Learning Grasp Strategies with Partial Shape Information

Ashutosh Saxena, Lawson L.S. Wong and **Andrew Y. Ng**
Computer Science Department,
Stanford University, Stanford, CA 94305
{asaxena,lsw,ang}cs.stanford.edu

Abstract

We consider the problem of grasping novel objects in cluttered environments. If a full 3-d model of the scene were available, one could use the model to estimate the stability and robustness of different grasps (formalized as form/force-closure, etc); in practice, however, a robot facing a novel object will usually be able to perceive only the front (visible) faces of the object. In this paper, we propose an approach to grasping that estimates the stability of different grasps, given only noisy estimates of the shape of visible portions of an object, such as that obtained from a depth sensor. By combining this with a kinematic description of a robot arm and hand, our algorithm is able to compute a specific positioning of the robot's fingers so as to grasp an object.

We test our algorithm on two robots (with very different arms/manipulators, including one with a multi-fingered hand). We report results on the task of grasping objects of significantly different shapes and appearances than ones in the training set, both in highly cluttered and in uncluttered environments. We also apply our algorithm to the problem of unloading items from a dishwasher.

Introduction

We consider the problem of grasping novel objects, in the presence of significant amounts of clutter. A key challenge in this setting is that a full 3-d model of the scene is typically not available. Instead, a robot's depth sensors can usually estimate only the shape of the visible portions of the scene. In this paper, we propose an algorithm that, given such partial models of the scene, selects a grasp—that is, a configuration of the robot's arm and fingers—to try to pick up an object.

If a full 3-d model (including the occluded portions of a scene) were available, then methods such as form- and force-closure (Mason and Salisbury 1985; Bicchi and Kumar 2000; Pollard 2004) and other grasp quality metrics (Pelossof et al. 2004; Hsiao, Kaelbling, and Lozano-Perez 2007; Ciocarlie, Goldfeder, and Allen 2007) can be used to try to find a good grasp. However, given only the point cloud returned by stereo vision or other depth sensors, a straightforward application of these ideas is impossible, since we do not have a model of the occluded portions of the scene.

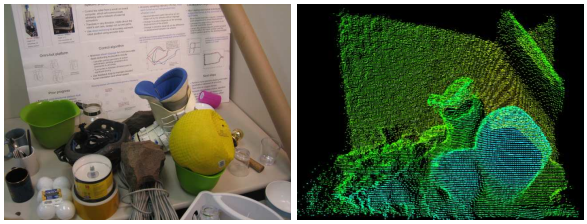

Figure 1: Image of an environment (left) and the 3-d point-cloud (right) returned by the Swissranger depth sensor.

In detail, we will consider a robot that uses a camera, together with a depth sensor, to perceive a scene. The depth sensor returns a "point cloud," corresponding to 3-d locations that it has found on the front unoccluded surfaces of the objects. (See Fig. 1.) Such point clouds are typically noisy (because of small errors in the depth estimates); but more importantly, they are also incomplete.[1]

This work builds on Saxena et al. (2006a; 2006b; 2007; 2008) which applied supervised learning to identify visual properties that indicate good grasps, given a 2-d image of the scene. However, their algorithm only chose a 3-d "grasp point"—that is, the 3-d position (and 3-d orientation; Saxena et al. 2007) of the center of the end-effector. Thus, it did not generalize well to more complex arms and hands, such as to multi-fingered hands where one has to not only choose the 3-d position (and orientation) of the hand, but also address the high dof problem of choosing the positions of all the fingers.

Our approach begins by computing a number of features of grasp quality, using both 2-d image and the 3-d point cloud features. For example, the 3-d data is used to compute a number of grasp quality metrics, such as the degree to which the fingers are exerting forces normal to the surfaces of the object, and the degree to which they enclose the object. Using such features, we then apply a supervised learning algorithm to estimate the degree to which different configurations of the full arm and fingers reflect good grasps.

We test our algorithm on two robots, on a variety of objects of shapes very different from ones in the training set, including a ski boot, a coil of wire, a game controller, and

[1] For example, standard stereo vision fails to return depth values for textureless portions of the object, thus its point clouds are typically very sparse. Further, the Swissranger gives few points only because of its low spatial resolution of 144×176.

