Using Near-Field Stereo Vision for Robotic Grasping in Cluttered Environments

Adam Leeper, Kaijen Hsiao, Eric Chu, and J. Kenneth Salisbury

Abstract. Robotic grasping in unstructured environments requires the ability to adjust and recover when a pre-planned grasp faces imminent failure. Even for a single object, modeling uncertainties due to occluded surfaces, sensor noise and calibration errors can cause grasp failure; cluttered environments exacerbate the problem. In this work, we propose a simple but robust approach to both pre-touch grasp adjustment and grasp planning for unknown objects in clutter, using a small-baseline stereo camera attached to the gripper of the robot. By employing a 3D sensor from the perspective of the gripper we gain information about the object and nearby obstacles immediately prior to grasping that is not available during head-sensor-based grasp planning. We use a feature-based cost function on local 3D data to evaluate the feasibility of a proposed grasp. In cases where only minor adjustments are needed, our algorithm uses gradient descent on a cost function based on local features to find optimal grasps near the original grasp. In cases where no suitable grasp is found, the robot can search for a significantly different grasp pose rather than blindly attempting a doomed grasp. We present experimental results to validate our approach by grasping a wide range of unknown objects in cluttered scenes. Our results show that reactive pre-touch adjustment can correct for a fair amount of uncertainty in the measured position and shape of the objects, or the presence of nearby obstacles.

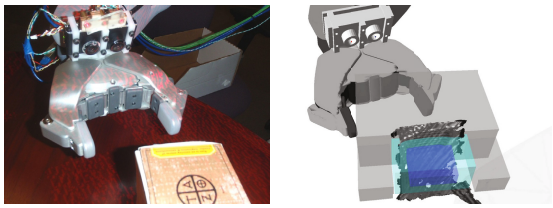
Adam Leeper · Eric Chu · J. Kenneth Salisbury
Stanford University, Stanford, CA
e-mail: aleeper@stanford.edu, echu508@stanford.edu,
jks@robotics.stanford.edu

Kaijen Hsiao
Willow Garage, Menlo Park, CA
e-mail: hsiao@willowgarage.com

O. Khatib, V. Kumar, and G. Sukhatme (eds.), *Experimental Robotics*,
Springer Tracts in Advanced Robotics 79,
DOI: 10.1007/978-3-642-28572-1_18, © Springer-Verlag Berlin Heidelberg 2014

1 Introduction

The work described herein is part of a project to enhance the reliability of robotic grasping in unstructured and cluttered environments. As robots move out of the factory and into human workspaces, new algorithms and sensors are needed to overcome uncertainties in sensing due to noise, calibration, and occlusion. As part of this effort, we propose a novel, small stereo camera attached to the gripper of a robot to provide local 3D sensing, and an algorithm for adjusting or re-planning grasps as needed in the vicinity of a desired grasp pose (Fig. 1).



(a) (b)

Fig. 1 From a pregrasp pose near the object, data from a gripper-mounted stereo camera is used to adjust the final grasp pose to more closely align with the target object, in this case, a tea box. In (b) the gripper's adjusted grasp pose is displayed using grey boxes.

Grasping is a sensitive task, and errors in sensing and kinematics in unstructured environments lead to particular persistent failure cases in current, state-of-the-art methods for grasping objects, such as [1, 2, 3, 4, 5, 6, 7]. The robot platforms used in the aforementioned systems employ laser scanners and stereo cameras for mid- or long-range sensing (1-5 meters), mounted on the body or head of a robot. Whether due to practical design limitations (space, wiring) or merely the familiar paradigm that humans "see" from the head, the result is a large kinematic chain between the sensor and the robot end-effector. A common failure mode for these systems results from the accumulation of errors in the sensed object position and errors in the forward kinematics of the robot arm.

Another limitation of sensors mounted to the base of the robot arises from having an inherently limited field of view, as well as occlusions caused by the environment or the robot itself. Common everyday scenarios such as retrieving an object inside a container or on a high shelf can be impossible to achieve if the object is obscured from the robot's primary sensors. Another complication is the motion of objects; the robot has little hope of achieving a grasp if it blocks its own sensors when attempting to reach out for an object and is unable to update the object model while the object is moving significantly.

We propose that these challenges can be mitigated using a stereo camera pair on a robotic hand. Even if the object is in plain view of a robot's primary sensors, a gripper stereo pair can add data to the 3D point cloud from a different perspective, reducing modeling ambiguities. One common example where grasp planning using

just a single view from the robot's head can run into trouble is shown in Figure 2; using gripper-mounted stereo cameras allows one to prevent obvious mistakes of this sort. More importantly, this sensor can expand current capabilites in robotic manipulation by allowing the robot hand to explore areas that are out of sight of the other sensors on the robot.

