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

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

Roberto Calandra[1], Andrew Owens[1], Dinesh Jayaraman[1], Justin Lin[1], Wenzhen Yuan[2], Jitendra Malik[1], Edward H. Adelson[2], and Sergey Levine[1]

***Abstract*—For humans, the process of grasping an object relies heavily on rich tactile feedback. Most recent robotic grasping work, however, has been based only on visual input, and thus cannot easily benefit from feedback after initiating contact. In this paper, we investigate how a robot can learn to use tactile information to iteratively and efficiently adjust its grasp. To this end, we propose an end-to-end action-conditional model that learns regrasping policies from raw visuo-tactile data. This model – a deep, multimodal convolutional network – predicts the outcome of a candidate grasp adjustment, and then executes a grasp by iteratively selecting the most promising actions. Our approach requires neither calibration of the tactile sensors, nor any analytical modeling of contact forces, thus reducing the engineering effort required to obtain efficient grasping policies. We train our model with data from about 6,450 grasping trials on a two-finger gripper equipped with GelSight high-resolution tactile sensors on each finger. Across extensive experiments, our approach outperforms a variety of baselines at (i) estimating grasp adjustment outcomes, (ii) selecting efficient grasp adjustments for quick grasping, and (iii) reducing the amount of force applied at the fingers, while maintaining competitive performance. Finally, we study the choices made by our model and show that it has successfully acquired useful and interpretable grasping behaviors.**

***Index Terms*—Deep Learning in Robotics and Automation; Grasping; Perception for Grasping and Manipulation; Force and Tactile Sensing**

I. Introduction

GRASPING is a deeply interactive task: we initiate contact by reaching our fingers toward an object, adjust the placement of our fingers, and balance contact forces as we lift. During this process, the feedback provided by the sense of touch is paramount, as demonstrated by human experiments [1]. Nonetheless, incorporating touch sensing into robotic grasping has thus far proved challenging, due to hardware limitations (e.g., sensor sensitivity and cost) and the difficulty of integrating tactile inputs into standard control schemes. Consequently, the predominant input modalities currently used in the robotic grasping literature are vision and depth.

However, vision does not easily permit the measurement of and reaction to ongoing contact forces, thus significantly hindering the potential benefits of interaction. As a result, vision-based grasping approaches have largely relied on selecting a grasp configuration (location, orientation, and forces) in advance, before making contact with the object.

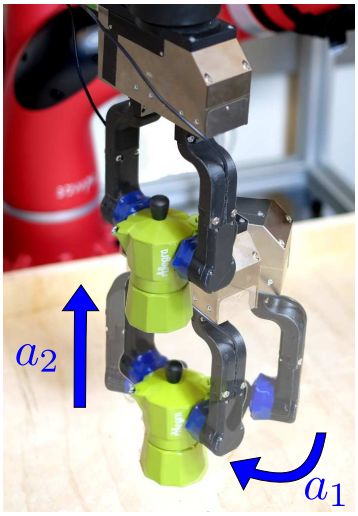


Figure 1: We propose an action-conditional model that iteratively adjusts a robot's grasp based on raw visuo-tactile inputs.

In the quest for interactive grasping, we study how tactile sensing can be integrated into a grasping system that can probe an object and then reactively adjust its grasp to achieve the highest chance of success. Our method is based on learning an action-conditioned grasping model, trained end-to-end in a self-supervised manner by using a robot to autonomously collect grasp attempts. In contrast to prior self-supervised grasping work [2], [3], however, our model incorporates rich touch sensing from a pair of GelSight sensors (see Fig. 1). Incorporating tactile sensing into action-conditional models, however, is not straightforward. The robot only receives tactile input intermittently, when its fingers are in contact with the object and, since each regrasp attempt can disturb the object position and pose, the scene changes with each interaction. In contrast, grasping methods that use vision typically do not interact repeatedly with the object, but simply drive the arm toward a chosen grasp pose and then attempt a single grasp.

Our contributions are as follows: (1) we introduce a new multi-modal action-conditional model for grasping using vision and touch; (2) we show that our model is effective at grasping novel objects, in comparison to unconditional models and vision-only variations; (3) we analyze the learned grasping policy and show that it produces interpretable and useful grasping behaviors; (4) we demonstrate that our model

Manuscript received: February 24, 2018; Revised May 17, 2018; Accepted June 17, 2018.

This paper was recommended for publication by Editor Tamim Asfour upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by Berkeley DeepDrive, NVIDIA, Amazon, the Toyota Research Institute, and the MIT Lincoln Labs.

[1]Roberto Calandra, Andrew Owen, Dinesh Jayaraman, Justin Lin, Jitendra Malik and Sergey Levine are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA {roberto.calandra, owens, dineshjayaraman, justinlin98, malik}@berkeley.edu, svlevine@eecs.berkeley.edu

[2]Wenzhen Yuan and Edward H. Adelson are with the Massachusetts Institute of Technology, USA yuan_wz@mit.edu, adelson@csail.mit.edu

Digital Object Identifier (DOI): 10.1109/LRA.2018.2852779

permits explicit constraints on contact forces, allowing us to command the robot to "gently" grasp an object with significantly reduced force. Since it incorporates raw visuo-tactile inputs, our approach requires neither calibration of the tactile sensors, nor any analytical modeling of contact forces, hence significantly reducing the engineering effort required to obtain efficient grasping policies.

