Application of Vision-Based Object Recognition and Grasp Strategy Determination for Multi-Functional Robotic Grippers

Ki Hong Im [a*] and Pil Geun Jang [a]
[a]*Korea Atomic Energy Research Insititute, 111, Daedeok-Daero 989Beon-Gil, Yuseong-Gu, Daejeon, KOREA*
[*]*Corresponding author: khim@kaeri.re.kr*

***Keywords :** Object Recognition, Grasp Strategy, YOLO, Resnet, Large Multimodal Model

1. Introduction

Robotic manipulation with manipulators and grippers has been widely adopted in fields such as manufacturing, logistics, and service industries. Traditionally, task-specific grippers—such as suction devices for packaging or parallel-jaw grippers for assembly—have proven effective within their designed domains. However, as robotic systems are increasingly deployed in unstructured and dynamic environments, there is a growing demand for versatile platforms capable of handling a broad variety of objects with different geometries, materials, and poses. Such application scenarios are widely observed, ranging from object handling on factory conveyor belts to debris manipulation in nuclear environments. Under these conditions, conventional single-mode grippers exhibit inherent limitations in terms of flexibility and reliability.

To address this challenge, multi-functional grippers have been introduced, combining operation modes such as suction and finger-based grasping to expand their applicability. These devices enable a single robot to manipulate both flat, smooth objects and irregular, deformable ones. Yet, the potential of multi-functional grippers cannot be fully realized without intelligent perception and decision-making. The system must not only recognize objects accurately but also evaluate their categories and poses in order to determine the most suitable handling strategy in real time.

This study focuses on implementing such perception-driven functionality. We present a vision-based framework that integrates YOLO [1,2] with a large multimodal model (LMM) to improve recognition accuracy while filtering out untrained objects to reduce misclassification. Building upon this perception layer, a ResNet-based model determines whether suction or finger-based grasping is optimal for each recognized object by considering both its category and pose. These functions are tested with target objects and tens of untrained objects, and the resulting accuracy of each function is evaluated by multiple vision tests on each object.

2. Overview of the Proposed System

2.1 Multi-Functional Gripper

By employing a multi-functional gripper, stable and efficient grasping can be achieved for a wide range of objects and object poses. As illustrated in Fig. 1, the gripper is designed with two operation modes—a multi-finger grasping mode and a suction mode—which can be switched as needed or used in a hybrid manner. The multi-finger grasping mode is particularly suitable when the object has an uneven surface, irregular shape, or limited contact area, whereas the suction mode is more effective when the object surface is smooth or the object height is relatively low.

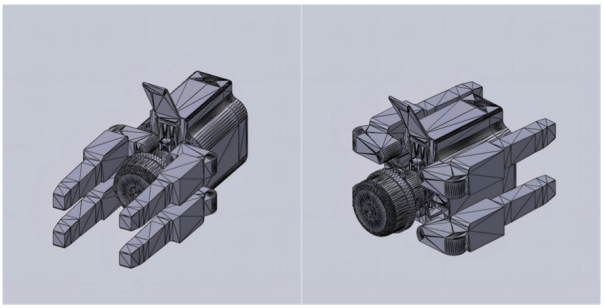

Fig. 1. Grasping mode and Suction mode of the multi-functional gripper

2.2 Object Recognition

Object images were acquired using an RGB-D camera to train object recognition models, including a custom YOLOv8 model. As shown in Fig. 2, the dataset consisted of 11 target objects: seven representative items commonly found on a production-line conveyor belt, and four objects with the shape of machined metallic components. A total of more than 9,000 images were collected for custom YOLOv8 training. In addition, a commercially available large multimodal model (LMM) was employed with a domain-specific corpus to enhance recognition of specific object groups during the query process.

Fig. 2. Example objects from the eleven target classes used in this study

In operation, real-time images obtained from the RGB-D camera are simultaneously processed by the custom YOLOv8 model and the LMM. The final decision is derived from the combined outputs of the two processes and categorized into three outcomes: Classified, Unclassified, and No Detection. Classified indicates that the recognized object belongs to a trained class; Unclassified indicates that the object is detected but belongs to an untrained class; and No Detection indicates that no object was detected. To support the recognition pipeline and experimental evaluation, a dedicated custom GUI program was developed for efficient visualization and data logging.

