More Than a Feeling: Learning to Grasp and Regrasp using Vision and Touch

Roberto Calandra\textsuperscript{1}, Andrew Owens\textsuperscript{1}, Dinesh Jayaraman\textsuperscript{1}, Justin Lin\textsuperscript{1}, Wenzhen Yuan\textsuperscript{2}, Jitendra Malik\textsuperscript{1}, Edward H. Adelson\textsuperscript{2}, and Sergey Levine\textsuperscript{1}

- Use haptic feedback to determine how to re-grasp for better grasping outcomes
Gelsight as a haptic sensor

(a) [Image of a hand holding an Oreo cookie against a glass surface]
(b) [Image of an Oreo cookie with a textured pattern]
(c) [Image of an Oreo cookie lying flat on a surface]

(a) [Diagram showing the setup with LEDs, camera, and support plate]
(b) [Image of a transparent box with a camera and LEDs inside]
(c) [Image of a circular device with multiple LEDs and a camera]

Ted Adelson’s lab at MIT
Gelsight as a haptic sensor

- Hardness sensing is a valuable capability for a robot.
- Object hardness is generally measured by touch, but there are strict conditions, like the precise control of contact movement.
- There are several attempts to measure hardness, but they work only under specific conditions. For hemispherical objects, we develop a model to estimate hardness. For flat objects, we press the sensor against a set of silicone samples, as shown in Figure 1.

The sensor is pressed on the touched objects, thereby improving its object recognition capability. The touched objects, thereby improving its object recognition capability. The touched objects, thereby improving its object recognition capability. The touched objects, thereby improving its object recognition capability.

We have attempted to expand the robot's ability to estimate object hardness with an optical based touch sensor GelSight. The sensor takes high-resolution tactile images of the object hardness with an optical based touch sensor GelSight. We describe the features that show object variation and which makes them easier to recognize, such as human and animal bodies, cushions, sponges, food, and fabrics.

Among the physical properties of objects, hardness is particularly important. Many objects have distinct hardness patterns that make them easier to recognize, such as human and animal bodies, cushions, sponges, food, and fabrics.

Abstract

- Hardness sensing is a valuable capability for a robot.
- Object hardness can be described in terms of deformation as a function of force, suggesting the need for accurate force sensing in the process. However, humans are surprisingly good at estimating hardness with a passive fingertip, via cutaneous touch alone, evidently based on the deformation pattern of the fingertip [4]. We wish to replicate this capability in a robot fingertip, allowing more convenient hardness estimation when the contact force is unknown or poorly estimated.

The movement is intentionally imprecise; it is performed by a human holding the sensor, and get a set of data during the press, and show some example results in Figure 2. The sensor is pressed on the touched objects, thereby improving its object recognition capability. The touched objects, thereby improving its object recognition capability. The touched objects, thereby improving its object recognition capability.

We have attempted to expand the robot's ability to estimate object hardness with an optical based touch sensor GelSight. The sensor takes high-resolution tactile images of the object hardness with an optical based touch sensor GelSight. We describe the features that show object variation and which makes them easier to recognize, such as human and animal bodies, cushions, sponges, food, and fabrics.

For hemispherical objects, we develop a model to estimate hardness. For flat objects, we press the sensor against a set of silicone samples, as shown in Figure 1.

We have attempted to expand the robot's ability to estimate object hardness with an optical based touch sensor GelSight. The sensor takes high-resolution tactile images of the object hardness with an optical based touch sensor GelSight. We describe the features that show object variation and which makes them easier to recognize, such as human and animal bodies, cushions, sponges, food, and fabrics.

The movement is intentionally imprecise; it is performed by a human holding the sensor, and get a set of data during the press, and show some example results in Figure 2. The sensor is pressed on the touched objects, thereby improving its object recognition capability. The touched objects, thereby improving its object recognition capability. The touched objects, thereby improving its object recognition capability.
Gelsight as a haptic sensor

Ted Adelson's lab at MIT
Robotic Setup

- 7-DoF Sawyer Arm + Weiss WSG-50 Parallel Gripper
- Two GelSight Sensors
- Microsoft Kinect2 Sensor
Overall Architecture

The model architecture is shown in Fig. 3. The action-conditional visuo-tactile model is composed of four input branches: camera image, two GelSight images, and tactile networks. We then jointly optimize the model with the 3D motion transformed into the gripper's coordinate system. Moreover, we also provided the network with the end effector pose, a 4-dimensional vector that includes position and angle.

The action is a 5-dimensional vector consisting of a 3D motion, in-plane rotation, and change in force. Likewise, the probability of success of the grasp is obtained from a multi-layer perceptron consisting of two fully-connected layers with 1024 hidden units each. This network is a multi-layer perceptron consisting of two fully-connected layers with 1024 hidden units each. The first layer of this fusion network contains 1024 hidden units and produces a grasp success probability.

