Tactile-Visual Integration for Task-Aware Grasping

Mabel M. Zhang*, Andreas ten Pas†, Renaud Detry‡, and Kostas Daniilidis*

*GRASP Laboratory
University of Pennsylvania
{zmen@seas, kostas@cis}.upenn.edu

†Computer and Information Science
Northeastern University
atp@ccs.neu.edu

‡Jet Propulsion Lab
California Institute of Technology
renaud.j.detry@jpl.nasa.gov

I. Introduction

Vision-based grasping has seen extensive studies [2]. There are two ways in general to approach data-driven grasping. Traditionally, grasping is done in a pipeline dependent on the knowledge of the object [12]. First, the object identity and pose are visually estimated. Then, grasp candidates trained on CAD models by some quality measure are retrieved and transformed from the object frame into the robot frame. Grasps are executed in descending quality, while pruning unreachable grasps in terms of inverse kinematics and collision-aware motion trajectory planning.

A second approach is independent of object identity. Given a scene, the grasp detector is simply given the raw camera input and predicts grasp candidates by geometry only [26, 17, 27, 22, 29, 21, 9, 34, 30]. The advantage is that it does not depend on correct identity and pose estimation, eliminating the risk of error propagation. With that, however, comes the disadvantage that the grasp detection is completely unaware of object semantics, and is thus only useful for pick and place tasks such as emptying a basket.

A common disadvantage for both approaches is that neither take object functionality into account. In cases of common tool use, such as hammer, pliers, or key, the object must be picked up in a certain orientation in order to execute its functionality. On the other hand, when the task is simply transportation, the object can be picked up in any orientation.

1) Task-driven grasping: This shortcoming has been addressed in several ways. A direct extension to the first approach is to add constraints to the grasp candidates based on the given task [31]. A more direct alternative is to compute grasps by simultaneously taking into account object identity and functionality. To this end, affordance estimation and task-driven grasping have been studied [20, 10, 28, 1]. More recently, deep learning has enabled grasp detection that takes object identity into account without explicit recognition [18].

2) Touch-based grasping: So far, all cases above are vision-based grasping. Recently, improvement in tactile sensing brought touch back into the light for perception [25] and grasping [5, 3, 13, 15, 23, 6, 8, 16, 7]. Other than exclusively touch-based grasping, touch is also an effective complement to vision-based manipulation [11, 24, 14, 19]. The latest visuotactile integration leverages convolutional neural networks (CNNs) for their capability of end-to-end derivation from raw sensor readings directly to prediction [33, 4].

We develop a new representation for visuotactile integration suitable for feature-embedding in CNNs and evaluate grasp success. To assess the tactile modality and grasp success, we train a data set of grasps in simulation, using an array of contact sensors on a gripper (Fig. 1), and lift the object into midair. In addition, we are able to measure task success, using an existing visual semantics predictor [10] to give hints of task constraints, in the form of probabilistic heat maps.

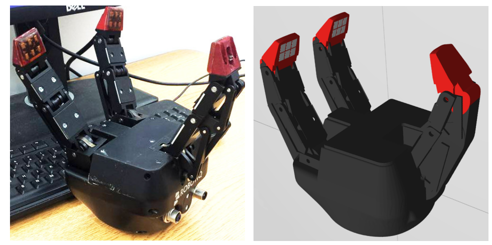


Fig. 1: Robotiq gripper with TakkTile sensors, in real world and simulation.

We show preliminary results that demonstrate the plausibility of touch, even sparse contacts, in improving grasp success. We aim to evaluate task success in the future. Furthermore, we seek to compare and evaluate for the optimal 2D visuotactile representation. The significance of these projected observations is that, first, tactile sensors are exclusively either high resolution or affordable. We target the latter type, both for accessibility and for independence on sophisticated sensors and therefore wider adaptability. Second, since touch is a 3D modality, 2D representations inevitably lose information. However, because of the exponential growth of CNN parameters and the sparse nature of touch, 2D image is a compact representation that makes sparse inputs more meaningful.

In the future, we plan on transferring the model learned in simulation to the real robot. To this end, the simulation is built to resemble the real environment, and the simulated contact sensors have the same resolution and are at the same locations as on the real gripper. We expect that the model may need to be retrained or fine-tuned for the real robot.

II. Visuotactile Representation

Our goal of visuotactile grasp prediction presents two problems: spatial correspondence between modalities, and representation for learning. An obvious answer is point clouds [11, 16], which is straight-forward for reconstruction. However, we are interested in higher-level abstractions.