II. Related Work

A. Learning to Grasp

A significant body of work in robotics has studied analytic grasping models, which use known or estimated models of object geometry, environments, and robot grippers, and which typically make use of manually defined grasping metrics [4], [5], [6]. While these methods provide considerable insight into the physical interactions in grasping, their actual performance depends on how well the real-world system fits the assumptions of the analytic model. Model misspecification and unmodeled factors can substantially reduce their effectiveness. As an alternative, data-driven approaches have sought to predict grasp outcomes from human supervision [7], [8], simulation [9], [10], [11], or autonomous robotic data collection [2], [3], typically using visual or depth observations. Among these works, the most related to ours is [3], which also proposes to use an action-conditional model. However, these prior works (with a few exceptions that we discuss below) do not consider tactile sensing, focusing instead on vision and 3D geometry, which afford a limited ability to reason about contact forces, pressures, and compliance. Critically, most of these methods rely on selecting grasp configurations in advance, before ever coming into contact with the target object. In contrast, we show that it is possible to exploit rich tactile feedback *after contact* to iteratively adjust and improve robotic grasps. For an overview of learning for robot grasping, we refer the reader to [12].

B. Tactile Sensors in Grasping

A variety of tactile sensors have been developed [13], mainly measuring force and torque, or the pressure distribution over the sensor. Multiple works [14], [15], [16], [17], [18] suggested the use of tactile sensors to estimate grasp stability. While these works estimate the stability of an ongoing grasp, we focus instead on *selecting* grasp adjustments to produce a stable new grasp. [19] incorporated tactile readings into dynamics models of objects for a dexterous hand, thereby adapting the grasp. Works such as [20], [21], [22] extracted features from tactile signals to detect/predict slip, so as to adaptively adjust the grasping force. Researchers have also proposed robotic systems that integrate visual and tactile information for grasping using model-based methods [23], [24], [25], [26], [27], which improved grasping performance over single-modality inputs. However, these approaches require accurate models of the robot and the objects to grasp, and often also calibrated tactile sensors. Along similar lines, [28] proposed a regrasping policy based on tactile sensing (without visual input) and a learned stability metric, which uses a heuristic transition function to predict future tactile readings. Our approach does not require any prior model or transition function, as it learns entirely end-to-end from raw inputs.

Closer to our approach are [29], [30], which proposed to learn regrasping using tactile sensors. In contrast to our approach, [30] directly optimizes a policy. Optimizing a policy requires the data collection to be on-policy and to be intertwined with the policy update; our approach does not directly optimize a policy, but learns an action-conditioned model. As a result our approach can use any data collected. Additionally, by using an action-conditioned model, we can change the objective of the policy at evaluation time (as in the case of reducing the grasping force demonstrated in Sec. VI-D), while changing the objective for a policy learning method would require re-training the policy, and thus require repeating the data collection process. Another difference with these works is that, in [29], [30], the features used from the tactile sensors are manually designed by applying PCA and extracting the first 5 principal components. Our approach, although using substantially higher resolution tactile inputs, does not require any manual engineering of features. Finally, our experiments consider a substantially wider range of objects than demonstrated by [30], with 65 training objects, and a detailed evaluation on 22 previously unseen test objects.

Closely related is also our previous work [18], where we proposed a visuo-tactile model from raw inputs for classifying grasp outcomes. The main difference to the present work is that [18] does not make use of the learned visuo-tactile model to actively select the next grasp to perform, but simply to evaluate the stability of an ongoing grasp. For grasp selection, this method executes random grasps iteratively until it arrives at a grasp that is stable according to the learned model. While this allows for evaluation of the correlation between touch sensing and grasp outcome, it does not by itself provide a practical method for grasp selection: in our experiments, we found that this prior approach could require as many as 50 random regrasp attempts to yield a stable grasp. Furthermore, by including the grasping force as part of the action, our approach allows for the grasping force to be modulated during the evaluation to achieve secondary objectives, such as minimum-force grasps.

Concurrently to our work, [31] also proposed a tactile regrasping method based on the GelSight sensor. This method simulates transformations to tactile readings based on rigid body dynamics, while our approach is entirely data-driven and self-supervised, which means that we do not require assumptions about dynamics or environment structure. An in-depth exploration of the tradeoffs between data-driven and analytic approaches would an interesting future topic of study. Another concurrent work [32], explores grasping with a 3-axis force sensor, but reports comparatively low success rates, focusing instead on tactile localization without vision. Our method uses rich touch sensing that is aware of texture and surface shape, simultaneously incorporates multiple modalities, and can flexibly accommodate additional constraints, such as minimum-force grasps.

The main contribution of this paper is a practical approach that exploits visual and tactile sensing to grasp successfully and efficiently i.e., with as few regrasps as possible. We do

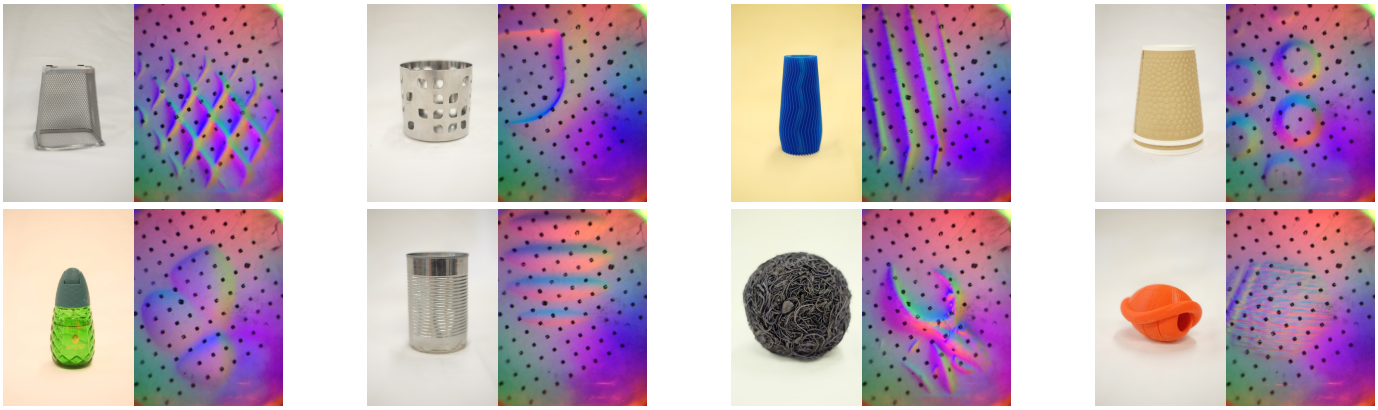


Figure 2: Examples of raw tactile data collected by one of the GelSights (*right*) for different training objects (*left*).

