Combining Grasp Pose Detection with Object Detection

Andreas ten Pas†, Kate Saenko‡, Robert Platt†
† College of Computer and Information Science, Northeastern University
‡ Department of Computer Science, University of Massachusetts, Lowell

1 Introduction

Recently, researchers have proposed various grasp detection methods that can be used to localize grasp configurations without estimating object pose [17, 7, 12, 16, 15, 2, 3, 1, 5, 10, 8, 20, 4]. These methods take as input a noisy and partially occluded RGBD image or point cloud and produce as output pose estimates of viable grasps. The underlying idea is to treat grasp perception analogously to object detection in computer vision. Given large amounts of grasp training data, a classifier or regression system is trained to detect parts of an image or a point cloud that can be grasped. Because these methods detect grasps independently of object identity, they typically generalize grasp knowledge to new objects well. Recently, we have proposed a variation on grasp detection that detects grasps in $SE(3)$ with high accuracy. This method achieved a 93% grasp success rate in dense clutter [4] (grasps succeeded as a fraction of grasp attempts) – one of the highest reported.

A key drawback of grasp pose detection in general is that it does not immediately provide a way to grasp a specific object of interest. Grasp detectors will typically find any viable grasp – not necessarily one on the desired object. This suggests that grasp detection might be integrated with object detection. One might think about this as a mixture of two experts: a grasp detection expert that estimates the probability that a grasp candidate is a true grasp; and an object detection expert that estimates the probability that a grasp candidate is located on the object of interest. Grasps would be ranked according to the product of these two experts to rank the grasps.

2 Combined Grasp and Object Detection

In order to detect grasps on a specific object of interest, our method takes the product of an object detector and a grasp detector. This is illustrated in Figure 1. The object detector estimates the probability that a grasp candidate is located on the object of interest, and the grasp detector estimates the probability that a grasp candidate is a good grasp. Given a set of grasp candidates, we use the product of these two experts to rank the grasps.

Generating Grasp Candidates

Our system generates a large number of grasp candidates (several thousand) by searching for 6-DOF hand configurations that satisfy certain conditions related to the geometry of a grasp [20, 4]. These become candidates both for the grasp detector and for the object detector.

Object Detection

Our object detector is a fully convolutional variation of Alexnet [9] that has been trained to detect eleven different objects from our lab. We started from the Alexnet reference model and weights in Caffe [6], and fine-tuned its last inner product layer. We obtained $\sim 11k$ RGB training images using the following semi-automated procedure. First, objects were placed in front of the robot on a table in a small number (three or four) of different configurations. Then, for each object, the robot took $\sim 1k$ RGBD images from different perspectives by using an RGBD camera (an Asus Xtion Pro) mounted to the robot gripper. We synthetically augmented the training set by randomly cropping, rotating, and adding zero-mean Gaussian noise with a small standard deviation. Our network was trained using stochastic gradient descent with momentum on $\sim 500k$ images gen-
erated this way and validated on \( \sim 5k \) images. We converted our trained classifier into a detector by replacing the last three inner product layers by fully convolutional layers and transplanted the network’s learned parameters into these new layers [13], resulting in a fully convolutional version of Alexnet. Given an input image, this network creates an object detection heatmap. The intensity of a pixel in this heatmap corresponds to the estimated probability that a grasp centered at that point belongs to the object of interest (see Figure 2(d)).

**Grasp Detection** Grasp candidates are classified either as grasps or not using a variation of the Lenet architecture [11]. Our approach takes a multi-view representation [19] of the volume that would be contained within the grasp and classifies it using Lenet. The network is pretrained using \( \sim 200k \) grasp examples obtained from approximately 400 CAD models contained in the 3DNET dataset [21]. Then, the network is finetuned using data from approximately 55 objects from the BigBird dataset [18]. See [4] for more details.

**Integrated Object-Grasp Detector** Figure 2 illustrates the end-to-end process of detecting grasps on an object of interest. First, we use the object detector to create a heatmap that identifies the approximate location of the object of interest (Figure 2 (d)). Second, we generate several thousand grasp samples in the vicinity of likely object locations using data from the point cloud [20, 4]. (Figure 2 (a) shows the regions of interest in magenta; Figure 2 (b) shows the sampled grasp candidates.). Third, we prune low-scoring grasp candidates and rank the remainder based on their object detection scores (Figures 2 (c) and (d) show the locations of the detected grasps relative to an RGB image and the corresponding object detection heatmap). Finally, we select a grasp to execute using the utility function described in [4].

### 3 Experiments and Discussion

We evaluated our method on the Baxter research robot. We used Baxter’s right 7-DOF arm, with the off-the-shelf parallel-jaw gripper, constrained to a 3 to 7cm aperture. Two Asus Xtion Pro RGBD-cameras were mounted to Baxter’s waist. In each trial of the experiment, we randomly selected 6 out of the 11 objects, and placed them on a table in front of the robot with at least 1cm distance between the objects and within the workspace of the robot’s right arm. The user then entered the name of one of the objects on the table, and the robot used our method to detect and grasp that object. This continued until all objects had been grasped. The robot was allowed to attempt to identify and grasp an object at most 3 times. We measured the objection detection success rate (grasps attempted on the desired object as a fraction of total attempts) and the grasp success rate (grasps succeeded as a fraction of grasp attempts). Over the 10 trials conducted for this experiment, we obtained a 94% object detection success rate (4 object detection failures out of 75 attempts) and a 90% grasp success rate (5 grasp failures out of 54 attempts; in the remaining attempts, the inverse kinematics solver did not find a solution).

We view the results reported above as promising. Our objective is to grasp objects of interest reliably from dense clutter where standard object segmentation methods are infeasible. This paper takes the first step by proposing and demonstrating a method that does not require precise object segmentation. Our object heatmap can be viewed as an approximate semantic segmentation that suffices to distinguish on-target grasps from those off-target. As this paper shows, we obtain good results for objects that are close (1 cm), but not in dense clutter. In the future, we hope to replicate these results in dense clutter. To accomplish this, we hope to use better semantic segmentation methods [13] and to jointly detect grasps and objects in a way that makes each detection more accurate.
Acknowledgements

This work was supported in part by NSF under Grant No. IIS-1427081, NASA under Grant No. NNX16AC48A and NNX13AQ85G, and ONR under Grant No. N000141410047. Kate Saenko was supported by NSF Award IIS-1212928.

References