Even at close range, stereo vision has some advantages over other pre-touch sensing methods. Unlike optical infrared [8] and electric-field [9] sensing, stereo vision is more robust to varying target material properties. Namely, the calibration of stereoscopic triangulation is not dependent on capacitance, reflectivity, or other material properties, as long as the target surface has some visible features for the stereo block matching algorithm. Stereo vision can be made more robust to lighting conditions and object texture by carrying a textured light source; inspired by the texture projector used in [10], we implement a simple laser diffraction pattern projected from the gripper to aid the gripper stereo sensor. While this is not an optimal texture due to its regular pattern of dots, it is a simple, compact solution.

(a) (b)

Fig. 2 (a) A view from the robot body (green cone) cannot see the protruding surface of the object. In a fairly typical planned grasp on the partial view, the gripper moves as the arrows indicate. Viewing the object from the gripper's perspective (orange cone) gives more complete data in the direction of interest. (b) An example is pictured with a large detergent bottle; the view from the front leaves significant ambiguity about the depth of the bottle, whereas the view from the gripper shows that the finger tips will not clear the object.

2 Related Work

Much of the work in robotic grasp planning searches for grasp points on an object with a known 3D model [11, 5]. Toward grasping unknown objects, systems such as [1, 2, 6] search for stable grasp points in scenes using only available 2D and 3D sensor data. These systems assume that what the robot senses and what its gripper will encounter share a measure of similarity, which is not always the case. Hence we turn to look at works that attempt to recover from sensing and planning errors.

Many strategies have been explored for reacting to imperfect grasp actions. There is a great deal of work on grasping using visual servoing, using both monocular and stereo cameras mounted on the head and on the end-effector. For cases where the object is known or at least can be segmented from the background and visually

tracked, visual servoing with end-effector cameras can ensure the robust execution of grasps that are known to be good if executed correctly, by incrementally correcting the position of the object relative to the gripper during the grasp approach. An excellent survey on the subject can be found in [12].

For cases where the object and gripper are not easily tracked visually, simple methods for grasp adjustment using a variety of contact and force sensors have been explored in systems such as [13, 14, 15, 16, 7, 17, 3, 18]. Such methods can be very effective at recovering from some grasp errors, and even in this work we use the tactile-sensor-based reactive grasping behaviors described in [18] as a final grasp adjustment on top of the methods described. However, for many common objects, espcially those that are light, fragile, or top-heavy, premature contact with a robot finger can cause total failure due to slippage or tipping even in cases where the grasp pose is only off by a few millimeters. Pre-touch sensing of some sort is required to successfully recover from grasping errors in such cases.

Work in pre-touch sensing has shown that various short-range sensors are useful to assure good grasps before disturbing objects. Hsiao et. al [8] developed small IR emitter-receiver pairs to determine the pose of surfaces in close proximity to a robotic finger to do online pre-touch finger adjustments starting from a given grasp pose. Such optical sensors feature a very close sensing range ($<$1 cm to 4 cm) but difficulties in calibration and inherent limitations of IR on certain materials pose some challenges.

In [9], Mayton et. al use electric field sensors on the fingers of a Barrett hand to gravitate toward objects and also adjust the fingers before making contact. This sensor has a larger range than many optical solutions, but its application is limited to objects with certain dielectric properties.

Although stereo cameras and laser rangefinders have previously been mounted on end-effectors [19, 12], the purpose of such sensors has thus far been only to either obtain 3D positions for visual features of known objects, or to create models of unknown objects. We propose to use the entire stereo point cloud from an end-effector-mounted stereo camera to adjust and even re-plan grasps entirely for unknown objects, by reasoning about previously-occluded geometries, potential collisions, and the likely quality of the grasp.

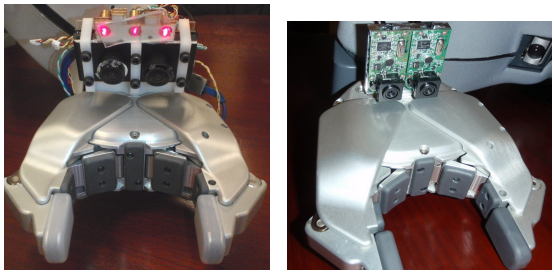
Furthermore, our stereo camera is more compact than many previous end-effector-mounted stereo cameras, and can be mounted on the end effector without losing much of its ability to be useful in constrained spaces. There are even smaller stereo cameras in the area of miniature stereoscopic vision, where most efforts have been focused at medical applications, such as stereo cameras with millimeter optical baselines mounted on the end of endoscopes [20, 21]. However, in these applications the goal is merely allowing the surgeon to see a tiny work area; such cameras are not suitable for our purposes.

