Robot-based Object Pose Auto-annotation System for Dexterous Manipulation

Sunme Park, Yujin Kim, Seungwon Lee, Mingi Jung, and Jongbum Park
Korea Electronics Technology Institute
Intelligent Robotics Research Center
Bucheon, South Korea
{pishp00200, kyj0301, seung, minngi, jbpark}@keti.re.kr

***Abstract*—To effectively employ robots in real-life scenarios, comprehending information about the surrounding objects is crucial for achieving dexterous manipulations. Since vision-based recognition relies on reliable high-quality annotated data, we present a 6D object pose auto-annotation method based on robot kinematics. This system directly measures object orientation and position through real-world robot operations by applying coordinate transformations among the robot, camera, and object. Leveraging robots enables automatic and precise object pose measurement while minimizing human effort. Depending on the robot's posture, capturing various facets of the object becomes straightforward. Furthermore, we account for occlusion caused by various robot hand grip styles to collect a diverse range of visible object portions while maintaining precise poses. The visualization code and data examples are publicly available at https://github.com/SunmePark/robot-based_annotation."**

***Index Terms*—Robot-based auto annotation, Object pose annotation system, 6D object pose estimation, Occlusion**

I. Introduction

Environmental perception is essential for service robots to interact with their surroundings and individuals. The robots are expected to accomplish dexterous tasks reducing human labor or assist individuals with disabilities or the elder by aiding them in their every day grasping activities. For the robot manipulation of objects, it is crucial to identify the presence of objects and their spatial arrangement. Accurate pose estimation is crucial to ensure safety in robot grasping. With the advancements in deep learning, numerous studies are currently underway in the field of vision-based object pose estimation. In this case, the high quality and large amount of pose-annotated data play a crucial role in accomplishing the learning performances [1].

In existing object pose datasets, there are image frames where a single object lies [2], multiple objects occlude each other [3] or are occluded by human hands [4]. However, when robots manipulate objects within their workspace, occlusions can occur in more complex forms due to the different finger mechanisms and sizes between the robot hand and the human hand. We propose an object pose dataset for dexterous manipulation considering occluded portions caused by the robot hand. The object images obscured by the robot hand provided with accurate pose data will help real robot operation utilizing reinforcement learning or imitation learning for dexterous tasks.

This work was supported by the Technology Innovation Program (20018112, Development of autonomous manipulation and gripping technology using imitation learning based on visualtactile sensing) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea)

For creating object pose annotation, there is two common approaches: real image annotation and using a synthetic image [5]. For the real image, humans manually mark keypoints in all RGB sequences [4] or partially annotate then refine poses by tracking [3]. There also exist several types of annotation tools for object pose estimation [6], [7]. However, it still entails significant costs in terms of human effort and is susceptible to human errors. When generating synthetic images, a gap with the real-world environment exists [4]. To address these issues, we propose an auto annotation method based on measuring object poses in real images by applying robot kinematics through real-world robot operation. Since robot shows high precision in manipulating workpieces [8], we calculate object poses from robot poses based on the relation between the camera, robot, and object.

The contributions of this paper are as follows.

1. We construct a 6d object annotation system utilizing robots to automatically annotate the poses of objects based on robot kinematics. Our system can improve annotation accuracy and labor intensity compared to conventional methods.

2. Various object orientations and locations can be collected according to the operating robot manipulation. Additionally, diverse occluded portions caused by the robot hand grasping are also considered which can be helpful to the robustness of object pose estimation task.

II. Robot-based Annotation

Depicted in Fig.1, our annotation system is constructed based on the pose relationship between the robot, object, and camera. Object pose that contains location and orientation in a 3D space is determined using robot kinematics. These poses are projected onto 2D images using the camera's intrinsic parameters, verifying the feasibility of the proposed annotation system. In this section, we use notation ${}_{\alpha}T_{\beta}$, which indicates transformation matrix from α to β.

