1 Motivations

1. They argue that previous approaches to grasping is lacking: analytical reasoning or learning-based approach by using human labelled data.

2. The analytical reasoning approach does not work because it does not pay attention to the mass distribution of object.

3. The human labeling approach does not work because it is hard to obtain negative data points. This is because successful grasps are not unique. And human tend to only label a subset of the set of successful grasps as successful because of bias (human tend to label the handle as the grasp location for objects like cups, even though other grasp locations might also be successful). Thus, a randomly sampled grasp locations and configurations can not be assumed to be a negative data points.

4. They thus argue for the need to automatically collect training data.

5. Another key contribution of this paper is they show large-scale learning experiments on real robots are now possible. The key here is autonomous dataset collection.

2 Problem Formulation

1. They only consider planar grasp. Thus, the network needs to pick (x_S, y_S, θ_S).

2. The input into the network is an image patch, that's larger than the projection of the gripper fingertip on the image to capture context.

3. They argue against the following ways of parameterizing the output space: predicting the value for x, y, θ directly, which is bad because the target values are not unique (there is no one unique successful grasp), and CNN is not good at regression, and 2-step classification, first predicting x, y then predict θ, which is not good because the x, y and θ are not independent.

4. They thus structure the output space to be a 18-dimensional vector where each dimension indicates the likelihood of whether the center of the patch is graspable at $0, 10, \ldots, 170$ degrees.

5. During testing, given an image, they randomly sample image patch of fixed-size, and estimate the 18-dimensional vector for each image patch. They select the maximum value over all image patches and dimensions, and select the grasp to be executed as the grasp corresponding to the image patch and angle with the highest value.

3 Initial data collection method

1. The workspace is initially setup with multiple objects of varying graspability placed randomly on a dull white background. They then execute multiple random trials where each trial proceeds as:

2. Image patches containing objects are extracted from the image of the workspace by using an off-the-shell background subtraction algorithm.

3. A random image patches containing objects is selected as the region of interest of the trial.

4. Given the region of interest, the robot arm moves to 25 cm above the object.

5. A random point is uniformly sampled from the space in the region of interest. This point indicates the x, y. An angle is chosen randomly in range $(0, \pi)$ since the two fingered gripper is symmetric.

6. Given the randomly sampled grasp configuration, the robot arm executes a pick grasp on the object. The object is then raised by 20 cm and annotated as success or failure depending on the gripper's force sensor readings.

4 Multi-stage training procedure

1. After training on the randomly collected dataset, they collect additional data on both seen and novel objects.
2. The grasp configuration is no longer chosen uniformly, but is sampled from a probability matrix constructed by using the trained CNN.
3. They aggregate the new dataset and existing dataset to fine-tune the previously trained model.

5 Experiments

6 Training dataset

1. The training dataset is collected over 150 objects with varying graspability.
2. During data collection, they use a cluttered table rather than objects in isolation.
3. They collect in total 50K grasp experience interactions.

7 Test dataset

1. The test dataset consists of both seen and unseen objects. The seen objects are now displayed in different conditions.
2. The first metric they use is binary classification of whether a grasp is successful. They argue this is important for: 1. reproducibility, since obtaining error on the same test set is not possible. 2. the test set is on real robot, so if the method performs well on the test set, it should work well on real robot. Their approach reaches an accuracy of 79.5%.
3. When the trained model is used to pick grasp and the grasp is executed, the grasp success rate for novel objects are 66% and for seen objects are 73%.

8 Other tibits

Reranking grasps:

1. The arm of the Baxter is imprecise.
2. To account for the imprecision, they use neighborhood analysis.
3. Given an image patch and angle (P, θ), they sample neighborhood patches.
4. The average of the highest angle probabilities of success is assigned as the new probability of success of (P, θ).
5. The top grasps have their probabilities of successful grasps recalculated this way.
6. The grasps with the highest re-computed probability of success is executed.
7. This reranking step is to ensure that if the execution of the grasp is off by a few milimeters, it should still be successful.

9 Questions

1. At the end of section A, they execute a pick grasp given a grasp configuration, does this mean there is another 'pick grasp' routine available?