3 Hardware Platform

The hardware used for the experiments in this paper is the PR2 personal robot, which has two 7-DOF arms with parallel-jaw grippers, two stereo cameras in the head, and an omni-directional base. For these experiments the PR2 was modified with a custom stereo camera attached to the gripper of one arm. At present, the smallest commercially available stereo camera has a stereo baseline of 6 cm, which is too large to fit on a gripper without losing a considerable amount of maneuverability, and has a minimum object sensing distance of at least 20 cm, which is unsuitable for near-field sensing.

3.1 Gripper Stereo

We created two custom gripper-mounted stereo sensors (Fig. 3). The first, emphasizing low cost and simplicity, is made from standard webcameras (Logitech C905). By removing the plastic casing, the stereo pair can be arranged with an optical baseline of 25mm. The stock lenses have an adjustable focus ideal for near-field sensing, and the 65 degree field of view of each camera is well suited to the task, providing a minimum sensing distance of about 8 cm. The total cost for our webcam sensor is under $150, not including the cost of mounting hardware. With a standard USB interface for each camera, stereo processing can be done easily in software.

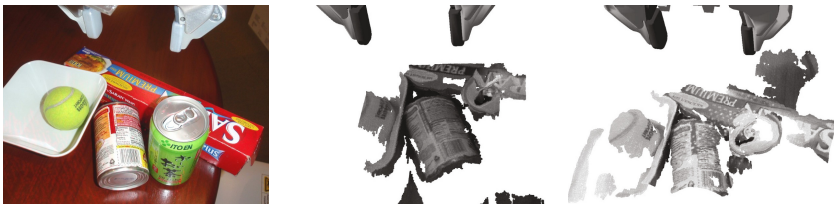


(a) (b)

Fig. 3 (a) The PR2 gripper used in these experiments is equipped with a custom stereo camera featuring synchronized, globally shuttered imagers. The textured laser projector is mounted to the top of the camera assembly. (b) An early prototype using low-cost logitech webcameras; we note this is a viable solution if more advanced hardware is not available.

While initial tests demonstrated that a useful result could be achieved with the low-cost webcams, there are significant limitations to having a rolling shutter (resulting in image skew when there is relative motion between scene and camera) and unsynchronized camera pairs. Thus, the stereo camera used for the experiments in this paper is made from the same hardware as the PR2's other custom cameras (model WGE100), giving us the advantages of a global shutter on each camera, sub-millisecond synchronization between the cameras, and synchronization with

the texture projector on the PR2. These features all result in improved stereo point clouds. Our gripper stereo sensor uses wide-angle lenses (2.5 mm focus) with a field of view of approximately 90 degrees. The stereo baseline is 30 mm, the minimum object distance is 10 cm, and images are collected at VGA (640 x 480) resolution.



(a) (b) (c)

Fig. 4 A cluttered table scene (a) is viewed from the gripper using the web cameras (b) and the custom WGE100 cameras (c). The stereo point clouds are of comparable quality on this static scene with no texture projected.

Our grasp sequence also uses tactile-based reactive grasp adjustment as described in [18] for final grasp correction. To this end, the fingertips of the gripper are equipped with a capacitive sensor consisting of 22 individual cells: a 5x3 array on the parallel gripping surface itself, 2 sensor elements on the tips of the fingertips, 2 elements on each side of the fingertip, and one on the back. These capacitive sensors measure the normal pressure applied in each sensed region.

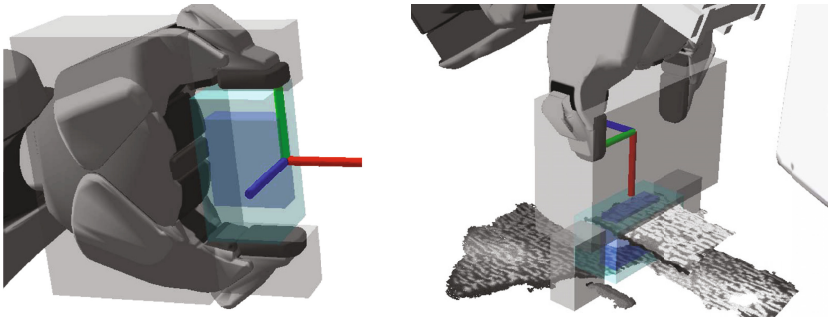
4 Grasp Adjustment

We begin with a brief overview of our strategy: given an intended grasp pose and a point cloud, we wish to determine a list of viable poses near the intended pose that we believe are likely to result in good, collision-free grasps. To that end, we search for poses that optimize a cost function based on point cloud features.

