Two Stage Vision System for Robotic Grasping and 6D Pose Estimation

Daniel Sawyer
Computer Science
University of South Florida
danielsawyer@mail.usf.edu

Tian Tan
Mechanical Engineering
University of South Florida
tiantan@mail.usf.edu

1. Introduction

Individuals with diminished physical capabilities must often rely on assistants to perform Activities of Daily Living (ADL). Comparing to having human assistant, assistive robots are more affordable and using assistive robot makes the user feel they have retrieved capabilities to live by themselves rather than have to live depend on other people. However, controlling a robot to perform activities of daily living are challenge even for healthy users. The solution for this is to give robot as much autonomy as possible and simplify the control as much as possible. In this project we will be focusing on building the vision system that enable the robot to perform pick and place autonomously while the user only need a touch on the screen to specify the object of interest. More specifically, the vision system will have two components, one is the object pose estimation (OPE) system which infers the object location and orientation in the scene and lead the robot to pre-grasp pose, another is the vision for grasp tuning (VGT) system which provides visual cues for self-adjustment in final grasping.

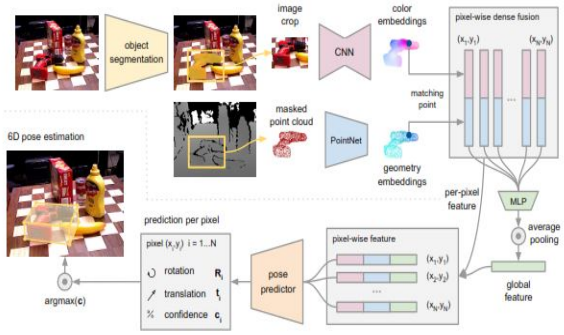
2. Problem statement

In this project we will be solving two computer vision problems, one is retrieving object 6D pose based on RGB-D data, the other is generate visual cues for grasp refinement using RGB data. For the OPE system we will implement a state of the art framework for 6D pose estimation of known objects called DenseFusion. We will modify the OPE framework by including our own object training set. For the VGT system we are planning to write our own program to detect the main object and find the center and edges of the object in a close up camera view.

3. Reading materials will be examined

DenseFusion 6D Object Pose Estimation by Iterative Dense Fusion will be our baseline for estimating pose. There will also be some form of object detection using either YOLOv3 or Faster R-CNN. We will also try and incorporate homography into our algorithms in order to try and increase performance. We will also be using the computer vision course's textbook.

Figure 1. DenseFusion framework



4. Evaluation

The system will be evaluated in three aspects. The 3D images taken from a depth camera will be run through our algorithms which will consist of different layers/modules that will process the image and output the information required in order for the robot to be able to interact with the real world and evaluate with the standard benchmarks and performance metrics. Vision for grasp tuning will be evaluated by comparing the result of detected object edges and center to human indicated ground truth edges and center. The overall system performance will be evaluated by the successful rate of grasping from subject testing.

Figure 2. Intel Realsense D435 RGBD Camera