We propose a concatenation of 2D image channels. The first channel is the depth image. Subsequent channels come from tactile contacts. Upon contact, the activated tactile sensors' 3D positions are obtained by forward kinematics. These positions are transformed into the camera frame and projected into the image using the intrinsics matrix, which gives the 2D pixels that correspond to the 3D positions. This completes the spatial correspondence between image and touch.

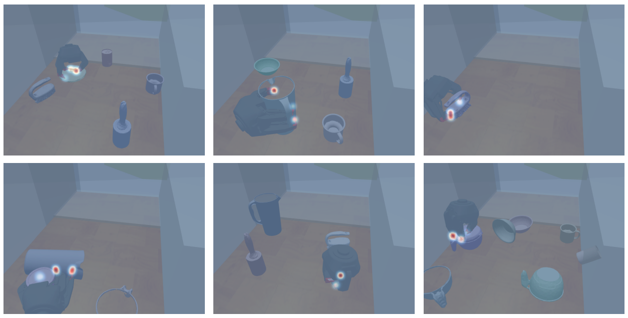


Fig. 2: Example tactile channels, shown as heat map overlaid on RGB for illustration. Top/bottom row: successful/unsuccessful grasps.

To account for object movement and calibration error, the exact image pixel is not used; instead, it will be blurred. A tactile map of the dimensions of the camera image is initialized to zeros. Pixels corresponding to contacts are given non-zero values. The resulting matrix is convolved with a maximum filter and a Gaussian filter. This has two effects. First, it removes the dependence on accurate 3D-to-2D correspondence, to allow small object movement. Second, it creates denser representation of the otherwise single-pixel contacts.

The number of channels in the tactile matrix depends on the representation. We propose several for evaluation: 1. raw depth z of contacts; 2. thickness $d = z_T - z_C$ between contact and camera depth; 3. normals of activated sensors, scaled by thickness, $d\hat{n}$. Fig. 2 shows examples in the 3-channel (xyz) normal and thickness representation. Each channel is visualized as a heat map; all three are overlaid.

Fig. 3 illustrates the process of visuotactile representation to grasp success prediction. We use an off-the-shelf TensorFlow CNN implementation and augment the first and last layers.

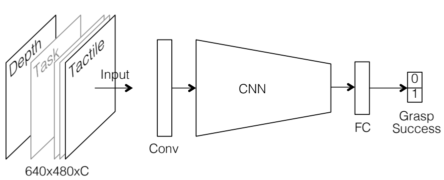
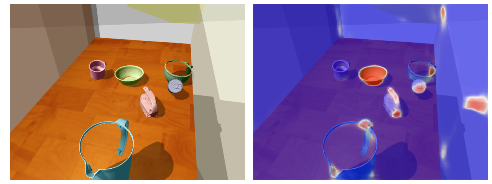


Fig. 3: Number of input channels varies for each tactile representation. Task channel is used for task success evaluation.

III. Tactile Grasp Dataset Collection

Grasps with tactile readings are collected in simulation. We spawn a random set of objects at random positions on a table (Fig. 4(a)). For each scene, the gripper executes a number of grasps, given by an off-the-shelf grasp planner *e.g.* [32]. Any vision-based planner that gives a wrist pose and a score can be used. Grasps with good and bad scores are executed, to produce positive and negative training examples.

The grasp collection process is as follows. For each grasp, the gripper moves to the goal wrist pose. The fingers are closed in pinch mode, to make maximum use of the fingertip sensors. At this point, the object is fixed. Tactile sensors are read, and the tactile map is constructed as in Section II. Fig. 5 shows example grasps. Then, the gripper is lifted 50 cm. After the lift, if the object is still with the gripper, the grasp is successful.



(a) (b)

Fig. 4: (a). A scene. (b). Semantic task map for *carry*, overlaid on RGB.

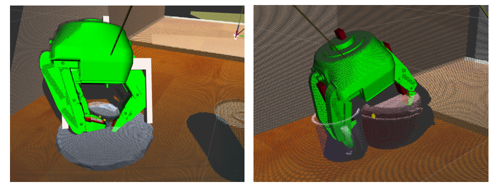


Fig. 5: Left/right: Successful/unsuccessful grasps. Rectangular gripper shape is goal pose; green gripper is actual gripper. On left, surface normals scaled by z-thickness are plotted as yellow vectors at the activated fingertip sensors.