Let us assign a right-handed world frame W with axes W_x and W_y parallel to the ground, and W_z vertical. The gripper has a local frame G, and is modeled using a simplified box model, defined in Fig. 5.

In the context of everyday grasping scenarios, an object in a stable pose on a table has only three degrees of freedom: translation in W_x and W_y, and orientation about W_z. Thus, we design our algorithm to adjust positioning of the gripper in these three dimensions, as well as vertical position along W_z. Note that the effect of adjusting orientation about W_z is that in a top grasp the algorithm appears to roll the gripper about G_x, while in a side grasp it appears to yaw the gripper from side to side about G_z.

Since our strategy is to adjust grasps, we assume the robot has already moved to a reasonable pregrasp pose near an object or group of objects. In the ROS grasp pipeline [4] this pose is typically 10 cm back from the intended grasp pose along G_x;



(a) (b)

Fig. 5 The gripper is modeled using a collection of simple boxes. The gripper's local frame, G, is defined with G_x forward (red), G_y along the axis between the finger tips (green), and G_z following by the right-hand-rule. The gray boxes are used to check for point cloud collisions with the gripper; the light green "space" box between the fingers is used to select which points contribute to the symmetry and centroid features; and the small, dark blue "target" box (nested inside the "space" box) is used to calculate the contribution from normals. In (b), the gripper model is displayed over the intended grasp pose.

thus, when we search for an adjusted grasp, we assume that the intended grasp pose is approximately 10 cm in front of the current gripper pose, although our search algorithm can compensate if the object turns out to be closer or farther away.

From a pregrasp pose in the vicinity of some objects, the robot can request a search for grasps on a wide area. This "global" search uses starting poses on a m-by-m grid of cell size c centered at the intended grasp pose (we used $m = 5$ and $c = 0.03$ m). On this grid, rotational perturbations of $\phi = \pm 45$ degrees about W_z are used as well. On the other hand, if the robot is more confident in its intended grasp pose, it can request a "local" optimization of its grasp. The algorithm attempts to optimize three grasp poses: the intended grasp pose and two poses offset by a 45 degree rotation about W_z. This results in a pose that is better aligned and centered over an object, but avoids having the robot decide to go grab another object within view of the gripper stereo.

With a list of candidate poses, we wish to optimize the poses by performing gradient descent using a cost function described in the next section. Since it can be computationally expensive to perform a full gradient descent on every candidate pose, our algorithm uses a brief, broad search before doing a deep search on the most promising candidates. The algorithm first eliminates poses that have no points within any boxes of the gripper model. The remaining poses are each stepped through n steps of gradient descent with fixed translational and rotational step sizes of t and r, respectively. Finally, the p poses with the best cost so far are optimized using a deeper gradient descent. (In these experiments: $n = 2$, $t = 0.002$m, $r = 0.05$rad, and $p \leq 5$.)

Because even doing a single gradient descent can be computationally expensive when searching in all possible directions at once, we further simplify the gradient

descent search by taking steps in each dimension separately. For each dimension in turn, we evaluate the poses that are 2 steps in each direction before selecting the best pose found along that dimension. Although this method may find slightly less optimal local minima than more thorough gradient descent methods, the fact that we are also searching over candidate poses on a broader search grid means that we need not search very broadly in our local gradient descent to find a good solution. We thus also limit the number of gradient steps in each local descent to a maximum of 25, to prevent the pose from drifting too far from its starting pose.

The adjustment algorithm is summarized concisely in Algorithm 1.

```
Input: point cloud, Cloud; intended pose, PoseIn; search extent, GlobalSearch.
Output: List of viable grasp poses, PosesOut.
if GlobalSearch then
    Candidates = Grid of starting poses around PoseIn with rotational offsets;
end
else
    Candidates = PoseIn + two rotational offsets;
end
foreach Pose in Candidates do
    if gripper model is in contact with any points in Cloud then
        Descend gradient for two steps;
        Add Pose to PriorityQueue;
    end
end
for best n Poses in PriorityQueue do
    Descend gradient until minimum;
    Add Pose to PosesOut;
end
```

Algorithm 1. The pose adjustment algorithm.