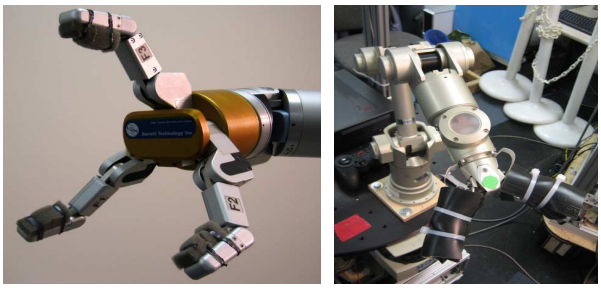

Figure 2: (Left) Barrett 3-fingered hand. (Right) Katana parallel plate gripper.

others. Even when the objects are placed amidst significant clutter, our algorithm often selects successful grasps.

Related Work

Space constraints prevent us from doing full justice to prior work, and here we will focus on prior work that performed real-world grasping experiments. For a more detailed treatment of related work, see, e.g., (Mason and Salisbury 1985; Bicchi and Kumar 2000; Saxena, Driemeyer, and Ng 2008).

In prior work that used vision for real-world grasping experiments, most were limited to grasping 2-d planar objects. For a uniformly colored planar object lying on a uniformly colored table top, one can find the 2-d contour of the object quite reliably. Using local visual features (based on the 2-d contour) and other properties such as form- and force-closure, (Coelho, Piater, and Grupen 2001; Chinellato et al. 2003; Bowers and Lumia 2003; Morales et al. 2004) computed the 2-d locations at which to place (two or three) fingertips to grasp the object. In more general settings (i.e., non-planar grasps), Edsinger and Kemp (2006) grasped cylindrical objects using a power grasp by using visual servoing and Platt et al. (2006) used schema structured learning for grasping simple objects (spherical and cylindrical) using power grasps; however, this does not apply to grasping general shapes (e.g., a cup by its handle) or to grasping in cluttered environments.

Description of Robots

Our experiments were performed on two robots. STAIR 1 uses a 5-dof harmonic arm (Katana, by Neuronics) with a parallel plate gripper, and STAIR 2 uses a 7-dof arm (WAM, by Barrett Technologies) with a three-fingered 4-dof hand.

The robot's vision system consists of a stereo camera (Bumblebee2, by Point Grey Research), and a SwissRanger camera (Swissranger 2005). The SwissRanger camera is a time-of-flight depth sensor that returns a 144×176 array of depth estimates, spanning a $47.5^\circ \times 39.6^\circ$ field of view, with a range of about 8m. Using an infrared structured light source, each pixel in the camera independently measures the arrival time of the light reflected back by the objects. Its depth estimates are typically accurate to about 2cm. However, it also suffers from systematic errors, in that it tends not to return depth estimates for dark objects, nor for surfaces lying at a large angle relative to the camera's image plane. (See Fig. 1 for a sample 3-d scan.)

Grasping Strategy

There are many different properties that make certain grasps preferable to others. Examples of such properties include form- and force-closure (to minimize slippage), sufficient contact with the object, distance to obstacles (to increase robustness of the grasp), and distance between the center of the object and the grasping point (to increase stability). In real world grasping, however, such properties are difficult to compute exactly, because of the quality of sensor data. Our algorithm will first compute a variety of features that attempt to capture some of these properties. Using these features, we then apply supervised learning to predict whether or not a given arm/finger configuration reflects a good grasp.

Definition of grasp: We will infer the full goal configuration of the arm/fingers that is required to grasp an object. For example, STAIR 1 uses an arm with 5 joints and a parallel plate gripper (with one degree of freedom); for this robot, the configuration is given by $\alpha \in \mathbb{R}^6$. The second robot STAIR 2 uses an arm with 7 joints, equipped with a three-fingered hand that has 4 joints (Fig. 2); for this robot, our algorithm infers a configuration $\alpha \in \mathbb{R}^{11}$. We will informally refer to this goal configuration as a "grasp."

We will then use a motion planner algorithm (Saha and Isto 2006) to plan a path (one that avoids obstacles) from the initial configuration to this goal configuration.

Probabilistic Model

We will use both the point-cloud R and the image I taken of a scene to infer a goal configuration α of the arm/fingers.