so by building predictive models that can predict the grasp outcome of a given action. Our experiments demonstrate that our action-conditioned predictive model substantially outperforms the results that can be obtained via grasp classification, illustrating the value of closed-loop regrasping. Finally, we demonstrate that our action-conditioned model can be used to optimize for gentler grasps, enabling the robot to determine grasps that can pick up an object with minimal force (hence avoiding damage to fragile objects). To the best of our knowledge, our work is the first to propose an action-conditioned model for learning to grasp from raw visuo-tactile inputs.

III. Hardware Setup

In our experiments we used a hardware configuration consisting of a 7-DoF Sawyer arm, a Weiss WSG-50 parallel gripper, and two GelSight sensors [33], one for each finger. Each GelSight sensor provides raw pixel measurements at a resolution of 1280x960 at 30 Hz over an area of $24\,\mathrm{mm} \times 18\,\mathrm{mm}$. Additionally, a Microsoft Kinect2 sensor was mounted in front of the robot to provide visual data. The GelSight sensor is an optical tactile sensor that measures high-resolution topography of the contact surface [34], [33]. The surface of the sensor is a soft elastomer painted with a reflective membrane, which deforms to the shape of the object upon contact. Underneath this elastomer is a camera (an ordinary webcam) that views the deformed gel. The gel is illuminated by colored lights, which light the gel from different directions. Additional visual cues of contacts are provided by the deformation of the grid of markers painted on the sensor surface, which can be used to compute the shear force and slip information [35]. One valuable property of the GelSight sensor is that the sensory data is provided on a regular 2D grid image format, hence we can use convolutional neural network (CNN) architectures initially designed for visual processing to process readings from the tactile sensor. Previous work on material property estimation with GelSight [36], [37] has successfully applied CNNs pretrained from natural image data. Examples of raw tactile data from the GelSight are shown in Fig. 2.

IV. Deep Visuo-Tactile Models for Grasping

We formalize grasping as a Markov decision process (MDP) where we greedily select the gripper actions that maximize the probability of successfully grasping an object. To address this, we solve the following prediction problem: given the robot's current visuo-tactile observations $\boldsymbol{s}_t$ at time t, and an action $\boldsymbol{a}$, we predict the probability that, after applying the action, the gripper will be in a configuration that leads to a successful grasp at time $t+1$. In Sec. IV-B, we describe how we use this prediction model to select optimal grasping actions.

Raw visuo-tactile observations $\boldsymbol{s}$ are acquired from tactile sensors and the RGB camera, as shown in Fig. 3. Each action $\boldsymbol{a}$ directs the gripper to a new pose relative to its current pose. For example, an action $\boldsymbol{a}$ might consist of moving the gripper to the left by $2\,\mathrm{cm}$, and rotating it by $15°$. More concretely, let $o(\boldsymbol{s}_t, \boldsymbol{a}) \in \{0, 1\}$ be the binary grasp outcome at time $t+1$ resulting from executing action $\boldsymbol{a}$ from grasp state $\boldsymbol{s}_t$: if $o(\boldsymbol{s}, \boldsymbol{a})$ is 1, the grasp is successful. At evaluation time, these outcome labels $o(\boldsymbol{s}_t, \boldsymbol{a}_t)$ are unknown and the robot must estimate them. At training time, the robot performs random trials as described in Sec. V to collect state-action-outcome tuples $(\boldsymbol{s}_i, \boldsymbol{a}_i, o_i) \in X$, which we will use to train an action-conditional model that can be used for selecting actions.

A. End-to-End Outcome Prediction

We would like to learn a function $f(\boldsymbol{s}, \boldsymbol{a})$ that directly predicts the success probability for a future grasp, given observations from the current grasp $\boldsymbol{s}$ and a candidate action $\boldsymbol{a}$. We parametrize f as a deep neural network, whose architecture is shown in Fig. 3. There are multiple design choices when designing deep models for multi-modal inputs [38]. In our experiments, we decided to employ a network processing the state $\boldsymbol{s}$, consisting of raw RGB inputs from the frontal camera and the two GelSight tactile sensors, in three deep stacks of convolutional layers. Additionally, the action $\boldsymbol{a}$ is processed in a two-layer, fully-connected stack (a multi-layer perceptron). We then use a late fusion approach to combine information from these modalities: the feature vectors produced by these four stacks are concatenated, and fed to a two-layer fully-connected network that produces the probability, $f(\boldsymbol{s}, \boldsymbol{a})$, that the input action from the current state results in a successful grasp at the next step. We train the network f on the training dataset X to minimize the loss $L_{dir}(f, X) = \sum_{(\boldsymbol{s},\boldsymbol{a},o)\in X} \mathcal{L}(f(\boldsymbol{s}, \boldsymbol{a}), o)$ where $\mathcal{L}$ is the cross-entropy loss. As input for the tactile CNNs, we rescale the