5 Grasp Evaluation

When evaluating a candidate grasp pose, our algorithm takes the following as input: a point cloud, a model of the gripper, and the height of the table surface. The point cloud can come from any number of combined stereo camera point clouds, the gripper model that we use is detailed in Fig. 5(a), and the height of the table surface can typically be found from the gripper stereo point cloud, if not known a priori; in our experiments we assume a priori knowledge of the table height, since the ROS grasping pipeline does table detection before trying to grasp objects.

We compute a weighted sum of a set of features based on these inputs to produce a numerical cost for each candidate grasp pose. The feature set has been chosen as follows:

1. *point_cost*: This feature attempts to ensure that there is part of an object well within the gripper; its value is the number of points inside the "target" box of the gripper model. The presence of many points inside this box indicates a likely graspable part of an object that is seen fairly well by the camera. The sides of the target box are recessed in from both front and side edges of the fingertips, so that we do not try to grasp just the edges of objects, and the box is also small to avoid the tendency of the gripper to rotate to be oriented diagonally with respect to thin, locally-box-like objects in order to capture more points. The formula is $point\_cost = (point\_count)^{1/4}$, so that the gradient descent strongly prefers having some points within the gripper, but does not value a few extra points as much when there are already many.
2. *normals*: This feature attempts to align the object in the gripper and avoid cases where the robot tries to grab the corner of a box. It does so by rewarding point normals that are parallel to G_x, G_y or G_z but not normals that make a 45 degree angle with any of these directions; its value is computed as $normals = \Sigma[1 - (cos(2\theta_i - \frac{\pi}{2}))^2]$ where the sum is over all points within the "target" box, and θ_i is the angle between G_y and the point normal.
3. *symmetry*: This feature attempts to locally align box-like objects with the gripper. Since the gripper stereo cannot see faces parallel to the finger tip pads, if we are not using a separate viewpoint in conjunction with the gripper point cloud normals will not help determine alignment about G_x. If an object is locally box-like and centered, with edges aligned with the gripper, then it will have an even distribution of points between the finger tips in the G_{yz} plane. We thus project any points within the "space" box to this plane and count how many points are in each quadrant, taking note of the minimum and maximum. The symmetry feature is calculated using $symmetry = \frac{min\_count}{max\_count} f(p)$, where $f(p) = min(point\_count, 20)$.
4. *centroid_error*: This feature encourages both centering the object within the hand and grasping deeper on the object. Its value is the euclidean distance in meters from the front-center of the gripper palm to the centroid of all points within the gripper model's "space" box, where the centroid is computed by $centroid = \frac{1}{n}\Sigma_{i=1}^{n} p_i$. By using the space box rather than the target box, we also penalize grasping near the edge of an object.
5. *collision_count*: This feature discourages collisions between the gripper and the environment; the value is the integer number of points within the palm or finger boxes of the gripper model.
6. *translation_cost*: This feature discourages gripper poses that are significantly translated with respect to the initial pose, to avoid grasping at the edge of the sensor's field of view; it is computed as the square of the euclidean distance in meters between the candidate pose and the initial pose.
7. *rotation_cost*: This feature discourages gripper poses that are significantly rotated with respect to the initial pose, and is computed as $rotation\_cost = (\frac{\phi}{\pi/2})^{10}$, where ϕ is the angle between the candidate and initial poses. The power of 10 creates a low cost for most rotations, but a quick increase near $\pm\frac{\pi}{2}$.

Our feature weightings were chosen empirically. The following weights were used: [-0.01, -0.3, -0.1, 10, 1, 1, 1]. Note that the best grasps will have a small (or negative) cost.

6 Experiments

Although it can be easy and effective to combine point clouds from the gripper stereo with point clouds from head stereo cameras, we wished to show off the capabilities of just the gripper stereo. Thus, in the experiments described here, we used only the point cloud from the gripper stereo. Also, to show the ability of our algorithms to correct for a large amount of grasp uncertainty/error, and to avoid the possibility that grasp planning based on the head stereo point clouds might select pregrasp poses that do not require much correction, we provided only two fixed pregrasp poses for the gripper in all of the experiments: one from the top, with the fingertips 18 cm above the table, and one from the side, with the fingertips 6 cm above the table.

In each grasping experiment, the robot makes two calls to the grasp adjustment service. The first call requests a "global" search, allowing the robot to find a grasp point that it can most likely grasp. It then moves the gripper to a pregrasp pose for this new grasp point, giving it a better view of the target area. It then makes a request for a "local" grasp adjustment before executing the grasp.

