

Stock Task Overview:
The WRS FCSC aims to advance technologies for automating convenience store operations, including product shelving and waste collection. The Stock Sub Task measures how quickly and accurately participants arrange pre-packaged items on shelves.

- **Items:** Arrange 54 rice balls (three types, 18 each) into designated shelf positions (e.g.,Fig. 2).
- **Scoring:** Correctly placing each rice ball earns *1* point. Using a standard container for transportation awards an additional *5* bonus points.

(a) Grounding DINO (b) Proposed method

Fig. 2 Detection Result

4.2. *Zero-Shot Object Detection*

We evaluated the proposed method's zero-shot object detection capabilities using a specialized dataset and compared its performance with Grounding DINO.

Experimental Setup:
- **Dataset:** Created a dataset of 10 test images featuring 202 densely packed rice balls to simulate real-world retail environments.
- **Models Compared:**
  - **Proposed Method:** Uses color-based point prompts with SAM2 for segmentation.
  - **Grounding DINO:** Employs textual prompts (e.g., "a rice ball") for object detection.
- **Evaluation Metrics:** Measured detection accuracy by the number of correctly identified rice balls compared to the ground truth.

5. Experimental Results

5.1. *WRS FCSC Stock Task*

We evaluated the proposed method using the HSR in the WRS FCSC Stock Sub Task. Our team, HMA, scored *7* points, as illustrated in Table. 1 Specifically, we earned *5* points by utilizing a standard container for item transportation and *2* points by accurately placing items.

Challenges Identified:
- Slow Operation Speed: A single cycle of picking, moving, and placing a rice ball took approximately one minute, limiting the maximum score achievable to *15* points.
- Shelf Height Issue: The shelf height was lower than the HSR's usual operational range, leading to frequent grasping failures due to an unconventional gripping method.

5.2. *Zero-Shot Object Detection*

Table. 2 shows that the proposed method detected 133 out of 202 rice balls, achieving an accuracy of 65.8% relative to visible rice balls. In contrast, Grounding DINO identified only 32 rice balls, corresponding to a 15.8% detection rate. Fig. 2 shows the detection results for Image ID 10. As shown in Figure 2 (a), the detection results of Grounding DINO reveal that language-based object recognition tends to identify the entire group of rice balls as a single entity rather than detecting each individual rice ball separately.

6. Conclusion

In this study, we introduce a high-performance object detection method for densely packed objects using color-based point prompts within the WRS FCSC. We validated our approach by participating in the WRS FCSC. In our preliminary experiments, language-based object recognition achieved a detection rate of 15.8%. In contrast, our proposed method reached a detection rate of 65.8%, outperforming the language-based approach by approximately 4.2 times in detecting densely packed rice ball. Our team placed seventh in this competition. The proposed approach efficiently identifies numerous objects without requiring the additional training typically necessary for CNN or Transformer-based detectors. Implementing and testing the method on the HSR confirmed its practical potential, highlighting its suitability for real-world autonomous retail applications.

Acknowledgements

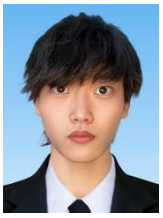
This research is based on results from a JPNP16007 project commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This research received support from JSPS KAKENHI Grant Number 23H03468 and 23K18495, as well as from JST ALCA-Next Grant Number JPMJAN23F3.

References

1. P. Kmecl, M. Munih, and J. Podobnik, Towards Autonomous Retail Stocking and Picking: Methods Enabling Robust Vacuum-Based Robotic Manipulation in Densely Packed Environments. *Sensors*, 2024, *24*, 6687.
2. K. Wada, New robot technology challenge for convenience store, *2017 IEEE/SICE International Symposium on System Integration (SII)*, Taipei, Taiwan, 2017, pp. 1086-1091.
3. L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, 2021, *J Big Data* 8, 53.
4. 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
5. N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár and C. Feichtenhofer, SAM 2, arXiv:2408.00714
6. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Look Only Ones, arXiv:1506.02640
7. 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
8. 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.

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.

Prof. Hakaru Tamukoh

Hakaru Tamukoh received the B.Eng. degree from Miyazaki University, Japan, in 2001, and the M.Eng. and Ph.D. degrees from the Kyushu Institute of Technology, Japan, in 2003 and 2006, respectively. He was a Postdoctoral Research Fellow at the Kyushu Institute of Technology, from April 2006 to September 2007. He was an Assistant Professor with the Tokyo University of Agriculture and Technology, from October 2007 to January 2013. He is currently a Professor with the Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology. His research interests include digital hardware design, neural networks, and home service robots.