The data we provide is the pose of objects $\{O\}$ in relation to the camera $\{C\}$ ${}_{C}T_{O}$. This can be obtained by multiplying the transformation matrices between the camera and the robot base $\{B\}$ ${}_{C}T_{B}$, and between the robot and the object ${}_{B}T_{O}$ as shown in (1).

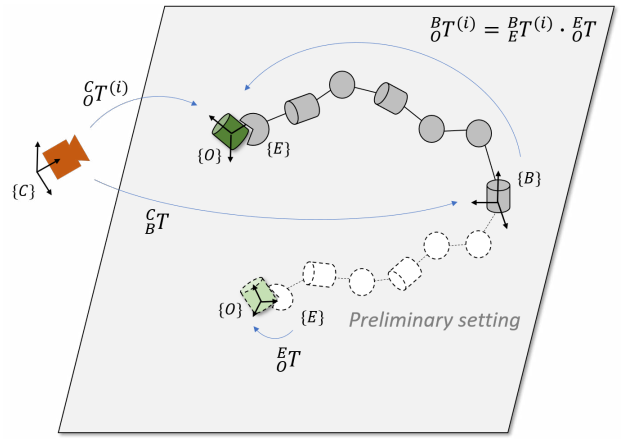

Fig. 1. Scheme of robot-based annotation system

$$ {}_{C}T_{O}^{(i)} = {}_{C}T_{B} \cdot {}_{B}T_{O}^{(i)} \tag{1} $$

${}_{C}T_{O}^{(i)}$ represents the 6d object pose with respect to the camera at i-th frame. ${}_{C}T_{B}$ is a fixed transformation matrix between the static camera and the robot base, determined during the preliminary setup. ${}_{B}T_{O}^{(i)}$ represents the object pose with respect to the robot base, obtained for each frame as the robot's posture changes. It can be calculated using the following equation.

$$ {}_{B}T_{O}^{(i)} = {}_{B}T_{E}^{(i)} \cdot {}_{E}T_{O} \tag{2} $$

${}_{B}T_{E}^{(i)}$ represents the end-effector $\{E\}$ pose obtained from the robot controller for the i-th frame. ${}_{E}T_{O}$ can be initially determined during the preliminary setup, attaching checkerboard to object. Upon the initial object grasp, robot posture was utilized to calculate ${}_{E}T_{O}$ using (3).

$$ {}_{E}T_{O} = ({}_{B}T_{E}^{(0)})^{-1} \cdot {}_{B}T_{C} \cdot {}_{C}T_{O}^{(0)} \tag{3} $$

For the first frame 0, the robot grasps the object and aligns it in a way that the checkerboard is visible to the camera. Then we measure end-effector's pose ${}_{B}T_{E}^{(0)}$ for this posture and the camera's extrinsic parameters for the checkerboard $\{CK\}$ on object to derive ${}_{C}T_{O}^{(0)} = {}_{C}T_{CK} \cdot {}_{CK}T_{O}$ at the same time. With ${}_{B}T_{C}$ obtained by inverting ${}_{C}T_{B}$, we can consequently specify ${}_{E}T_{O}$. Further details are explained in Sec III-B.

III. Experiment

A. Hardware configuration

1) Robot manipulator: The xArm6 with a payload of 5kg from Ufactory is used as a robot manipulator. It has 6 degrees of freedom and we obtain an end-effector pose to the robot base ${}_{B}T_{E}$ through Python robot controller SDK.

2) Robot hand: We utilized a dexterous gripper, Allegro Hand which has four fingers with 16 degrees of freedom. The grasp control was achieved through ROS. The four fingers were all employed to grasp the object firmly, ensuring the prevention of slip occurrences.