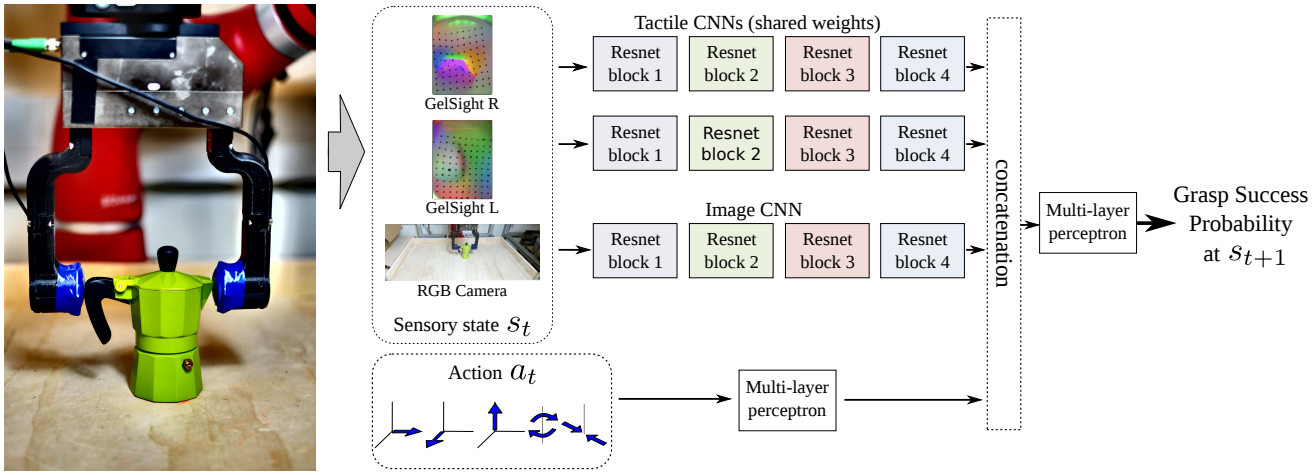


Figure 3: Action-conditioned visuo-tactile model network architecture.

original GelSight RGB images to 256×256, and subsequently (for data augmentation) sample random 224×224 crops. This kind of image resolution is standard for CNN-based object recognition in computer vision, though it is substantially lower than the native resolution of the GelSight. Although we did not investigate the effect of image resolution on performance, this is an interesting question for future work.

a) Network design: We process each image using a convolutional network. Specifically, we use the penultimate layer of a 50-layer deep residual network [39]. We further emphasize 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 is a multi-layer perceptron consisting of two fully-connected layers with 1024 hidden units each. This network takes as input vector representations of the action and pose. The action is a 5-dimensional vector consisting of a 3D motion, in-plane rotation, and change in force. Likewise, the end effector pose is a 4-dimensional vector represented by position and angle. Moreover, we also provided the network with the 3D motion transformed into the gripper's coordinate system. To fuse these networks, we concatenate the outputs of the four input branches (camera image, two GelSight images, and the action network), and then pass them through a two-layer fully-connected network that produces a grasp success probability. The first layer of this fusion network contains 1024 hidden units. Our model architecture is shown in Fig. 3.

b) Training: To speed up training, we pretrain these networks using weights from a model trained to classify objects on ImageNet [40], and we tie the weights of the two tactile networks. We then jointly optimize the model with a batch size of 16 for 9,000 iterations (using a dataset of 18,070 examples), lowering the learning rate by a factor of 10 after 7000 iterations.

B. Regrasp Optimization

Once the action-conditional model f has been learned, we use it to select the action that maximize the expected probability of success of the grasp *after* performing the action

$$\boldsymbol{a}_t^* = \arg\max_{\boldsymbol{a}} f\left(\boldsymbol{s}_t, \boldsymbol{a}\right) . \tag{1}$$

We perform this optimization using stochastic search: we randomly sample potential actions and predict the success probability using the learned model f, and then select the action with the highest success probability. Although this optimization can be computationally expensive (in our experiments, approximately $0.6\,\mathrm{s}$ for 5000 samples), in practice we find that it performs well.

V. Data Collection

To collect the data necessary to train our model, we designed a self-supervised automated data collection process. In each trial, depth data from the front Kinect was used to approximately identify the starting position of the object and enclose it within a cylinder. We then set the end-effector (x, y) coordinates to the position of the center of the cylinder plus a small random perturbation, and set its height to be a random value between the floor and the height of the cylinder. Its orientation ϕ was set uniformly at random. Moreover, we randomized the gripping force F to collect a large variety of behaviors, from firm, stable grasps, to occasional slips, to overly gentle grasps that fail more often. After moving to the chosen position and orientation, and closing the gripper with the desired gripping force, the gripper attempt to lift the object and wait in the air for $4\,\mathrm{s}$. If the object was still in the gripper at the end of this time, the robot would place the object back at a randomized position, and a new trial would start.

The labels for this data (i.e., whether the grasp was successful) were also automatically generated using deep neural network classifiers (running two instances, one for each finger) trained to detect contacts using the raw GelSight images observed[1]. We performed additional manual labeling on a small set of samples for which the automatic classification was borderline ambiguous (e.g., if both sensor were not confident of the presence of contacts after lifting), or in the rare cases when a visual inspection would indicate a wrong

[1]This model was initially trained using manually collected data, and iteratively fine-tuned in a self-supervised manner using the very same automatically collected, but manually labeled, data.

label. Overall, we collected 6,450 grasping trials from over 65 training objects.

As the gripper moves from one position to another, the locations that it moves to along the way can provide additional data points for training. We use this idea to augment the dataset with additional examples. When the robot is gripping an object, we create a state-action pair with zero translation or rotation, corresponding to the action of the robot keeping the gripper in the same position (a useful possible action for regrasping). Similarly, we create a state-action pair at the moment that the robot has released the gripper but has not yet moved. In this case, the action is the same as when the gripper is in contact with the object. After this augmentation, our dataset contains 18,070 examples.