We enumerate our experiments for later reference:

1. We positioned the gripper in the top pregrasp position, with the finger tips 18 cm above the table surface. A simple cylindrical object (a 305 g can of soup, 10 cm tall by 6.5 cm diameter) was placed on the table at various positions, and the robot was asked to execute a grasp. This experiment turned out to be largely a test of the field-of-view of the sensor; if the sensor could see any points on the soup can it would move over toward it and find a good grasp on the second call to the adjustment service.
2. We tested the algorithm's ability to discriminate grasp orientation by grasping a large tea box in a random orientation. For the top-pregrasp position, the box was laid down with its centroid within ± 5 cm of the gripper's centerline and in a random orientation (0 to 360 degrees). For the side-pregrasp position, the box was placed upright with its narrower side facing toward the gripper within ± 30 degrees. Ten trials were performed for each pregrasp pose. To make things more challenging, three of the side-grasp trials included an obstacle, as illustrated in Fig. 7(a).
3. We tested the algorithm's ability to handle cluttered scenes, where many of the objects are directly touching or even sitting on top of each other. We placed cluttered scenes with 2 to 9 objects on the table and then asked the robot to clear as many objects as possible using either side- or top-pregrasp positions. The objects were drawn from a wide range of common household and office items, including 29 unique objects: several 450mL rectangular juice bottles, two small boxes (soap, band-aids), the box-shaped PR2 run-stop, a large box of tea bags,

a 305g can of soup, two different sizes of soda cans, a stapler, a tape measure, a white-board eraser, a tape dispenser, a roll of duct tape, a long-stem plastic glass, a 590mL detergent bottle, a box of cereal, a 340g rectangular can of Spam, a round plastic bowl, three different shapes of shampoo bottles, a tall can of shaving cream, a small bottle of vitamins, a toothpaste tube, a bottle of creamer, a bottle of rubbing alcohol, a tin tea can, a 1-quart (946mL) carton of milk, and a 4 cm foam ball.

7 Results and Discussion

The results of experiment 1 are shown in Figure 6; in a top-grasp pose with the fingers 18 cm above the table, our sensor and algorithm corrected for a positioning error on the cylindrical can of at least 10 cm in all directions. Outside of this range the sensor could not see enough points to properly run the algorithm.

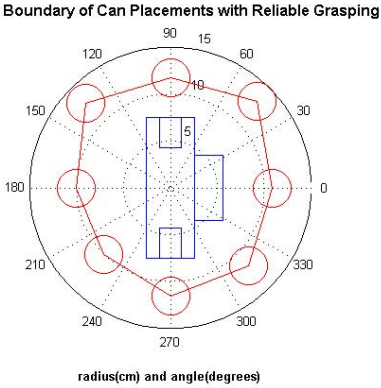


Fig. 6 Basin of attraction for successful object pick up. The gripper always starts in the position and orientation shown at center, with the camera on the right (0 degrees). A can placed anywhere inside the red polygon will be found and picked up by the gripper; outside of this polygon, the object is out of the view of the gripper stereo.

In experiment 2, the robot successfully aligned to the box and executed the grasp properly in 10 out of 10 attempts for both top and side grasps. An example of this adjustment is shown in Fig. 1. Throughout the experimental process we found that the system is very good at aligning to objects. In the cases where an obstacle was added, the algorithm successfully moved the gripper up to avoid the can when grasping the box; this process is illustrated in Fig.7.

In experiment 3, we asked the robot to clean up a variety of cluttered table scenes, including 8 side-grasp and 7 top-grasp scenes; Fig. 8 shows a sampling of these scenes. In 15 scenes using 29 unique objects, the robot succeeded in removing 56 of 64 objects, for a success rate of 87.5%. Figure 9 shows a full clearing sequence for one of the scenes.

Figures 8(e), 8(f), and8(g) represent crowded arrangments of various bottles and boxes the robot might encounter on a counter or refrigerator. These scenes show off the ability of the gripper stereo to find good grasp points from the side, whereas

(a) (b) (c)

Fig. 7 (a): The pregrasp pose is aligned with the box, but a soup can stands in the way. The can would not be visible from the robot head, and so this is another common problem scenario. (b): The algorithm adjusts the positioning to avoid the soup can while remaining properly aligned with the box. (c): The completed grasp.

(a) 3/4 (b) 5/6 (c) 3/3

(d) 5/5 (e) 7/9 (f) 6/6

(g) 4/4 (h) 3/3 (i) 3/3

Fig. 8 A sampling of the scenes the robot attempted to clear. The fraction of objects successfully cleared from the table is noted below each picture. We attempted 7 top-grasp and 8 side-grasp scenes.