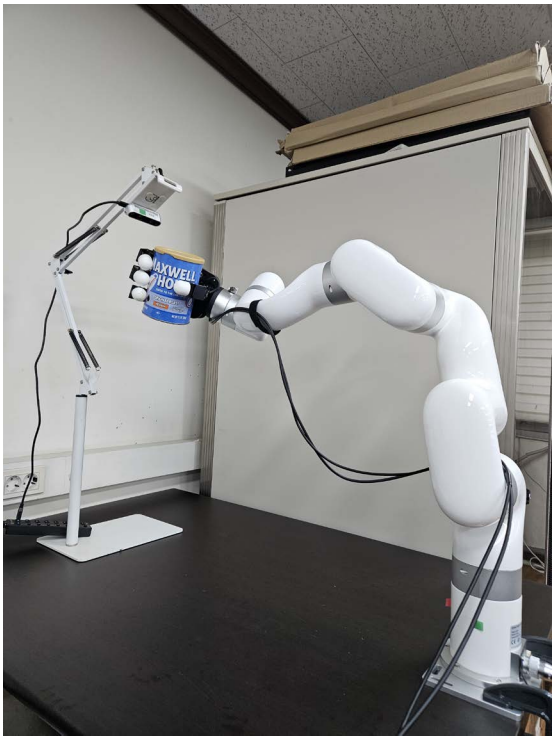

Fig. 2. Laboratory setup

3) RGBD camera: For capturing images, the Realsense RGBD camera d435i was employed. Images were captured at a resolution of 480x640, with synchronized RGB and depth image output. Intrinsic parameters were also provided.

4) Object: The selected target objects primarily encompass YCB objects, widely utilized in previous object pose estimation studies. We attached a checkerboard to the center of the object's base. We conducted on various shapes of objects such as mustard bottle, coffee can, and so on.

B. Preliminary setting

1) Camera-robot Base calibration: First of all, the pose relationship between the camera and the robot should be determined. Both the camera and the robot were statically positioned and we used a checkerboard affixed to the robot's end-effector. Through a captured image by the camera, extrinsic parameters were extracted which refers to ${}_{C}T_{CK}$. The transformation matrix from the robot base to the camera ${}_{B}T_{C}$ was subsequently derived by multiplying the end-effector's pose ${}_{B}T_{E}$ and inverse of ${}_{C}T_{E} = {}_{C}T_{CK} \cdot {}_{CK}T_{E}$ where ${}_{CK}T_{E}$ can be manually obtained.

2) Object-End-effector calibration: Since ${}_{B}T_{O}$ cannot be directly obtained, as derived in (2), we acquire ${}_{E}T_{O}$, the transformation matrix from the robot's end-effector to the object. To derive ${}_{E}T_{O}$, we attached checkerboard $\{CK\}$ to the objects $\{O\}$. Given the flexibility afforded by the position and orientation of the checkerboard, we have attached it to the underside center of the object, ensuring that the z-axis aligns upward. The robot hand firmly grasps the rigid object to prevent slipping and aligns itself such that the checkerboard is

visible to the camera. During this process, the pose of the end-effector ${}_{B}T_{E}^{(0)}$ is measured, and through camera-checkerboard extrinsic calibration ${}_{C}T_{CK}^{(0)}$, ${}_{E}T_{O}$ is determined as explained in (3) and saved for each grasp.

C. Object Pose Acquisition

Following the preliminary setting, the robot hand's grasp should remain unchanged. In each subsequent frame, by altering the robot's posture, we acquire ${}_{B}T_{E}^{(i)}$. Given ${}_{E}T_{O}$ and ${}_{C}T_{B}$, we obtain ${}_{C}T_{O}^{(i)}$ as derived in (1) and (2). This process automatically annotates the object pose as the robot's posture changes, thereby reducing human effort. As the robot can capture images from various angles while maintaining its grasp, diverse object poses can be collected.

The 3D bounding box (bbox) can be obtained using the transformation from the attached checkerboard to each of its corner points. Fig. 3 shows the projection of the 3D bbox onto a 2D image using calculated pose and camera intrinsic parameters.

