Grasp Detection for Human-to-Robot Object Handover

Lucas Wohlhart
JOANNEUM RESEARCH ROBOTICS
Lakeside B08a, EG
9020 Klagenfurt am Wörthersee
lucas@wohlhart.at

ABSTRACT

This project presents an attempt to apply current state-of-the-art methods for grasp pose estimation to human-to-robot handover scenarios. The implemented method shall enable a robotic mobile manipulator to perform antipodal grasps on previously unknown objects presented by a human collaborator.

1. INTRODUCTION

Grasping is to be considered one of the fundamental object manipulation tasks a robot has to perform. In a **human-robot collaboration** scenario with a human giver handing over an object to a robot receiver the perception task is to determine the desired object transfer point and a corresponding grasp pose. This has proven to be challenging especially when facing unknown objects in unstructured environments. Driven by applications in fields such as warehouse automation or flexible manufacturing, recent advances in object agnostic robotic bin picking, mainly inspired by vision-based deep learning techniques, suggest that currently proposed methods are increasingly capable of solving these **grasp synthesis** tasks.

Mahler et al. [1] trained a neural network, dubbed grasp quality CNN (GQ-CNN), to learn the evaluation of a grasp success probability. The model is trained on the Dexnet-2.0 dataset; an extensive collection of synthesized RGB-D images annotated with corresponding grasp configurations. By iteratively ranking and resampling grasp candidates this method has shown to yield good proposals for unknown real world objects. Morrison et al. [2] propose a fully convolutional generative grasp CNN (GG-CNN) estimating individual maps for grasp quality, gripper angle and gripper width from a given 2 image. The resulting best grasp is determined by choosing the gripper configuration corresponding to the highest success probability encountered in the grasp quality map.

2. METHOD

Our method builds on the idea of estimating grasp configuration maps as in GG-CNN and extends the approach by adding a semantic segmentation layer to enforce scene understanding. This acts as guidance to focus on the region of interest for the object handover task and avoid estimating grasps that would collide with the hand of the human collaborator. The proposed fully convolutional neural network architecture is based on a U-Net inspired structure featuring encoder and decoder each comprised of four residual network blocks connected by an atrous spatial pyramid pooling layer to foster scale invariance. At the input stage, the network is fed a depth map acquired by an RGB-D sensor. The multi-headed output consists of a pixelwise semantic segmentation classifying as background, hand or object, a grasp center point quality map, a grasp angle map and a gripper opening width map. To obtain training data we extend the pipeline of DexNet 2.0 by combining it with the hand pose estimation data synthesis approach of Riegler et al. [3]. This enables us to render depth images and segmentation masks and to annotate corresponding grasp rectangles as introduced by Jiang et al.[4] for scenes in which a human presents an object to hand over in various poses and viewpoints.

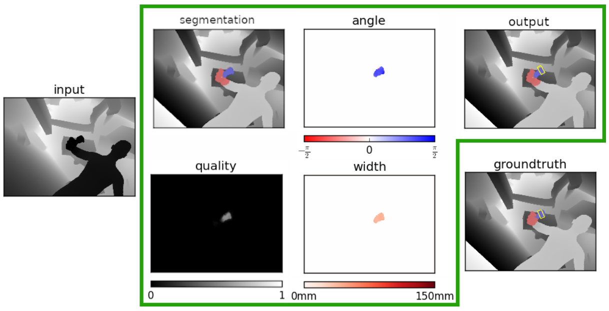

Figure 1: Left: input depth image. Green box: intermediate segmentation and grasp maps, resulting output estimated grasp configuration. Bottom-right: ground truth segmentation and grasp configuration

We are currently constructing a data acquisition pipeline to capture real world ground truth using RGB-D sensors, to bridge the sim-to-real gap due to noisy sensors by fine-tuning the trained model on such data.

3. REFERENCES

[1] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg. Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics.

[2] D. Morrison, P. Corke, and J. Leitner. Closing the Loop for Robotic Grasping: A Real-time, Generative Grasp Synthesis Approach.

[3] G. Riegler, D. Ferstl, M. Rüther, and H. Bischof. A framework for articulated hand pose estimation and evaluation.

[4] Yun Jiang, S. Moseson, and A. Saxena. Efficient grasping from RGBD images: Learning using a new rectangle representation.