To evaluate semantic task success, we use an existing vision-based per-pixel classifier [10], which outputs a probabilistic heat map (Fig. 4(b)). Its values indicate whether a pixel, if in contact with the gripper, is compatible with a given task – *carry, pour, handover,* or *open*. For example, for *pour*, the gripper should avoid regions near the opening. On the other hand, for *carry*, the gripper is free to lift at the opening. Binary task-compatibility is labeled per-vertex in the CAD model, one model per task. Ground truth task success is thus obtained by the task labels of the contact points in the object frame.

Thus far, we have collected over 10,000 grasps in Gazebo on 10 computers in parallel. The bottleneck is in the gripper movement, which cannot be sped up. We will evaluate the simplest tactile representation first, and task success at last. Since the simulation mimics our real environment, including collision scene, we anticipate the obstacles in transferability to be in sensor noise, reachability, and network adaptability.

IV. Preliminary Results

To investigate the issue of sparse contacts in the whole scene, we first evaluate on an existing data set for a simplified case – overhead planar parallel grasps on cropped images [27]. We simulate contacts on Dex-Net Adv-Synth (188,300 image-grasp pairs) depth images by gradient along the grasp axis. The addition of tactile input, even as constant-peak blobs, yielded lower errors (Fig. 6). Meaningful peaks should improve futher.

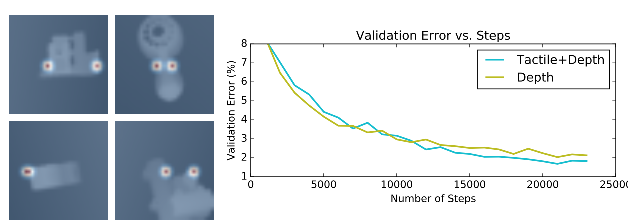


Fig. 6: Left: Tactile heat maps with blobs of constant peak $z = 1$, overlaid on depth image. Top/bottom: good/bad grasps. Right: Tactile+depth input yielded lower error rates than depth alone, as steps increase.

References

[1] Amir M. Ghalamzan E., Nikos Mavrakis, Marek Kopicki, Rustam Stolkin, and Ales Leonardis. Task-relevant grasp selection: A joint solution to planning grasps and manipulative motion trajectories. In *IROS*, 2016.

[2] Jeannette Bohg, Antonio Morales, Tamim Asfour, and Danica Kragic. Data-Driven Grasp Synthesis - A Survey. *TRO*, 30(2):289–309, April 2014.

[3] Jeannette Bohg, Karol Hausman, Bharath Sankaran, Oliver Brock, Danica Kragic, Stefan Schaal, and Gaurav S. Sukhatme. Interactive Perception: Leveraging Action in Perception and Perception in Action. *arXiv:1604.03670*, 2016.

[4] Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H. Adelson, and Sergey Levine. The Feeling of Success: Does Touch Sensing Help Predict Grasp Outcomes? In *CoRL*, volume 78, pages 314–323, 2017.

[5] Yevgen Chebotar, Karol Hausman, Oliver Kroemer, Gaurav S. Sukhatme, and Stefan Schaal. Regrasping using Tactile Perception and Supervised Policy Learning. In *AAAI Spring Symposium on Interactive Multi-Sensory Object Perception for Embodied Agents*, 2017.

[6] Hao Dang and Peter Allen. Stable Grasping under Pose Uncertainty Using Tactile Feedback. *AURO*, 2014.

[7] Hao Dang and Peter K. Allen. Learning grasp stability. In *ICRA*, 2012.

[8] Hao Dang and Peter K. Allen. Grasp adjustment on novel objects using tactile experience from similar local geometry. In *IROS*, 2013.

[9] Renaud Detry, Carl Henrik Ek, Marianna Madry, and Danica Kragic. Learning a Dictionary of Prototypical Grasp-predicting Parts from Grasping Experience. In *ICRA*, pages 601–608, May 2013.

[10] Renaud Detry, Jeremie Papon, and Larry Matthies. Task-oriented Grasping with Semantic and Geometric Scene Understanding. In *IROS*, 2017.

[11] Pietro Falco, Shuang Lu, Andrea Cirillo, Ciro Natale, Salvatore Pirozzi, and Dongheui Lee. Cross-modal Visuo-Tactile Object Recognition Using Robotic Active Exploration. In *ICRA*, pages 5273–5280, May 2017.