During the data collection and experimental evaluation, we replaced the gels of the two GelSight sensors multiple times due to wear and tear. Each gel is unique, and as a result produces slightly different inputs (e.g., grid of markers might not be evenly aligned). Moreover, with the progressive wear of the surface a single gel, the images can significantly change over time. In our experiments we noticed how, initially, replacing the gel would degrade the performance of the learned models. However, after collecting data with a few different gels, changing the gels did not seem to significantly affect performance anymore, hence suggesting that the model learned features that are reasonably invariant to the specific gel being used.

VI. Experimental Results

To validate our multi-modal grasping model, we first compare the performance of the model on the dataset we collected. Then, we test the model on an actual robot, and evaluate its generalization capabilities on additional (unseen) test objects. Moreover, we analyze the learned visuo-tactile model to gain some insight into its learned behavior and features. Finally, we demonstrate that it is possible to exploit our visuo-tactile action-conditioned model to minimize the applied forces while maintaining a high success rate. Videos showing the robotic grasping experiments (and other material) are available online at: https://sites.google.com/view/more-than-a-feeling

A. Model Evaluation

First, we ask: can our model successfully learn to predict future grasp success for novel objects? Recall that while previous works such as [18] have shown that it is possible to predict stability of ongoing grasps from visuo-tactile inputs, we seek to evaluate the stability of *future* grasps, conditional on a relative adjustment from the current grasp. We compare the predictive performance of a number of variations of our model, using our dataset of grasps (Sec. V). For this, we use K-fold ($K = 3$) cross-validation, partitioning the data by object instance. Does our model learn to use actions to predict future outcomes? This is critical, since we expect to use this model to search over possible actions during grasping on a robot. To test this, we evaluate the model in Fig. 3 without the action ("Tactile + Vision (no action)" in Tab. I) – an unconditional model similar to the one considered in [18] – which without having access to the action corresponds to computing the expectation over all the possible actions. We see that performance indeed drops significantly when action information is withheld, validating that the model learns to successfully evaluate the importance of different actions. Next, we test whether our model significantly outperforms variations where different components are ablated, such as the vision-only and tactile-only models. As seen in Tab. I, the full visuo-tactile model performs best – results for future-grasp prediction that are consistent with those reported in [18] for the task of evaluating current grasps.

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

Model	Accuracy (mean $\pm$ std. err.)
Chance	$62.80\% \pm 0.85\%$
Vision (+ action)	$73.03\% \pm 0.24\%$
Tactile (+ action)	$79.34\% \pm 0.66\%$
Tactile + Vision (+ action)	$\mathbf{80.28\% \pm 0.68\%}$
Tactile + Vision (no action)	$76.43\% \pm 0.42\%$

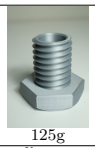
B. Robot Grasp Evaluation

Next, we evaluated the learned models on the robot. In these experiments, we had the robot grasp a given object after executing a series of regrasp actions. Each grasp begins by randomly sampling an end-effector position and angle with the manually engineered system used for the data collection of Sec. V, but without closing the fingers of the robot. Since we start from a configuration where the fingers are not in contact, it is impossible to fairly compare against the tactile-only variant of our model, which requires the robot to already be in contact with the object to select a meaningful action. Consequently, we compare with the vision-only variant of our model, which is similar to that in [3]. We then use the learned models to select the next grasp, by solving the optimization of Eqn. (1). For the action optimization, we consider translations in the interval $[-2, +2]$ cm, gripper rotations from $[-17°, 17°]$, and force values in $[4, 25]$ N. The optimization is performed by randomly sampling 4900 actions, plus 100 additional actions sweeping over the grasping force interval, but having the end-effector rotation and translation set to 0. Each action results in performing a translation and rotation of the end-effector, and in closing the fingers with the desired force. Moreover, if the predicted grasp success probability is above the desired threshold, the re-grasp also includes lifting the object. In our experiments, we set this threshold to 0.9. To ensure that the probabilities are well-calibrated, we applied Platt scaling [41] to its probability predictions, using a validation set containing approximately 1900 examples.

As a baseline, we also evaluated against an approach that fits a cylinder around the object using depth data and subsequently attempt to grasp the centroid of the object using a constant grasping force of 10N. Since we used this cylinder fitting approach as a component of our data collection procedure, it was manually engineered to perform well.

We first trained the models on 18,070 data points collected as described in Sec. V, and evaluated them on a test set of

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

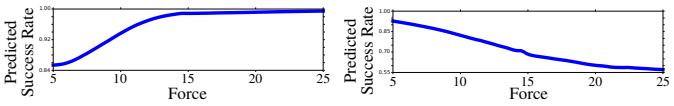
	Objects / Methods	 215g	 160g	 40g	 125g	 125g	 65g	 135g	 30g	 380g	 140g	 10g	Average grasp success
"Easy" set		% grasp success (# success / # trials)											
	Vision only	76% (38/50)	70% (7/10)	60% (6/10)	50% (5/10)	50% (5/10)	90% (9/10)	40% (4/10)	60% (6/10)	**90% (9/10)**	10% (1/10)	**100% (10/10)**	63.2%
	Tactile + Vision	**95% (95/100)**	**100% (10/10)**	**100% (10/10)**	**100% (10/10)**	90% (9/10)	**100% (10/10)**	**90% (9/10)**	**100% (10/10)**	80% (8/10)	**90% (9/10)**	90% (9/10)	**94.0%**
	Cylinder fitting	90% (18/20)	90% (18/20)	80% (16/20)	55% (11/20)	**100% (20/20)**	**100% (20/20)**	**90% (18/20)**	75% (15/20)	35% (7/20)	20% (4/20)	**100% (20/20)**	75.9%
	Objects / Methods	 230g	 120g	 195g	 50g	 70g	 85g	 38g	 165g	 65g	 340g	 110g	Average grasp success
"Hard" set		% grasp success (# success / # trials)											
	Vision only	60% (6/10)	80% (8/10)	30% (3/10)	30% (3/10)	80% (8/10)	40% (4/10)	60% (6/10)	50% (5/10)	50% (5/10)	50% (5/10)	20% (2/10)	50%
	Tactile + Vision	80 % (8/10)	**100% (10/10)**	**50% (5/10)**	80% (8/10)	**90% (9/10)**	**70% (7/10)**	**100% (10/10)**	40% (4/10)	**60% (6/10)**	**80% (8/10)**	60% (6/10)	**73.6%**
	Cylinder fitting	**95% (19/20)**	**100% (20/20)**	35% (7/20)	**100% (20/20)**	**90% (18/20)**	15% (3/20)	90% (18/20)	**85% (17/20)**	15% (3/20)	15% (3/20)	**95% (19/20)**	66.8%