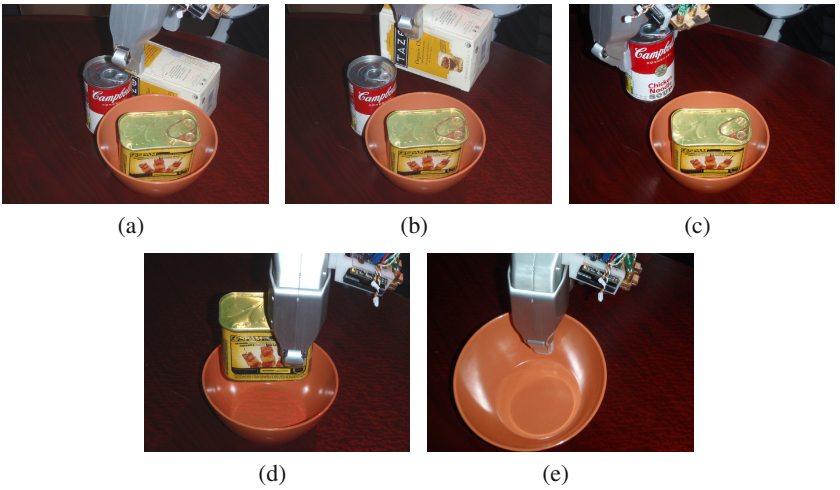


Fig. 9 The robot clears a cluttered table scene by finding viable grasp poses and removing objects one at a time.

the head stereo would be limited to seeing mostly the tops of the bottles where the grasping surfaces are smaller. The algorithm was good at adjusting the height of the grasp to avoid objects on either side of the target, such as grasping the brown part of the bottle in Fig. 8(f) to avoid the small white box on the right side.

There were three observed failure modes in our experiments. The most basic was that objects would escape the field of view of the sensor in the gripper's pre-defined starting position, which happened 3 times. Since many objects were touching or stacked on each other it was very difficult to avoid disturbing other objects at all during a grasp; usually the disturbance was minor and did not affect the feasibility of subsequent grasps, but other times, particularly for very light objects, it was enough to lead to subsequent failure. For example, the robot has no problem grasping the large, tan tea box in general (e.g. Fig. 8(d)), but in another top-grasp scene the box ended up slightly outside the search area of the algorithm. Such failures could be easily fixed just by having the robot use the head cameras to move the gripper near objects or clusters of objects first, rather than using the fixed pregrasp positions we have limited ourselves to for these experiments.

The second failure mode we observed was due to interacting with two objects at once, which typically happened with shallow objects lying on top of other objects. Since the algorithm wants to move the gripper forward as far as possible to get a more stable grasp, the fingers sometimes go too far forward while grasping a shallow object and instead grasp or disturb the supporting object. For example, in one scene a cordless phone on top of a rectangular box was tipped off when the gripper pushed down on the box. In two other cases the gripper successfully grabbed the object underneath, but without restraining the top object, so that the top object was dropped when the gripper moved. Fixing the general problem of possibly grasping more

than one object at a time when the objects appear to be contiguous in the point cloud would require more advanced reasoning about object segmentation, which is beyond the scope of this work.

The last failure mode was from poor grasp selections due to difficult point clouds, which happened twice. In one scene several rectangular juice bottles were placed side by side, which resulted in a relatively flat point cloud when viewed from the side. Faced with few choices, the algorithm chose a rather poor grasp point and failed to get the first bottle. When one of them was removed, the contour of the rest of the objects was easier for the algorithm to deal with. Another failure occurred during one of the attempts to grab a tape dispenser: because other objects were blocking grasps of the center of the dispenser, the algorithm chose a grasp at one end that was not strong enough to support the (fairly heavy) dispenser once it was lifted off the table.

8 Conclusions and Future Work

This work presents a novel, compact stereo sensor mounted to the gripper of a robot and an algorithm for using 3D data from this sensor to find grasp poses for unknown objects in cluttered scenes. We showed that the system is able to correct for fairly large position and alignment errors, by grasping a cylinder and a box in a large variety of positions and orientations. We also showed that the system is able to pick up a large variety of objects in cluttered scenes by finding local grasp points on objects using only 3D data.

