Paper Review
Robotic Grasping of Novel Objects

Alvin Sun

October 4, 2021

1 Paper Summary

This paper introduces a learning method for generating grasps for novel objects that are unseen during training. The learning method, which does not require a full 3D model of the object, takes some hand-crafted features precomputed from the image space as input and produces the likelihood of grasp points (also in image space). From multiple view points, the probabilities of grasp points are back-projected into a discretized 3D space. The 3D grasp location is then solved by optimizing the log-likelihood in that 3D space. Their experiments showed that even trained with synthetic datasets, the robotic manipulator is capable of grasping unseen and even challenging (e.g. translucent or reflective) objects with high success rate.

2 What I Learned

1. It is intriguing to me that this multi-view correspondence problem can be solved in an implicit way. Instead of explicitly solving for point correspondence, casting rays into 3D space for all high probability image space locations can also spits out solutions for the best grasping location given measurements from different viewpoints. This method seems to be fairly robust against inconsistent detections across multiple views.

2. Traditional (instead of deep) learning methods seems to generalize pretty well across synthetic and real-world domains. I am quite surprised how well the learned algorithm generalizes on unseen objects during inference.

3 Opinions

3.1 Up Votes

- I agree with their approach on training with synthetic dataset. Given their problem formulation of finding grasp locations in image space, it is indeed very hard and time-consuming to hand label real-world datasets that is diverse in viewpoints and scale. Ray tracing is quite a good way to render photo-realistic data.
- I also like how they evaluate on novel objects that are unseen by the training dataset. This really demonstrated the generalizability of the learning algorithm.
- I strongly agree with their feature design for incorporating multi-scale features for the images. The effectiveness of multi-scale features are reflected in many modern state-of-the-art computer vision algorithms as well.

3.2 Down Votes

Despite all the great results obtained through the experiments, there are, however, some flaws in the theoretical analysis of their approach. The assumed independence between multiple observations as well as the independence between grid locations along the same ray is quite unrealistic. Reducing $P(y_j = 1|C1, \cdots, C_N)$ to $\prod_{i=1}^{N} P(y_j = 1|C_i)$ is rather an engineering hack than a sound derivation in my opinion.

4 Evaluations

The goal of this project is to introduce a learning method for general grasping of novel objects without knowing a prior knowledge on the shape of the object. This a perfectly valid objective as previous work, learning or non-learning based, has been focusing mostly on grasps that can be generated given prior knowledge of the object shape. In reality, however, those shapes (usually in the form of 3D models) are unavailable. Another significance of this work is that it demonstrated the generalizability of a learning algorithm for grasping, which is huge in terms of being deployed to an industry settings. This algorithm really does not require too much prior knowledge on the object for the grasp to be successful.

In my opinion, this is a work of exceptional quality, as they both developed theories and carried out implementations on actual hardware that can be quantitatively evaluated. The result are also quite impressive given the ability of the learning algorithm to generalize for unseen and challenging situations. One slight shortcoming of this work comes with the independence assumption made in the derivation of Bayes-like inference formulae. These assumptions over simplifies the actual model to make a

model that is tractable to compute. Nonetheless, the impressive success rate of grasping a diverse set of objects with simple 2-plate gripper does show great potentials for this method to become a foundational grasp generation algorithm that can be extended for many more gripper configurations.

5 Questions

1. Are the ground truth grasp location hand picked for the synthetic training dataset?
2. Where does the Gaussian uncertainty modeling on the image space gets propagated? Or is that reason why they picked logistic regression for estimating the probability map?