Saxena et al. (2006a) classified each 2-d point in a given image as a 1 (candidate grasp) or 0. For example, for an image of a mug, it would try to classify the handle and rim as candidate grasping points. In our approach, we use a similar classifier that computes a set of image features and predicts the probability $P(y = 1|\alpha, I) \in [0, 1]$ of each point in the image being a candidate grasping point. However, a remaining difficulty is that this algorithm does not take into account the arm/finger kinematics; thus, many of the 2-d points it selects are physically impossible for the robot to reach.

To address this problem, we use a second classifier that, given a configuration α and a point-cloud R, predicts the probability $P(y|\alpha, R)$ that the grasp will succeed. This classifier will compute features that capture a variety of properties that are indicative of grasp quality. Our model will combine these two classifiers to estimate the probability $P(y|\alpha, R, I)$ of a grasp α succeeding. Let $y \in \{0, 1\}$ indicate whether $\{\alpha, R, I\}$ is a good grasp. We then have:

$$P(y|\alpha, R, I) \propto P(R, I|y, \alpha)P(y, \alpha) \quad (1)$$

We assume conditional independence of R and I, and uniform priors $P(\alpha)$, $P(I)$, $P(R)$ and $P(y)$. Hence,

$$\begin{aligned} P(y|\alpha, R, I) &\propto P(R|y, \alpha)P(I|y, \alpha)P(y, \alpha) \\ &\propto P(y|\alpha, R)P(y|\alpha, I) \end{aligned} \quad (2)$$

Here, the $P(y|\alpha, I)$ is the 2-d image classifier term similar to the one in Saxena et al. (2006a). We also use $P(y|\alpha, R; \theta) = 1/\exp\left(1 + \psi(R, \alpha)^T\theta\right)$, where $\psi(R, \alpha)$ are the features discussed in the next section.

Inference: From Eq. 2, the inference problem is:

$$\begin{aligned} \alpha^* &= \arg\max_\alpha \log P(y = 1|\alpha, R, I) \\ &= \arg\max_\alpha \log P(y = 1|\alpha, R) + \log P(y = 1|\alpha, I) \end{aligned}$$

Now, we note that a configuration α has a very small chance of being the optimal configuration if either one of the two terms is very small. Thus, for efficiency, we implemented an image-based classifier $P(y|I, \alpha)$ that returns only a small set of 3-d points with a high value. We use this to restrict the set of configurations α to only those in which the hand-center lies on one of the 3-d location output by the image-based classifier. Given one such 3-d location, finding a full configuration α for it now requires solving only an $n-3$ dimensional problem (where $n = 6$ or 11, depending on the arm). Further, we found that it was sufficient to consider only a few locations of the fingers, which further reduces the search space; e.g., in the goal configuration, the gap between the finger and the object is unlikely to be larger than a certain value. By sampling randomly from this space of "likely" configurations (similar to sampling in PRMs) and evaluating the grasp quality only of these samples, we obtain an inference algorithm that is computationally tractable and that also typically obtains good grasps.

Features

Below, we describe the features that make up the feature vector $\psi(R, \alpha)$ used to estimate a good grasp. The same features were used for both robots.

Presence / Contact: For a given finger configuration, some part of an object should be inside the volume enclosed by the fingers. Intuitively, more enclosed points indicate that the grasp contains larger parts of the object, which generally decreases the difficulty of grasping it (less likely to miss). For example, for a coil of wire, it is better to grasp a bundle rather than a single wire. To robustly capture this, we calculate a number of features—the number of points contained in spheres of different sizes located at the hand's center, and also the number of points located inside the volume enclosed within the finger-tips and the palm of the hand.

Symmetry / Center of Mass: Even if many points are enclosed by the hand, their distribution is also important. For example, a stick should be grasped at the middle instead of at the tip, as slippage might occur in the latter case due to a greater torque induced by gravity. To capture this property, we calculate a number of features based on the distribution of points around the hand's center along an axis perpendicular to the line joining the fingers. To ensure grasp stability, an even distribution (1:1 ratio) of points on both sides of the axis is desirable. More formally, if there are N points on one side and N' on the other side, then our feature would be $|N - N'|/(N + N')$. Again, to increase robustness, we use several counting methods, such as counting all the points, and counting only those points not enclosed by the hand.