11 previously unseen objects (that we call "Easy"). These objects significantly differed from the ones seen in the training set in terms of color, weight, shape, friction, etc. From the evaluations, we found that our visuo-tactile model significantly outperformed both the vision-only and the cylinder fitting models, achieving 94% accuracy. However, on the harder objects from the "Hard" test set, this learned model would not perform very well. Hence, we decided to collect more data on the training objects, but this time *on-policy* using the learned model. We thus collected a new dataset consisting of 25,404 datapoints, which we used to re-train both the Vision and Tactile+Vision models. After retraining, we evaluated the performance again on the "Hard" test set. In Tab. II, we can see how the visuo-tactile model again outperform the other two models. Based on these experiments, the largest improvements in performance of our model seem to happen in the presence of compliant objects, and objects where it is difficult to visually ascertain a good grasp, such as small or irregular objects. Another interesting result is that the vision-only model performs quite poorly. We hypothesize that the main cause is the relatively small size of the dataset. Prior work [3] used a smaller model and 40x more data. As such, it is likely that the performance of our tactile+vision model could also be further improved by collecting more data.

C. Understanding the Learned Visuo-Tactile Model

Our approach relies on a future grasp evaluation model learned entirely from data, without manual specification of heuristically useful behaviors. We now examine qualitatively: what strategies has our model learned and what behaviors does it produce?

1) Grasping Force: The first question we study is whether or not the model has learned the importance of modulating the amount of force F applied at the fingers for the grasp outcome. Naturally, a stronger grasp is typically more likely to succeed. To test this hypothesis, we placed the gripper in a state where it was in contact with a previously unseen object. We then asked the model to predict the probability of grasp success given various finger forces, keeping the other parts of the action vector fixed. Given this state and candidate actions, we computed the corresponding success rate prediction. As illustrated in Fig. 4, the model appears to have learned that



(a) Stable grasp (b) Unstable grasp

Figure 4: Predicted grasp success rate with varying the amount of force F. The model learned that, when stably in contact with the object, there is a correlation between force applied and success rate. However, for unstable grasps, the model learned that increasing the grasp force might misplace the object and result in an unsuccessful grasp.

there is a correlation between the force and the grasp outcome. However, further analysis shows that the model did not just learn to increase the force in all cases: for multiple situations having very high forces seems to reduce the predicted success rate. For example, we saw this occur when the robot grasped a cube whose corner was only half in contact with the fingers. Due to the shape of the fingers, applying large forces in this case would cause the object to be displaced and slip out of the fingers, and the model correctly predicts that lower forces should be preferred (see Fig. 4b).

2) Height and Center-of-Mass: A second important question is what the model learned with respect to the height of the grasp. For instance, it may be important to grasp close to the vertical center-of-mass of the object: objects that are held close to their top might slip away under even small perturbations. At the same time, objects that are grasped below the center-of-mass might be unstable and rotate around the contact, increasing the chance of slippage. Evaluating the model in different circumstances shows that the model learned that the probability of success increases when decreasing the height of the fingers (an example is shown in Fig. 7). The model did not however, seem to have learned any relevant correlation between the height of the object, or

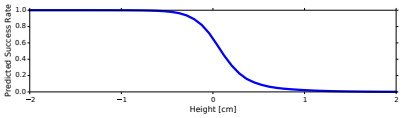
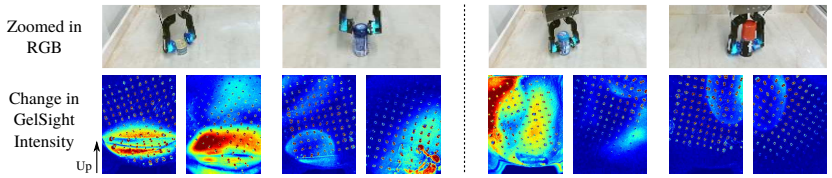


Figure 7: Example of predicted grasp success rate varying the height of the fingers. The model learned that decreasing the height of the fingers generally increases the success rate.



(a) Improvement from downward motion (b) No improvement

Figure 5: What does the model learn? Here we show examples where the network predicts that a downward motion will result in a grasp with (a) higher or (b) lower chance of succeeding. Notice that downward movement is predicted to be beneficial for cases where the fingers hold the top of an object, but not when they hold it by the bottom. To more clearly visualize the contact on the robot's fingertip, we show the change in intensity of the GelSight images.

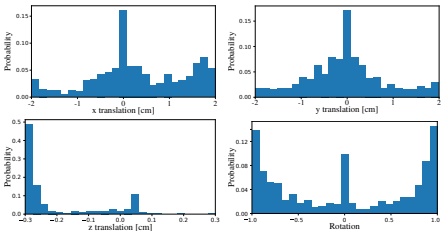


Figure 6: Histograms of the actions applied by the Tactile+Vision policy for the successful grasps. It can be noticed how the policy strongly favour moving downward.

the center-of-mass, and the preference for moving downward. In Fig. 5, we show examples, taken from our dataset, of cases in which the model strongly preferred a downward motion to a static or upward one. For this, we trained a variation of our model without the end effector pose, so that it cannot use the height above the table as a cue. We show held-out examples with the most (and least) predicted improvement in grasp success. The examples with the largest improvement in downward motion tend to be cases in which the top of the object has been gripped (which result in a visible bump in the bottom of the GelSight image). Fig. 6 shows histograms of the actions performed by the Tactile+Vision model for the successful grasps in Sec. VI-B. For the z-translation, almost 50% of the actions used the maximum downward motion allowed (i.e., 2cm), which clearly shows that the learned model acquired a strong preference for moving downward to produce stable grasps.

