Multi-view self-supervised deep learning framework for solving 6D pose estimation problem

Li Zhidong 12011506
Ren Shize 12012818
Jiang Meng 12011603
Zeng Yuqi 12011811
Zhao Xuda 1200325

***Abstract*—Automatic logistics is a very hot topic recently, and picking is a significant part to implement this technology. To study further, we searched the past project of a famous competition: Amazon Picking Challenge (APC). From the projects we chose a Multi-view self-supervised learning[1] for our project.Limited to the device, we decided to reappear the project by ROBOSUITE[2] simulation.**

I. Introduction

Robots have been increasingly used in various industries, including manufacturing, logistics, and healthcare, to automate repetitive and dangerous tasks, improve efficiency and safety, and reduce costs. To achieve these goals, robots need to be equipped with advanced sensing, perception, and control capabilities. Computer vision (CV) is a critical technology that enables robots to recognize and understand the environment, objects, and people around them, and make decisions and take actions accordingly.

In this article, we present a CV-based approach to designing a 6-dof shelf gripping model using a robotic arm, a gripper, and a shelf. We use the single-arm environment DEMO provided by ROBOSUITE[2] and add multiple cameras to observe the movement of the robot and object from different angles. We collect motion and posture data of the robot and object from different angles, as well as image data captured by the cameras, and use them to train a model for CV recognition and control of object gripping from different angles.

The CV recognition is achieved by using a multi-angle self-supervised machine learning algorithm, where we capture a set of multi-view images of objects from multiple perspectives, preprocess them, and perform self-supervised learning. We employ several techniques to enhance the training effectiveness, including image enhancement, loss function design, and weight sharing. We also use the Iterative Closest Point (ICP) algorithm to optimize the pose estimation results and obtain the final 6D pose estimation outcome.

Finally, we solve the forward and inverse kinematics of the robot arm by analyzing the coordinates obtained from vision to control the gripping of objects. Our approach demonstrates the potential of CV-based methods in designing and controlling robotic systems for various applications.

II. Problem Statement

The problem addressed in the paper is instance-level object segmentation on few object categories in a warehouse setting. The proposed approach leverages multi-view RGB-D data and self-supervised, data-driven learning to reliably estimate the 6D poses of objects under a variety of scenarios.

The dataset used in the paper is a well-labeled benchmark dataset of APC 2016 containing over 7,000 images from 477 scenes.

The expected results are accurate and reliable estimation of the 6D poses of objects in a warehouse setting.

The evaluation is done using F-score across all benchmark categories going from 1% to 10% to 100% of training data, and several key components of the vision system are evaluated to determine whether they increase performance in isolation.

III. Literature Review

This paper is highly relevant to the topic of multi-view self-supervised deep learning for solving the 6D pose estimation problem. The paper proposes a novel approach for estimating the 6D pose of objects in cluttered scenes by leveraging multiple camera views and a self-supervised deep learning framework.

The proposed approach is particularly useful for applications such as robotic manipulation, where it is essential to accurately estimate the pose of objects in a cluttered scene. The use of multiple camera views enables the system to capture a more comprehensive view of the scene, which improves the accuracy of the 6D pose estimation.

Moreover, the self-supervised deep learning framework used in this paper eliminates the need for annotated data, which is typically time-consuming and expensive to obtain. Instead, the system learns from the raw data by generating its own training labels based on geometric constraints.Our project will mainly refer to this article

IV. Technical Approach And Preliminary Results

A. Robot

Designing a 6-dof shelf gripping model requires a 6-dof robotic arm, a gripper, and a shelf. Currently, the single-arm environment DEMO provided by ROBOSUITE[2] is being used, and the viewing angle needs to be set. Multiple cameras need to be added to the simulation environment to observe the movement of the robot and object from different angles. ROBOSUITE's camera configuration function allows cameras to bundle a name with a set of properties to render images of the environment such as the pose and pointing direction, field of view, and resolution. Cameras are defined in the robot and arena models and can be attached to any body, inheriting from MuJoCo.

Data collection is performed by running the simulation to collect motion and posture data of the robot and object from different angles, as well as image data captured by the cameras. The data can be saved as CSV files or other formats using ROBOSUITE's[2] data recording and export functions. We can specify one or several cameras we want to use images from when we create the environment, and the images are generated and appended automatically to the observation.