Fig. 3 illustrates two representative cases: in one, both YOLO and LMM recognized the object as "APPLE," resulting in a final decision of Classified; in the other, YOLO misclassified a human hand as "ELEPHANT," while LMM correctly recognized it as "Hand," leading to a final decision of Unclassified. Experimental evaluation was conducted using 11 trained objects and 25 untrained objects. As summarized in Table Ⅰ, the average recognition success rate reached 99% for trained classes and 100% for untrained classes. Notably, the integrated YOLO–LMM approach showed a substantial improvement in classification accuracy compared to using YOLO alone.

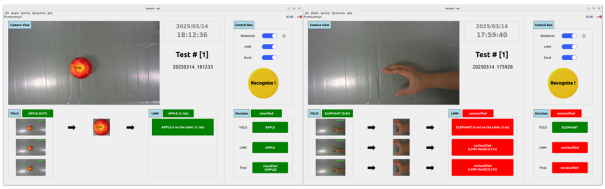

Fig. 3. Examples of final decision outcomes: Classified and Unclassified

Table I: Experimental results for trained and untrained classes using YOLO alone versus the integrated YOLO-LMM approach

	YOLO alone	Integrated YOLO-LMM approach
Trained Classes	99.091%	99.091%
Untrained Classes	79.6%	100%

2.3 Grasp Strategy Determination

Grasping strategies were predefined for the 11 trained object classes to ensure stable handling across diverse object categories and poses. A custom model based on ResNet was developed to predict the appropriate grasping strategy. When an object was categorized as Classified during the recognition stage, the same input image was subsequently processed by the ResNet model to determine whether suction or finger-based grasping should be applied.

To integrate and record both object recognition and grasp strategy decisions, a dedicated GUI program was employed. Fig. 4 shows an example in which an object recognized as Classified was assigned the strategy of finger-based grasping. As summarized in Table II, experimental results demonstrated an average prediction success rate of 98.9% for grasp strategy determination across the trained object classes.

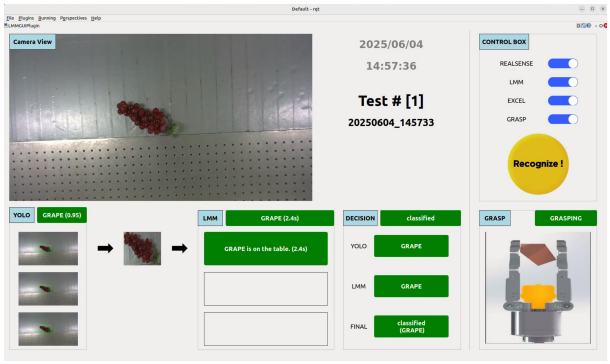

Fig. 4. Example of final decision in grasp strategy determination: Grasping

Table II: Grasping strategy prediction success rate for trained classes

	Grasping strategy prediction
Trained Classes	98.853%

3. Conclusions

In this study, object recognition accuracy for specific object groups was enhanced by integrating YOLO with a large multimodal model (LMM), thereby reducing the misclassification of untrained classes often observed in custom YOLO models. In addition, a ResNet-based model was employed to classify object poses and predict the corresponding grasping strategies, achieving a high prediction success rate. These results demonstrate that applying vision-based techniques to multi-functional grippers enables stable and efficient object handling.

Future research will extend this work by incorporating precise object pose estimation and integrating the recognition results from the YOLO–LMM framework with the grasping strategies predicted by ResNet to control a robotic manipulator. Ultimately, this will enable the gripper to perform robust and adaptive grasping of objects in real-world scenarios.

ACKNOWLEGEMENT

This work was supported by Robot Industry Core Technology Development Programs of the Ministry of Trade, Industry & Energy of KOREA(20018270)

REFERENCES

[1] YOLOv8, https://docs.ultralytics.com/models/yolov8/

[2] Im, Ki Hong, Pil Geun Jang, and Young Kyun Kim, "A Study on Integrating YOLO and Large Multimodal Models for Improved Object Recognition," Proc. of the Korean Nuclear Society Spring Meeting, 2025, Jeju, Korea.