Local Planarity / Force-Closure: One needs to ensure a few properties to avoid slippage, such as a force closure on the object (e.g., to pick up a long tube, a grasp that lines up the fingers along the major axis of the tube would likely fail). Further, a grasp in which the finger direction (i.e., direction in which the finger closes) is perpendicular to the local tangent plane is more desirable, because it is more likely to lie within the friction cone, and hence is less likely to slip. For example, when grasping a plate, it is more desirable that the fingers close in a direction perpendicular to the plate surface.

Figure 3: Snapshots of our robot grasping novel objects of various sizes/shapes.

To capture such properties, we start with calculating the principal directions of a 3-d point cloud centered at the point in question. This gives three orthonormal component directions u_i, with u_1 being the component with largest variance, followed by u_2 and u_3. (Let σ_i be the corresponding variances.) For a point on the rim of a circular plate, u_1 and u_2 would lie in the plane in which the plate lies, with u_1 usually tangent to the edge, and u_2 facing inwards. Ideally, the finger direction should be orthogonal to large variance directions and parallel to the small variance ones. For f_j as the finger direction ($j = 1$ for parallel gripper, and $j = 1, 2$ for three-fingered hand), we would calculate the following features: (a) Directional similarity, $s_{ij} = |u_i \cdot f_j|$, and (b) Difference from ideal, $(\frac{\sigma_1 - \sigma_i}{\sigma_1 - \sigma_3} - s_{ij})^2$.

Experiments

We performed three sets of extensive experiments: grasping with our three-fingered 7-dof arm in uncluttered as well as cluttered environments, and on our 5-dof arm with a parallel plate gripper for unloading items in a cluttered dishwasher.

Grasping single novel objects

We considered several objects from 13 novel object classes in a total of 150 experiments. These object classes varied greatly in shape, size, and appearance, and are very different from the plates, bowls, and rectangular blocks used in the training set. During the experiments, objects were placed at a random location in front of our robot. Table 1 shows the results: "Prediction" refers to the percentage of cases the final grasp and plan were good, and "Grasp success" is the percentage of cases in which the robot predicted a good grasp and actually picked up the object as well (i.e., if the object slipped and was not picked up, then it counted as failure).

Using the same robot, Saxena et al. (2007) considered power grasps only and required that objects be neatly placed on a "rack." However, we consider the significantly harder task of grasping randomly placed objects in any orientation. Further, many objects such as ski boots, helmets, etc. require more intricate grasps (such as inserting a finger in the ski boot because it is too big to hold as a power grasp).

For each object class, we performed 10-20 trials, with

Table 1: STAIR 2. Grasping single objects. (150 trials.)

Object Class	Size	Prediction	Grasp success
Ball	small	80%	80%
Apple	small	90%	80%
Gamepad	small	85%	80%
CD container	small	70%	60%
Helmet	med	100%	100%
Ski boot	med	80%	80%
Plate bundle	med	100%	80%
Box	med	90%	90%
Robot arm link	med	90%	80%
Cardboard tube	large	70%	65%
Foam	large	60%	60%
Styrofoam	large	80%	70%
Coil of wire	large	70%	70%

different instances of each object for each class (e.g., different plates for "plates" class). The average "Prediction" accuracy was 81.3%; and the actual grasp success rate was 76%. The success rate was different depending on the size of the object.[2] Both prediction and actual grasping were best for medium sizes, with success rates of 92% and 86% respectively. Even though handicapped with a significantly harder experimental trial (i.e., objects not neatly stacked, but thrown in random places and a larger variation of object sizes/shapes considered) as compared to Saxena et al., our algorithm surpasses their success rate by 6%.

Grasping in cluttered scenarios

In this case, in addition to the difficulty in perception, manipulation and planning became significantly harder in that the arm had to avoid all other objects while reaching for the predicted grasp; this significantly reduced the number of feasible candidates and increased the difficulty of the task.

In each trial, more than five objects were placed in random locations (even where objects touched each other, see some examples in Fig. 3). Using only the 2-d image-based method of Saxena et al. (with PRM motion planning), but not our algorithm that considers finding all arm/finger joints from partial 3-d data, success was below 5%. In a total of 40 experiments, our success rate was 75% (see Table 2). We believe our robot is the first one to be able to automatically grasp objects, of types never seen before, placed randomly in such heavily cluttered environments.