[12] Corey Goldfeder, Matei Ciocarlie, Hao Dang, and Peter K. Allen. The Columbia Grasp Database. In *ICRA*, Kobe, Japan, May 2009.

[13] Kaiyu Hang, Miao Li, Johannes A. Stork, Yasemin Bekiroglu, Florian T. Pokorny, Aude Billard, and Danica Kragic. Hierarchical Fingertip Space: A Unified Framework for Grasp Planning and In-Hand Grasp Adaptation. *TRO*, 32(4):960–972, August 2016.

[14] Robert Haschke. Grasping and Manipulation of Unknown Objects Based on Visual and Tactile Feedback. *Mechanisms and Machine Science*, 29:91–109, March 2015.

[15] Emil Hyttinen, Danica Kragic, and Renaud Detry. Learning the Tactile Signatures of Prototypical Object Parts for Robust Part-based Grasping of Novel Objects. In *ICRA*, 2015.

[16] Jarmo Ilonen, Jeannette Bohg, and Ville Kyrki. Fusing visual and tactile sensing for 3-D object reconstruction while grasping. In *ICRA*, 2013.

[17] Stephen James, Andrew J. Davison, and Edward Johns. Transferring End-to-End Visuomotor Control from Simulation to Real World for a Multi-Stage Task. In *CoRL*, volume 78, pages 334–343, 2017.

[18] Eric Jang, Sudheendra Vijaynarasimhan, Peter Pastor, Julian Ibarz, and Sergey Levine. End-to-End Learning of Semantic Grasping. In *CoRL*, 2017.

[19] Carlos A. Jara, Jorge Pomares, Francisco A. Candelas, and Fernando Torres. Control Framework for Dexterous Manipulation Using Dynamic Visual Servoing and Tactile Sensors' Feedback. *Sensors*, 14(1):1787–1804, January 2014.

[20] Mia Kokic, Johannes Andreas Stork, Joshua Alexander Haustein, and Danica Kragic. Affordance Detection for Task-Specific Grasping Using Deep Learning. In *Humanoids*, 2017.

[21] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *IJRR*, 34(4-5), March 2015.

[22] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection. *IJRR*, June 2017.

[23] Miao Li, Yasemin Bekiroglu, Danica Kragic, and Aude Billard. Learning of grasp adaptation through experience and tactile sensing. In *IROS*, 2014.

[24] Qiang Li, Robert Haschke, and Helge Ritter. A visuo-tactile control framework for manipulation and exploration of unknown objects. In *Humanoids*, Seoul, South Korea, November 2015.

[25] Shan Luo, Joao Bimbo, Ravinder Dahiya, and Hongbin Liu. Robotic tactile perception of object properties: A review. *Mechatronics*, 48:54–67, December 2017.

[26] Jeffrey Mahler and Ken Goldberg. Learning Deep Policies for Robot Bin Picking by Simulating Robust Grasping Sequences. In *CoRL*, 2017.

[27] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio, and Ken Goldberg. Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics. In *RSS*, Cambridge, Massachusetts, July 2017.

[28] Anh Nguyen, Dimitrios Kanoulas, Darwin G. Caldwell, and Nikos G. Tsagarakis. Detecting Object Affordances with Convolutional Neural Networks. In *IROS*, 2016.

[29] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours. In *ICRA*, pages 3406–3413, Stockholm, Sweden, May 2016.

[30] Mila Popović, Dirk Kraft, Leon Bodenhagen, Emre Başeski, Nicolas Pugeault, Danica Kragic, Tamim As-

four, and Norbert Krüger. A strategy for grasping unknown objects based on co-planarity and colour information. *RAS*, 58(5):551–565, May 2010.

[31] Dan Song, Carl Henrik Ek, Kai Huebner, and Danica Kragic. Task-Based Robot Grasp Planning Using Probabilistic Inference. *TRO*, 31(3), 2015. doi: 10.1109/TRO.2015.2409912.

[32] Andreas ten Pas, Marcus Gualtieri, Kate Saenko, and Robert Platt. Grasp Pose Detection in Point Clouds. *IJRR*, 36(13-14), October 2017.

[33] Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward Adelson. Connecting Look and Feel: Associating the visual and tactile properties of physical materials. In *CVPR*, 2017.

[34] Li Emma Zhang, Matei Ciocarlie, and Kaijen Hsiao. Grasp evaluation with graspable feature matching. In *Workshop on Mobile Manipulation: Learning to Manipulate, RSS*, 2011.