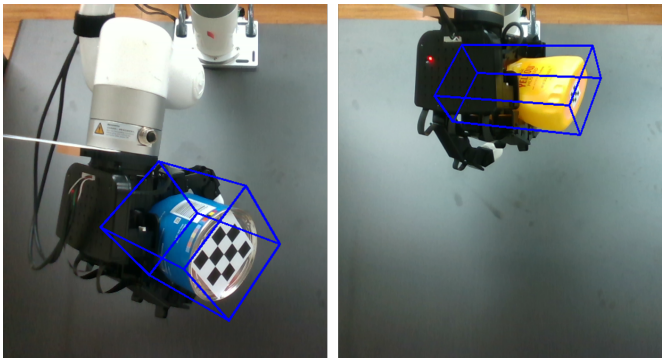

Fig. 3. Example of coffee can and mustard 3d bounding box

IV. Robot-based Object Pose Dataset

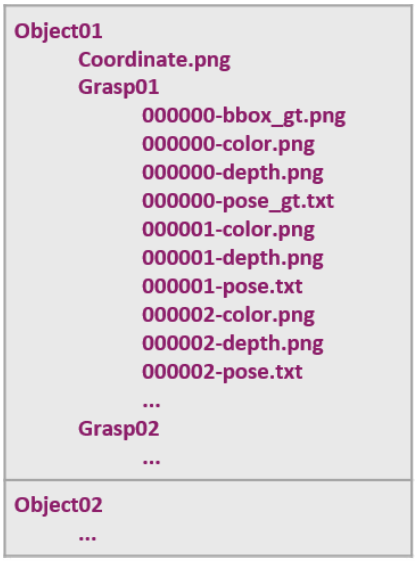

Fig. 4. Structure of Robot-based Object Pose Dataset

The data includes RGB images, depth images, camera information, object coordinate information, and object 6D poses, as illustrated in Fig. 4. '000000' refers to the first captured setting during calibration between object and end-effector where the checkerboard is visible (See in III-B2). The 'gt' suffix denotes ground truth specifically derived from camera's extrinsic parameter determined using a checkerboard attached to the object to validate calculation. Subsequently, we changed the robot's posture for each frame and stored RGBD image paired with a calculated pose. As we conducted multiple grasping postures, 'G' stands for grasp in case of grasping in various manners which affect the proportion of occlusion area.

V. Conclusion

In this study, we have introduced a novel method for annotating object poses based on the transformation relationship between the robot, camera, and objects. Object poses derived from the robot's posture can be automatically generated, reducing human errors and effort. Moreover, object poses can be diversified according to the robot's stances. The annotation process is capable of handling occluded portions caused by the robot hand. By providing real-world data, we anticipate that this approach will aid in reducing the sim-to-real gap.

In future research, we plan to construct datasets applicable to tracking since generating robot trajectories is relatively straightforward. We also aim to minimize constraints on the robot and camera positions, allowing for more diverse experimental backgrounds in the laboratory. Furthermore, given object size, we expect to generate pose annotation for the custom objects.

References

[1] A. Salari, A. Djavadifar, X. Liu, and H. Najjaran, "Object recognition datasets and challenges: A review," *Neurocomputing*, vol. 495, pp. 129–152, 2022.

[2] Y. Ze and X. Wang, "Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 469–27 483, 2022.

[3] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[4] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield *et al.*, "Dexycb: A benchmark for capturing hand grasping of objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9044–9053.

[5] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "Honnotate: A method for 3d annotation of hand and object poses," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3196–3206.

[6] A. Ahmadyan, L. Zhang, A. Ablavatski, J. Wei, and M. Grundmann, "Objectron: A large scale dataset of object-centric videos in the wild with pose annotations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7822–7831.

[7] Y. Xiang, W. Kim, W. Chen, J. Ji, C. Choy, H. Su, R. Mottaghi, L. Guibas, and S. Savarese, "Objectnet3d: A large scale database for 3d object recognition," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 2016, pp. 160–176.

[8] E. Matheson, R. Minto, E. G. Zampieri, M. Faccio, and G. Rosati, "Human–robot collaboration in manufacturing applications: A review," *Robotics*, vol. 8, no. 4, p. 100, 2019.