Transactions of the Korean Nuclear Society Autumn Meeting
Changwon, Korea, October 30-31, 2025

Vision-Guided Servoing of a Manipulator via Object Recognition and Pose Estimation

Ki Hong Im [a*], Pil Geun Jang [a] and Young Kyun Kim [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, Pose Estimation, YOLO, Servoing

1. Introduction

Vision-guided robotic manipulation has emerged as a key technology for enabling robots to interact with complex and dynamic environments. Unlike traditional preprogrammed motion strategies, vision-guided servoing enables continuous adaptation of manipulator trajectories based on real-time perception, thereby improving accuracy, flexibility, and robustness. This capability is critical in applications ranging from general industrial automation to extreme environments, including space exploration—where planetary rovers equipped with manipulators must autonomously identify and handle unknown objects under highly uncertain conditions—as well as the manipulation of contaminated objects in nuclear facilities.

A major challenge in this domain is the reliable estimation of object poses, as both recognition errors and sensor noise can significantly degrade servoing performance. To address this, object recognition and pose estimation must be tightly integrated into the control loop. In particular, robust recognition ensures correct object identification, while accurate six-degree-of-freedom (6-DoF) pose estimation provides the spatial information required to guide the end effector.

In this study, we develop a vision-guided servoing framework that combines a custom-trained YOLO-based object recognition module [1,2] with depth camera sensing to estimate object poses in real time. The detected object classes and their depth information are jointly utilized as inputs to a pose estimation process, producing accurate 6-DoF position and orientation data. This pose information is then employed to continuously update the trajectory of the manipulator's end effector, enabling it to track the object's motion and achieve adaptive servoing. To further enhance stability, noise components in the estimated pose—particularly orientation jitter—are filtered through a smoothing process, ensuring reliable control performance even under fluctuating sensory conditions.

The proposed approach demonstrates that by integrating deep-learning-based recognition with depth-based pose estimation and stabilizing noisy pose signals, a manipulator can achieve precise and robust servoing. Experimental validations confirm its feasibility for real-world applications, highlighting its potential for both terrestrial industrial tasks and space robotics scenarios where adaptability and reliability are paramount.

2. Model Preparation and Evaluations

2.1 Test Environment

Fig. 1. Images of the 19 target objects used in this study

Fig. 1 illustrates the 19 target objects used in this study. For each object, approximately 250 images were prepared, including both original and augmented data, and a custom YOLOv8 model was trained using a few-shot learning approach. For pose estimation, we adopted a zero-shot model that requires 3D object models and processes RGB-D data as input, enabling its use without prior training on specific object categories.

The 3D object model files required for pose estimation were generated using the Neural Radiance Fields (NeRF) method. Specifically, 10 images of each object were captured under fixed camera and object conditions, and NeRF was applied to reconstruct the models. To enhance the accuracy of reconstruction, YOLO-based segmentation was incorporated. Mask information provided by YOLO was used to extract only the corresponding regions in the depth maps, thereby eliminating irrelevant background data and producing clean object models. Representative examples of the generated models are shown in Fig. 2.

The hardware configuration for this study employed an RGB-D camera to capture images for both object recognition and pose estimation, while a Universal Robots UR10 manipulator was utilized to implement the visual servoing system.

Fig. 2. Object model files generated using the NeRF method with YOLO

2.2 Object Recognition and Pose Estimation

The first RGB-D frame acquired from the camera was processed using the custom-trained YOLO model for object recognition, where the label and mask information were extracted and passed to the pose estimation module. The pose estimation process consists of two stages: register and track.

In the register stage, the label information obtained from YOLO was used to load the corresponding pre-generated object model file. With the model and mask data, the pose estimation algorithm [3] computed the 6-DoF position and orientation of the target, as illustrated in Fig. 3. In the subsequent track stage, no additional model upload was required. Instead, the information stored during the register stage was reused, allowing continuous pose estimation without repeated object recognition or model loading. This design significantly improved real-time performance. However, the 6-DoF estimates obtained during the track stage exhibited noticeable noise, particularly in orientation.

Fig. 3. Example of pose estimation model execution

Fig. 4 shows the position and orientation results when both the object and the camera remained stationary. While the position data were relatively stable, the orientation values displayed significant fluctuations. Fig. 5 provides enlarged plots of the x- and y-orientation components, highlighting the most severe noise.

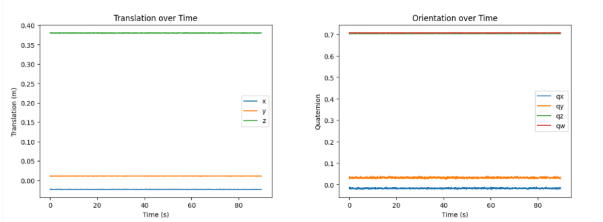

Fig. 4. Results of pose estimation model execution under stationary object and camera

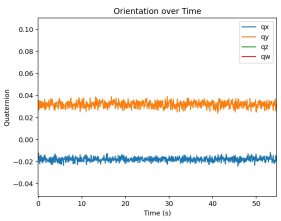

Fig. 5. Enlarged graph of the x and y components of orientation

To mitigate this issue, a Kalman filter and a moving average filter were applied to both position and orientation estimates. As shown in Fig. 6, the filtering process substantially reduced orientation noise, producing smoother and more stable pose data suitable for servoing control.

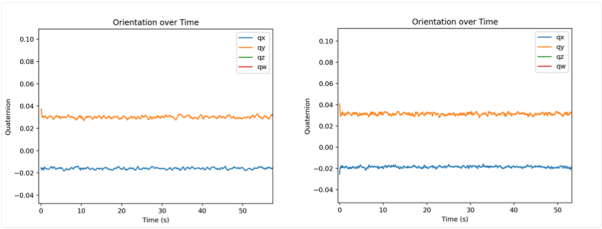

Fig. 6. Enlarged graph of the x and y components of orientation after applying Kalman filter and moving average filter

2.3 Visual Servoing Test

Visual servoing was implemented to control the manipulator's end effector in real time using the filtered 6-DoF pose information of the target object. The estimated position and orientation values were provided as input to the manipulator, enabling the end effector to track the object's motion while maintaining a constant offset in both horizontal and vertical directions. By combining YOLO-based recognition, depth-informed pose estimation, and noise-filtered signals, the system achieved stable vision-guided servoing performance.

3. Conclusions

This study presented a vision-guided servoing framework that integrates custom object recognition and pose estimation for robotic manipulators. A few-shot trained YOLO model was used to recognize 19 target objects, while a zero-shot pose estimation model employing pre-generated object models provided 6-DoF information from RGB-D input. By applying Kalman and moving average filters, noise in both position and orientation estimates was effectively reduced, enabling stable visual servoing of the manipulator.

The proposed approach demonstrates the feasibility of combining learning-based recognition, depth-informed pose estimation, and noise filtering to achieve reliable real-time servoing. These results suggest strong potential for application in both industrial automation, where robust and adaptive manipulation is required, and in space robotics, such as planetary rovers equipped with manipulators operating under uncertain environments. Future work will focus on extending this framework to dynamically moving objects and implementing robust grasping strategies with robotic grippers for complex manipulation tasks.

ACKNOWLEGEMENT

This work was funded by the Ministry of Trade, Industry and Energy and the Korea Planning and Evaluation Institute of Industrial Technology (No.RS-2024-00432390)

REFERENCES

[1] YOLOv8, https://docs.ultralytics.com/models/yolov8/
[2] Im, Ki Hong, Pil Geun Jang, 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.
[3] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In CVPR, 2024.