Paper: Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning

Summary

In this paper, the authors present a method for learning hand-eye coordination for robotic grasping, using deep learning to build a grasp success prediction network, and a continuous servoing mechanism to use this network to continuously control a robotic manipulator. By training on over 800,000 grasp attempts from 14 distinct robotic manipulators with variation in camera pose, they achieve invariance to camera calibration and small variations in the hardware. Unlike other methods, their approach does not require calibration of the camera to the robot, instead of using continuous feedback to correct any errors resulting from discrepancies in calibration.

Their method consists of two components: a grasp success predictor, which uses a deep convolutional network to determine how likely a given motion is to produce a successful, and a continuous servoing mechanism that uses the CNN to continuously update the robot's motor commands. By continuously choosing the best-predicted path to a successful grasp, the servoing mechanism provides the robot with fast feedback to perturbations and object motion, as well as robustness to inaccurate actuation.

The grasp prediction CNN was trained using a dataset of over 800,000 grasp attempts, collected using a cluster of similar robotic manipulators. Even though the hardware parameters of each robot were initially identical, each unit experienced different wear and tear over the course of data collection, interacted with different objects, and used a slightly different camera pose relative to the robot base. These differences provided a diverse dataset for learning continuous hand-eye coordination for grasping.

To evaluate their continuous grasping system, they conducted a series of quantitative experiments with novel objects that were not seen during training. They found out that the success rate of their continuous servoing method exceeded the baseline and prior methods in all cases. For the evaluation without replacement, their method cleared the bin completely after 30 grasps on one of the 4 attempts and had only one object left in the other 3 attempts. The hand-engineered baseline struggled to accurately resolve graspable objects in clutter. They also evaluated the performance of their model under no replacement condition. They found that the grasp success rate continued to improve as more data accumulated, and a high success rate was not observed until at least halfway through the data collection process. Finally, they also report that their system learned to grasp softer objects by embedding the finger into the center of the object, while harder objects were grasped by placing the fingers on either side.

Strengths

- They explain their network architecture well.
- They also include drawbacks in their approach and some future works.

Weaknesses

- I would have liked to see how their model's performance improves as time increases.

Confusions

- I was a little confused about how they were relating their approach with RL.
- Why do they pass two images as an input to their CNN?
- How does replacement and without replacement experiment that they mention may yield different results?

Discussions

- Can we talk a little more about visual servoing? What kind of methods has been used for leaning visual servoing? How has deep reinforcement learning influenced this field?
- How successful supervised learning been in this problem domain?