Once the action-conditional model has been learned, we use it to select the action that maximizes the expected probability of success. Although this is an interesting question for future work, we do not investigate the effect of image resolution on performance, as it is substantially lower than the native resolution of the GelSight. Although we did not perform additional manual labeling on a small set of samples for which the automatic classification was borderline ambiguous (e.g., if both sensors were not observed to be in contact), we also automatically generated using deep neural networks successful grasps.

To collect the data necessary to train our model, we determined the starting position of the object and randomize the gripping force, orientation, and closing the gripper with an overly gentle grasp that fails more often. After moving to the desired gripping position and orientation, we wait in the air for 4 seconds (for data augmentation) sample random original GelSight RGB images to automatically collect, but manually labeled, data. This model was initially trained using manually collected data, and we iteratively fine-tuned in a self-supervised manner using the very same augmented data.

The labels for this data (i.e., whether the grasp was successful) were also automatically generated using deep neural networks. We perform additional manual labeling on a small set of samples for which the automatic classification was borderline ambiguous (e.g., if both sensors were not observed to be in contact), or in the rare cases when a visual inspection would indicate a wrong confident of the presence of contacts after lifting), or in the rare cases when a visual inspection would indicate a wrong confident of the presence of contacts after lifting), or in the rare cases when a visual inspection would indicate a wrong confident of the presence of contacts after lifting).

V. D. Collection

To collect a large variety of behaviors, from firm, stable grasps, to occasional slips, to deformations in each GelSight image through background subtraction i.e., we pass the neural network the difference of the GelSight images before and after contact. The action network, and subsequently, the action network), and then pass them through a two-layer fully-connected network that produces a grasp success probability using the learned model.
Data Collection

- Estimate Starting Position
- Set end-effector coordinates and height
- Set orientation and randomize force
- Lift and wait for 4s
Results

Table I: K-fold (K=3) cross-validation accuracy of the different models trained with 18,070 data points.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (mean ± std. err.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chance</td>
<td>62.80% ± 0.85%</td>
</tr>
<tr>
<td>Vision (+ action)</td>
<td>73.03% ± 0.24%</td>
</tr>
<tr>
<td>Tactile (+ action)</td>
<td>79.34% ± 0.66%</td>
</tr>
<tr>
<td>Tactile + Vision (+ action)</td>
<td><strong>80.28% ± 0.68%</strong></td>
</tr>
<tr>
<td>Tactile + Vision (no action)</td>
<td>76.43% ± 0.42%</td>
</tr>
</tbody>
</table>

Table II: Detailed grasping results using different policies for the "Easy" and "Hard" test objects.

```
<table>
<thead>
<tr>
<th>&quot;Easy&quot; set</th>
<th>Objects</th>
<th>Methods</th>
<th>Vision only</th>
<th>Tactile + Vision</th>
<th>Cylinder fitting</th>
<th>Average grasp success</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>% grasp success (# succes / # trials)</td>
<td>% grasp success (# succes / # trials)</td>
<td>% grasp success (# succes / # trials)</td>
<td>% grasp success (# succes / # trials)</td>
</tr>
<tr>
<td></td>
<td>215g</td>
<td>160g</td>
<td>40g</td>
<td>125g</td>
<td>125g</td>
<td>65g</td>
</tr>
<tr>
<td></td>
<td>60% (6/10)</td>
<td>70% (7/10)</td>
<td>50% (5/10)</td>
<td>50% (5/10)</td>
<td>100% (10/10)</td>
<td>100% (10/10)</td>
</tr>
<tr>
<td></td>
<td>95% (95/100)</td>
<td>100% (10/10)</td>
<td>100% (10/10)</td>
<td>100% (10/10)</td>
<td>100% (10/10)</td>
<td>100% (10/10)</td>
</tr>
<tr>
<td>&quot;Hard&quot; set</td>
<td>230g</td>
<td>120g</td>
<td>195g</td>
<td>50g</td>
<td>70g</td>
<td>85g</td>
</tr>
<tr>
<td></td>
<td>60% (6/10)</td>
<td>80% (8/10)</td>
<td>30% (3/10)</td>
<td>30% (3/10)</td>
<td>80% (8/10)</td>
<td>40% (4/10)</td>
</tr>
<tr>
<td></td>
<td>80% (8/10)</td>
<td>100% (10/10)</td>
<td>50% (5/10)</td>
<td>50% (5/10)</td>
<td>90% (9/10)</td>
<td>70% (7/10)</td>
</tr>
<tr>
<td></td>
<td>95% (19/20)</td>
<td>100% (20/20)</td>
<td>35% (7/20)</td>
<td>100% (20/20)</td>
<td>90% (18/20)</td>
<td>15% (3/20)</td>
</tr>
</tbody>
</table>
```
Overall Architecture
Results (Analysis)

(a) Stable grasp

(b) Unstable grasp

(a) Improvement from downward motion

(b) No improvement
Thank you