Interdisciplinary Project - Vision-Guided Robotic Grasping of Semi-Transparent Plastic Tubes

Haberger David Dylan 11705993
Domain-specific lecture: 330.273 Assistance Systems in Manufacturing 2
Main Supervisor: Dr.techn. J.-B. Weibel MSc
Co Supervisor: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Andreas Rauber

TU Wien

Abstract

Robotic manipulation of semi-transparent, non-rigid tubes, common in medical and pharmaceutical manufacturing, is hindered by the lack of reliable depth data and object models for pose estimation. This project addresses this challenge in two ways. A custom annotation tool, leveraging zero-shot object segmentation with the Segment Anything Model (SAM) and voxel carving, will be developed to efficiently generate ground truth data. This data is then used for training a YOLO object detection model to detect the position of the tube on a plane. The trained model's performance and robotic grasping will be evaluated in a real-world setting against a non learning traditional computer vision approach.

1 Motivation and Research Question

Using robotic systems to manufacture and test medical and pharmaceutical articles comes with great benefits, decreasing cost while increasing production, consistency and traceability. Some objects used in these applications however have properties that make robot manipulation comparably hard. One such object often found in these settings is semi-transparent non-rigid tubes. To manipulate an object, one first has to locate it, most methods of pose estimation depend not only on RGB images but also on reliable depth data which is not given due to the tubes translucent nature. Most methods also need a model of the object. In the case of the tube, neither is available. Another problem is the generation/annotation of the ground truth data needed to test and train the methods. To get ground truth masks from non-rigid objects, images have to be annotated pixel-wise by hand, which is an unfeasible approach to generating lots of data.

In this project, two things are investigated. First, a tool is to be created and evaluated that allows for faster annotation of segmentation masks using zero-shot object segmentation. Secondly, the effectiveness of using pixel-level object detection for grasping non-rigid semi-transparent tubes from a table plane is to be evaluated by training a model using the generated ground truth. The resulting pipeline is compared to an alternative approach using traditional non-learning Computer vision methods.

When generating a dataset with a robot-mounted camera, taking multiple images from different views per scene, the relative camera poses are known. To speed up the annotation process first a few masks are created from different perspectives, then voxel carving should be used to auto-prompt a zero-shot image segmentation model (Segment Anything) to automatically generate the segmentation masks for the remaining views.

The gained ground truth should then be used to train an object detection model (YOLO) to detect the tube location in an image. Assuming that the tube is lying on a flat surface, this information is sufficient to perform a simple top grasp with a robot arm if the surface plane is detected correctly.

2 Methodology

This project will follow the CRISP-DM methodology:

1. **Business Understanding:** The problem is defined as the difficulty in pose estimation and ground truth generation for non-rigit semi-transparent tubes, hindering robotic manipulation in medical and pharmaceutical manufacturing.
2. **Data Understanding:** An existing dataset of robot-captured images will be analyzed to understand the variations in scene configurations.
3. **Data Preparation:**
   - **Annotation Tool Development:** A tool leveraging Segment Anything Model (SAM) and voxel carving will be developed to accelerate mask annotation.
   - **Ground Truth Generation:** The developed tool will be used to generate set of segmentation masks, which will be manually refined to ensure a valid grundtruth.
4. **Modeling:**
   - **Object Detection Model Training:** A YOLO model will be trained on the generated dataset
   - **Baseline Comparison:** A traditional computer vision model will be used for comparison.
5. **Evaluation:**
   - **Annotation Tool:** Its effectiveness will be evaluated by measuring annotation time reduction, mask accuracy AP, IoU, F-Score, and the impact of initial manual annotations on performance.
   - **Object Detection Model:** Both the YOLO and baseline models will be evaluated using AP, IoU, F-Score.
   - Physical grasping experiments will assess real-world success rates.

3 Expected results

- **Annotation Tool:**
  - A annotation-tool that significantly accelerates the annotation process for non-rigid, semi-transparent objects, for robot captured datasets.
  - At least 50% reduction in annotation time compared to manual methods.
- **Object Detection Model (YOLO):**
  - A reliable model capable of accurately detecting the tube's location on a plane in real-time
  - At least 80% success rate in physical grasping experiments.
  - At least comparable performance to baseline computer vision approach.