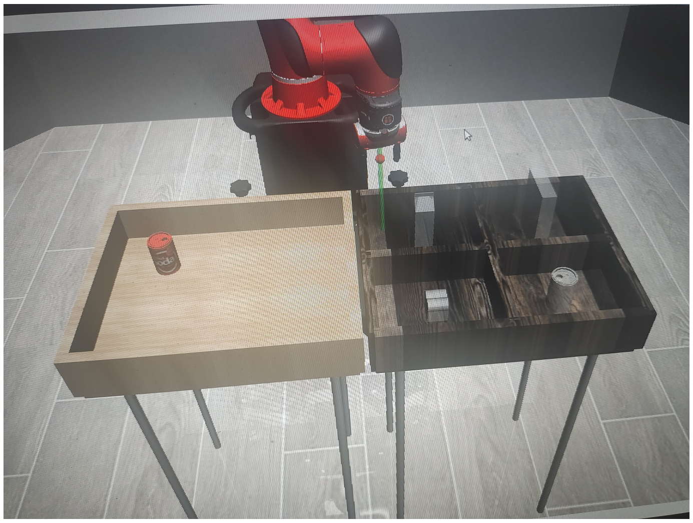

Fig. 1. Environment configuration according to the DEMO.

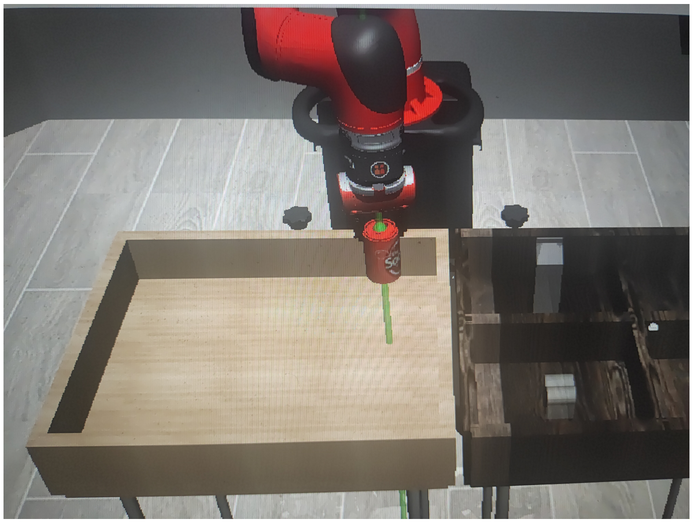

Fig. 2. Try to use the keyboard for clipping.

B. CV

Multi-view image generation Capture a set of multi-view images of objects from multiple perspectives. Then, the images are preprocessed, including background removal, alignment, and cropping. Next, we use this preprocessed set of images to perform self-supervised learning.

CV recognition is achieved by training the model using multi-angle self-supervised machine learning algorithm on the collected image data as input. This allows the machine to recognize and control the gripping of objects from different angles.

C. Self-supervised learning

In a pair of images from different views, one image is used as the ”query image” and the other as the ”positive sample”. Then, by rendering the query image using a 3D model, a ”synthetic image” is obtained as the ”negative sample”. Next, the three images are input into the deep neural network for training, with the goal of minimizing the distance between the positive and negative samples. During the training process, they used some techniques to improve the training effect, including image enhancement, loss function design, and weight sharing.

D. Implementation

1) Pose estimation: During the training process, we employ several techniques to enhance the training effectiveness, including FCN,image augmentation, loss function design, and weight sharing. Also utilized Iterative Closest Point (ICP) algorithm to optimize the pose estimation results, which eventually led to the final 6D pose estimation outcome.

2) Control: Solve the forward and inverse kinematics of the robot arm by analyzing the coordinates obtained from vision.

ACKNOWLEDGEMENTS

REFERENCES

[1] Andy Zeng, Kuan-Ting Yu, Shuran Song, Daniel Suo, Ed Walker Jr. au2, Alberto Rodriguez, and Jianxiong Xiao. Multi-view Self-supervised Deep Learning for 6D Pose Estimation in the Amazon Picking Challenge, 2017.

[2] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A Modular Simulation Framework and Benchmark for Robot Learning, 2022.