The most immediate way to extend this work is to remove the pre-defined pre-grasp locations and allow the robot to determine where there are clusters of points that should be explored. Another immediate way to extend this work would be to combine point clouds from multiple views; we have already successfully used a single gripper stereo cloud in combination with the narrow-field-of-view stereo camera in the PR2 head, and plan to incorporate many more viewpoints from the gripper stereo to obtain a more complete picture of the scene. Because the gripper stereo has much more freedom of movement than sensors mounted to a robot body, it has great promise for scene exploration and reconstruction. We intend to continue exploring the best ways to make use of the gripper stereo sensor, both in autonomous operation and also to aid in assisted teleoperation of manipulation tasks.

Acknowledgments. This work was financially supported in part by Disney Research.

References

1. Saxena, A., Wong, L.L.S., Ng, A.Y.: Learning grasp strategies with partial shape information. In: AAAI (2008)
2. Rao, D., Le, Q., Phoka, T., Quigley, M., Sudsang, A., Ng, A.: Grasping novel objects with depth segmentation. In: IROS (2010)
3. Maldonado, A., Klank, U., Beetz, M.: Robotic grasping of unmodeled objects using time-of-flight range data and finger torque information. In: IROS (2010)

4. Ciocarlie, M., Hsiao, K., Jones, E.G., Chitta, S., Rusu, R., Sucan, I.: Towards reliable grasping and manipulation in household environments. In: ISER (2010)
5. Srinivasa, S., Ferguson, D., Weghe, M.V., Diankov, R., Berenson, D., Helfrich, C., Strasdat, H.: The robotic busboy: Steps towards developing a mobile robotic home assistant. In: 10th International Conference on Intelligent Autonomous Systems (2008)
6. Rusu, R.B., Holzbach, A., Diankov, R., Bradski, G., Beetz, M.: Perception for mobile manipulation and grasping using active stereo. In: Humanoids, Paris (2009)
7. Jain, A., Kemp, C.: EL-E: an assistive mobile manipulator that autonomously fetches objects from flat surfaces. In: Autonomous Robots (2010)
8. Hsiao, K., Nangeroni, P., Huber, M., Saxena, A., Ng, A.Y.: Reactive grasping using optical proximity sensors. In: ICRA, pp. 2098–2105 (May 2009)
9. Mayton, B., LeGrand, L., Smith, J.R.: An electric field pretouch system for grasping and co-manipulation. In: ICRA, pp. 831–838 (May 2010)
10. Konolige, K.: Projected texture stereo. In: ICRA, pp. 148–155 (May 2010)
11. Miller, A.T., Knoop, S., Christensen, H.I., Allen, P.K.: Automatic grasp planning using shape primitives. In: IEEE International Conference on Robotics and Automation, pp. 1824–1829 (2003)
12. Kragic, D., Christensen, H.: Survey on visual servoing for manipulation. Technical report, ISRN KTH/NA/P-02/01-SE, Computational Vision and Active Perception Laboratory (2002)
13. Hsiao, K., Kaelbling, L., Lozano-Perez, T.: Task-driven tactile exploration. In: RSS (2010)
14. Platt Jr., R., Fagg, A.H., Grupen, R.A.: Nullspace composition of control laws for grasping. In: IROS (2002)
15. Dollar, A., Jentoft, L., Gao, J., Howe, R.: Contact sensing and grasping performance of compliant hands. In: Autonomous Robots (2010)
16. Prats, M., Martinet, P., Lee, S., Sanz, P.: Compliant physical interaction based on external vision-force control and tactile-force combination. In: MFI (2008)
17. Natale, L., Torres-Jara, E.: A sensitive approach to grasping. In: Proceedings of the Sixth International Workshop on Epigenetic Robotics (2006)
18. Hsiao, K., Chitta, S., Ciocarlie, M., Jones, E.G.: Contact-reactive grasping of objects with partial shape information. In: IROS (2010)
19. Torabi, L., Gupta, K.: Integrated view and path planning for an autonomous six-dof eye-in-hand object modeling system. In: IROS (2010)
20. Falk, V., McLoughlin, Guthart, G., Salisbury, K., Walther, T., Gummert, J., Mohr, F.: Dexterity enhancement in endoscopic surgery by a computer controlled mechanical wrist. Minimally Invasive Therapy and Allied Technologies 8 (4), 235–242 (1999)
21. Mintz, D., Falk, V., Salisbury Jr., J.K.: Comparison of Three High-End Endoscopic Visualization Systems on Telesurgical Performance. In: Delp, S.L., DiGoia, A.M., Jaramaz, B. (eds.) MICCAI 2000. LNCS, vol. 1935, pp. 385–394. Springer, Heidelberg (2000)