Object 6D Pose Recognition and Grasping and Placing Based on Machine Learning

Dongrui Li 12112208
Zhaokai Chen 12212220
Songran Wang 12010114
Haoyang Xiang 12010611
Zuo Xu 12111625
Yifei Luo 12012039

I. Introduction

With the increasingly widespread application of machine learning, its usage spans across numerous domains. Specifically, there is a growing demand for object 6D pose recognition and grasping in various sectors such as the logistics and waste management industries. In the logistics sector, robots equipped with advanced 6D pose recognition technology can automate package handling, thereby enhancing sorting efficiency and reducing labor costs. Similarly, in waste management, the precise identification and grasping capabilities of objects in 6D pose can assist robots in accurately categorizing recyclable and non-recyclable materials, fostering environmental awareness and enabling effective resource recovery. These applications combine machine learning technologies with practical needs, presenting new opportunities for development in diverse industries.

In this article, we propose using a computer vision-based machine learning approach to achieve accurate object pose recognition. Specifically, our goal is to apply this method to the robotic arm environment built using Robosuite. In this setup, multiple cameras are strategically positioned to capture images of objects from different angles. By extracting information from these images and utilizing a trained model, we achieve precise identification of object poses. Additionally, by leveraging the forward and inverse kinematics equations of the robotic arm and planning predefined trajectories, we are able to control the robotic arm to grasp and place objects at specified locations. This method is specifically designed to meet the demands of grasping and placing tasks within a simulated environment, thereby enhancing task efficiency and accuracy.

II. Problem Statement

To achieve object 6D pose recognition, we utilize image data obtained from multi-view RGB-D data and preprocess these images by denoising and enhancing contrast. We combine the multiple angle images using the DenseFusion method and employ supervised and iterative learning to train the 6D pose recognition model.

The data used in this article is sourced from The Linemod dataset. It is a widely utilized object recognition dataset featuring 15 objects such as bottles, cups, televisions, keyboards, etc. The dataset provides multiple perspectives of RGB-D images and corresponding 3D models. These images and models serve as valuable resources for training and evaluating object recognition algorithms.
We aim to develop a robust model capable of accurately recognizing the 6D pose of objects and integrating it with a robotic arm to achieve precise grasping and placing of objects.

For the accuracy of object pose recognition, we will conduct supervised learning during the machine learning phase. The model will return the error value of its recognition on known data. Secondly, we plan to assess the accuracy of object pose estimation in the simulation phase by evaluating the smoothness of grasping objects with the gripper. Additionally, we will evaluate the entire grasping process and placement results to determine the accuracy of the robotic arm in achieving precise grasping and placing of objects.

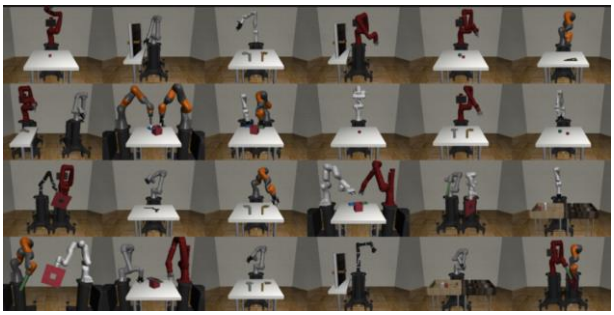
III. LITERATURE REVIEW

Researchers from Stanford Vision and Learning Lab and Stanford People, AI Robots Group, led by Wang, have conducted research on 6D pose estimation and grasping using DenseFusion. The model takes RGB-D images as input and predicts the 6D pose of each object in the frame. This network is implemented using PyTorch, with the rest of the framework written in Python. Semantic segmentation/instance segmentation methods can be chosen according to specific needs. In their research, they provide implementations of the DenseFusion model, iterative refinement model, and a basic SegNet semantic segmentation model used in real robot grasping experiments.

IV. TECHNICAL APPROACH AND PRELIMINARY RESULTS

A. Robotic Arm Platform

We plan to use the single-arm environment DEMO provided by ROBOSUITE as the simulation environment. This environment consists of a 6-DOF robotic arm, a gripper, and a shelf. To enhance the observation perspectives, we will add multiple cameras to the simulation environment. We will utilize ROBOSUITE's camera configuration functionality to define the properties of these cameras, such as pose, pointing direction, field of view, and resolution, to render the environment's images. During the motion simulation, the cameras will capture images of the robotic arm and objects. We will use ROBOSUITE's data recording and export functionality to save this image data as CSV files or other formats.



B. CV

Preprocessing of multi-angle image data includes several steps: background removal, image alignment, cropping, and image enhancement. During background removal, irrelevant background information is eliminated from the images to reduce noise and enhance the visibility of the target objects. Subsequently, image alignment ensures that images from different perspectives are aligned in the same spatial coordinate system for further processing. Following that, cropping is performed to focus on the regions of interest in the images, reducing computational overhead and improving training efficiency. Finally, image enhancement techniques such as adjusting brightness, contrast, and color balance are applied to enhance the quality and features of the images.

Next, the preprocessed multi-angle image data is used as input for training with the DenseFusion algorithm. DenseFusion is a network architecture used for 6D pose estimation, predicting object poses from RGB-D data through end-to-end learning. During training, the iterative refinement model is employed to continuously optimize model parameters, enabling better adaptation to multi-angle image data and improving the accuracy and robustness of CV recognition.

C. DenseFusion

DenseFusion, proposed by Li Feifei et al. in 2019, is a network designed for 6D pose estimation. It operates as an end-to-end structure, taking RGB-D data as input and generating the 6D pose of an object as output. Its key innovation lies in introducing a pixel-level dense fusion approach to seamlessly integrate color and geometric features.

In practice, DenseFusion adopts a conventional network architecture for processing RGB-D data, replacing the depth map with PointNet. However, the initial network, PoseNet, exhibits subpar performance when faced with occluded objects. To address this limitation, a new network was trained with a similar structure, but its input was altered to the output of the initial network.

D. The Iterative Refinement of the Model

The iterative refinement of the model refers to the process of gradually improving and optimizing the model through multiple iterations during training. Initially, the model is initialized and trained using the training data, updating its param- eters through the backpropagation algorithm. Subsequently, the model is evaluated, and based on the evaluation results, its hyperparameters such as learning rate and regularization parameters are adjusted. This process is repeated until the model reaches a predefined stopping condition.

E. Implementation

Integrating the learned model into the simulation environment and interacting with other elements within the environment involves feeding multi-angle images recorded by cameras in the simulation environment into the learned model to predict the pose of objects. Subsequently, trajectory planning is performed on the robotic arm to facilitate the grasping and placing of objects.

ACKNOWLEDGMENTS

V. REFERENCES

[1] Wang C ,Xu D ,Zhu Y , et al. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. [J]. CoRR, 2019, abs/1901.04780

[2] Hui Q ,Shanhe G . Network architecture based on improved DenseFusion algorithm research on the recognition and grasping method of robotic arm[C], 2023:

[3] A. Zeng, K. T. Yu, S. Song, D. Suo, and J. Xiao. Multiview self-supervised deep learning for 6d pose estimation in the amazon picking challenge. IEEE, 2017.

[4] Tielin Zhang, Yang Yang, Yi Zeng, and Yuxuan Zhao. Cognitive template-clustering improved linemod for efficient multi-object pose estimation. Cognitive Computation, 12(4):834–843, 2020.