Table 2: STAIR 2. Grasping in cluttered scenes. (40 trials.)

Environment	Object	Prediction	Grasp Success
Terrain	Tube	87.5%	75%
Terrain	Rock	100%	75%
Kitchen	Plate	87.5%	75%
Kitchen	Bowl	75%	75%

Table 3: STAIR 1. Dishwasher unloading results. (50 trials.)

Object Class	Prediction Good	Actual Success
Plate	100%	85%
Bowl	80%	75%
Mug	80%	60%

[2]We defined an object to be small if it could be enclosed within the robot hand, medium if it was approximately 1.5-3 times the size of the hand, and large otherwise (some objects were even 4ft long).

Applying algorithm on a different robotic platform

One of the properties of the algorithm is that it is agnostic to particular robot platforms. We applied our algorithm on STAIR 1, and attempted to grasp kitchen items from a cluttered dishwasher with the presence of 3 or more objects placed randomly (see Table 3). In a total of 50 experiments, even with more clutter than in Saxena et al. (2006a; 2007) (where objects were placed neatly in dishwasher), our algorithm gave comparable results. Our algorithm is therefore generalizable to different robots.

We have made our grasping movies available at:
http://stair.stanford.edu/multimedia.php

References

Bicchi, A., and Kumar, V. 2000. Robotic grasping and contact: a review. In *ICRA*.

Bowers, D., and Lumia, R. 2003. Manipulation of unmodeled objects using intelligent grasping schemes. *IEEE Trans Fuzzy Systems* 11(3).

Chinellato, E.; Fisher, R. B.; Morales, A.; and del Pobil, A. P. 2003. Ranking planar grasp configurations for a three-finger hand. In *ICRA*.

Ciocarlie, M.; Goldfeder, C.; and Allen, P. 2007. Dimensionality reduction for hand-independent dexterous robotic grasping. In *IROS*.

Coelho, J.; Piater, J.; and Grupen, R. 2001. Developing haptic and visual perceptual categories for reaching and grasping with a humanoid robot. *Robotics and Autonomous Systems* 37:195–218.

Edsinger, A., and Kemp, C. 2006. Manipulation in human environments. In *Int'l Conf Humanoid Robotics*.

Hsiao, K.; Kaelbling, L.; and Lozano-Perez, T. 2007. Grasping POMDPs. In *ICRA*.

Mason, M., and Salisbury, J. 1985. *Robot Hands and the Mechanics of Manipulation*. MIT Press, Cambridge, MA.

Morales, A.; Chinellato, E.; Sanz, P. J.; del Pobil, A. P.; and Fagg, A. H. 2004. Learning to predict grasp reliability for a multifinger robot hand by using visual features. In *Int'l Conf AI Soft Comp*.

Pelossof, R.; Miller, A.; Allen, P.; and Jebara, T. 2004. An svm learning approach to robotic grasping. In *ICRA*.

Platt; Grupen; and Fagg. 2006. Improving grasp skills using schema structured learning. In *ICDL*.

Pollard, N. S. 2004. Closure and quality equivalence for efficient synthesis of grasps from examples. *IJRR* 23(6).

Saha, M., and Isto, P. 2006. Motion planning for robotic manipulation of deformable linear objects. In *ICRA*.

Saxena, A.; Driemeyer, J.; Kearns, J.; and Ng, A. Y. 2006a. Robotic grasping of novel objects. In *NIPS*.

Saxena, A.; Driemeyer, J.; Kearns, J.; Osondu, C.; and Ng, A. Y. 2006b. Learning to grasp novel objects using vision. In *ISER*.

Saxena, A.; Wong, L.; Quigley, M.; and Ng, A. Y. 2007. A vision-based system for grasping novel objects in cluttered environments. In *ISRR*.

Saxena, A.; Driemeyer, J.; and Ng, A. Y. 2007. Learning 3-d object orientation from images. In *NIPS workshop on Robotic Challenges for Machine Learning*.

Saxena, A.; Driemeyer, J.; and Ng, A. Y. 2008. Robotic grasping of novel objects using vision. *IJRR* 27(2):157–173.

Swissranger. 2005. Mesa imaging. In *http://www.mesa-imaging.ch*.