D. Minimum Force Grasp

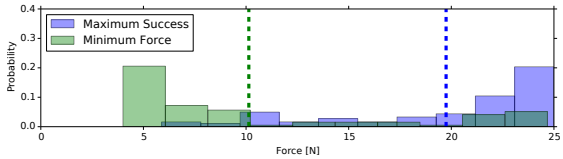
One of the benefits of training an action-conditional grasp outcome prediction model, in contrast to the static grasp classification model in prior work [18], is that we can predict how successful a given grasp will be if we modify the strength of the grasp. Humans typically do not use the strongest grasp possible, but rather employ the minimum amount of contact force, out of consideration for energy consumption and object fragility. Our model also allows us to directly optimize for grasps with either a constraint on the contact force, or via a weighted combination of contact force and grasp success probability. In this experiment, we modified the optimization in Eqn. (1) as a constrained optimization problem such that the selected action would instead minimize the use of force, but while still having an expected success rate $> 90\%$ (if such an action existed, otherwise it would revert to the standard optimization task).

We evaluated the success rate and applied the force of grasps optimized for either pure grasp success or the minimum force objective on the 'Green tea cup' object. After evaluating 100 grasps for each criterion using the Tactile+Vision model, we observed a fairly similar grasp success rate, with $95/100$ successful grasp for the maximum success optimization and $94/100$ for the minimum force grasps. However, we can see in Fig. 8a that, for the successful grasps, the force distribution of the minimum force grasp optimization was substantially lower compared to the maximum success criterion (mean of 10 vs 20 N). Similar results were obtained also when evaluating the Vision only model, as shown in Fig. 8b. This time, both criteria achieved a success rate of 76% (out of 50 trials), which is lower than the Tactile+Vision model. However, the force distribution of the minimum force grasping policy was substantially lower compared to the maximum success criteria at 6 vs 18 N. These results suggest that using a minimum force optimization with our learned model can effectively reduce the amount of force exerted when grasping, without impacting performance. We believe that this is an important result that show the quality of the learned visuo-tactile model, and further motivate the use of tactile sensors in applications which require handling of fragile objects (i.e., glass or fruit, such as strawberries).

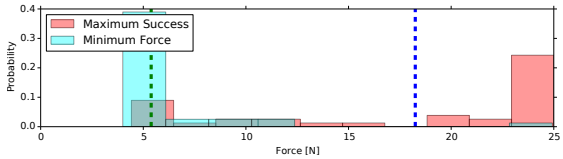
VII. Conclusions

Touch sensing is an inherently active sensing modality, and it is natural that it would be best used in an active fashion, via feedback controllers that incorporate tactile inputs *during* the grasping process. Designing such controllers is challenging, particularly with complex, high-bandwidth tactile sensing combined with visual inputs. In this paper, we introduced a novel action-conditional deep model capable of incorporating raw inputs from vision and touch. By using raw visuo-tactile information, this model can continuously re-plan what action to take so as to best grasp objects. To train this model, we collected over 6,000 trials from 65 training objects. The learned model is capable of grasping a wide range of unseen objects, and with a high success rate. Moreover, we demonstrated that with an action-conditioned model, we can easily decrease the amount of force exerted when grasping, while preserving a similar chance of success.

Our method has multiple limitations that could be addressed in future work. First, our action-conditioned model only makes single-step predictions, and does not perform information-gathering actions. Second, we consider relatively coarse actions – A model using fine-grained actions could more delicately manipulate the object before the grasp, and potentially react to slippage during the lift-off. Finally, it would be valuable to extend our approach to more realistic cluttered environments. Together, addressing these limitations would require a transition to more continuous feedback control



(a) Tactile+Vision



(b) Vision only

Figure 8: Histogram and mean (dashed lines) of the forces applied in the successful grasps. (a) Although the success rates for the two Tactile+Vision policies are similar (95% maximum success vs 94% minimum force), the mean force applied is significantly reduced when using the minimum force policy (10 vs 20 N). (b) The success rates for the Vision only policies is lower at 76%, but again the mean force applied is significantly reduced when using the minimum force policy (6 vs 18 N).

strategy (potentially using torque control), which is an exciting avenue for future work.

REFERENCES

[1] R. S. Johansson and J. R. Flanagan, "Coding and use of tactile signals from the fingertips in object manipulation tasks," *Nature Reviews Neuroscience*, vol. 10, no. 5, 2009.

[2] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *International Conference on Robotics and Automation (ICRA)*, 2016.

[3] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *IJRR*, pp. 421–436, 2016.

[4] K. B. Shimoga, "Robot grasp synthesis algorithms: A survey," *IJRR*, vol. 15, no. 3, 1996.

[5] C. Goldfeder and P. K. Allen, "Data-driven grasping," *Autonomous Robots*, vol. 31, no. 1, 2011.

[6] A. Rodriguez, M. T. Mason, and S. Ferry, "From caging to grasping," *IJRR*, vol. 31, no. 7, 2012.

[7] I. Kamon, T. Flash, and S. Edelman, "Learning to grasp using visual information," in *International Conference on Robotics and Automation (ICRA)*, vol. 3, 1996.

[8] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *IJRR*, vol. 34, no. 4-5, 2015.

[9] D. Kappler, J. Bohg, and S. Schaal, "Leveraging big data for grasp planning," in *International Conference on Robotics and Automation (ICRA)*, 2015.

[10] E. Johns, S. Leutenegger, and A. J. Davison, "Deep learning a grasp function for grasping under gripper pose uncertainty," in *International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[11] J. Mahler, F. T. Pokorny, B. Hou, M. Roderick, M. Laskey, M. Aubry, K. Kohlhoff, T. Kröger, J. Kuffner, and K. Goldberg, "Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards," in *International Conference on Robotics and Automation (ICRA)*, 2016.

[12] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis – a survey," *Transactions on Robotics*, 2014.

[13] H. Yousef, M. Boukallel, and K. Althoefer, "Tactile sensing for dexterous in-hand manipulation in robotics—a review," *Sensors and Actuators A: physical*, vol. 167, no. 2, 2011.

[14] Y. Bekiroglu, J. Laaksonen, J. A. Jorgensen, V. Kyrki, and D. Kragic, "Assessing grasp stability based on learning and haptic data," *Transactions on Robotics*, vol. 27, no. 3, 2011.

[15] J. Schill, J. Laaksonen, M. Przybylski, V. Kyrki, T. Asfour, and R. Dillmann, "Learning continuous grasp stability for a humanoid robot hand based on tactile sensing," in *BioRob*. IEEE, 2012.

[16] H. Dang and P. K. Allen, "Stable grasping under pose uncertainty using tactile feedback," *Autonomous Robots*, vol. 36, no. 4, Apr 2014.

[17] D. Cockburn, J. P. Roberge, T. H. L. Le, A. Maslyczyk, and V. Duchaine, "Grasp stability assessment through unsupervised feature learning of tactile images," in *International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 2238–2244.

[18] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, "The feeling of success: Does touch sensing help predict grasp outcomes?" *Conference on Robot Learning (CORL)*, 2017.

[19] M. Li, Y. Bekiroglu, D. Kragic, and A. Billard, "Learning of grasp adaptation through experience and tactile sensing," in *International Conference on Intelligent Robots and Systems (IROS)*, 2014.

[20] A. Bicchi, M. Bergamasco, P. Dario, and A. Fiorillo, "Integrated tactile sensing for gripper fingers," in *Int. Conf. on Robot Vision and Sensory Control*, 1988.

[21] J. M. Romano, K. Hsiao, G. Niemeyer, S. Chitta, and K. J. Kuchenbecker, "Human-inspired robotic grasp control with tactile sensing," *Transactions on Robotics*, vol. 27, no. 6, 2011.

[22] F. Veiga, H. van Hoof, J. Peters, and T. Hermans, "Stabilizing novel objects by learning to predict tactile slip," in *Conference on Intelligent Robots and Systems (IROS)*, 2015.

[23] P. K. Allen, A. T. Miller, P. Y. Oh, and B. S. Leibowitz, "Integration of vision, force and tactile sensing for grasping," *Int. J. Intelligent Machines*, vol. 4, pp. 129–149, 1999.

[24] Y. Bekiroglu, "Learning to assess grasp stability from vision, touch and proprioception," Ph.D. dissertation, KTH Royal Institute of Technology, 2012.

[25] C. A. Jara, J. Pomares, F. A. Candelas, and F. Torres, "Control framework for dexterous manipulation using dynamic visual servoing and tactile sensors' feedback," *Sensors*, vol. 14, no. 1, 2014.

[26] Y. Bekiroglu, A. Damianou, R. Detry, J. A. Stork, D. Kragic, and C. H. Ek, "Probabilistic consolidation of grasp experience," in *International Conference on Robotics and Automation (ICRA)*, 2016.

[27] D. Guo, F. Sun, B. Fang, C. Yang, and N. Xi, "Robotic grasping using visual and tactile sensing," *Information Sciences*, vol. 417, 2017.

[28] E. Hyttinen, D. Kragic, and R. Detry, "Estimating tactile data for adaptive grasping of novel objects," in *International Conference on Humanoid Robotics (Humanoids)*, Nov 2017, pp. 643–648.

[29] Y. Chebotar, K. Hausman, O. Kroemer, G. S. Sukhatme, and S. Schaal, "Generalizing regrasping with supervised policy learning," in *International Symposium on Experimental Robotics*. Springer, Cham, 2016.

[30] Y. Chebotar, K. Hausman, Z. Su, G. S. Sukhatme, and S. Schaal, "Self-supervised regrasping using spatio-temporal tactile features and reinforcement learning," in *International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[31] F. R. Hogan, M. Bauzá, O. Canal, E. Donlon, and A. Rodriguez, "Tactile regrasp: Grasp adjustments via simulated tactile transformations," *arXiv preprint arXiv:1803.01940*, 2018.

[32] A. Murali, Y. Li, D. Gandhi, and A. Gupta, "Learning to grasp without seeing," *arXiv preprint arXiv:1805.04201*, 2018.

[33] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, 2017.

[34] M. K. Johnson and E. Adelson, "Retrographic sensing for the measurement of surface texture and shape," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[35] W. Yuan, R. Li, M. A. Srinivasan, and E. H. Adelson, "Measurement of shear and slip with a gelsight tactile sensor," in *International Conference on Robotics and Automation (ICRA)*, 2015.

[36] W. Yuan, C. Zhu, A. Owens, M. A. Srinivasan, and E. H. Adelson, "Shape-independent hardness estimation using deep learning and a gelsight tactile sensor," in *International Conference on Robotics and Automation (ICRA)*, 2017.

[37] W. Yuan, S. Wang, S. Dong, and E. Adelson, "Connecting look and feel: Associating the visual and tactile properties of physical materials," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[38] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *International Conference on Machine Learning (ICML)*, 2011, pp. 689–696.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[41] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.