Visual-Tactile Geometric Reasoning

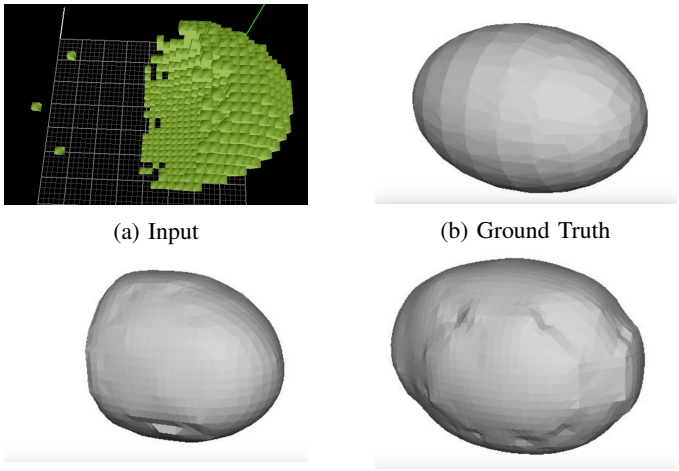
Jacob Varley, David Watkins, and Peter Allen

I. Introduction

This work utilizes a 3D Convolutional Neural Network (CNN) to enable 3D geometric reasoning by incorporating both tactile and depth information to infer occluded geometries. Despite recent advances [13][14] accomplishing robotic manipulation tasks based on data captured from a single depth sensor remains difficult. The difficulty arises from the fact that a single depth image does not always provide enough information to accurately predict occluded geometry. For instance the partial view of an egg from Fig. 1a is very difficult to discern from a bowl which was also present during training.

The idea of integrating additional sensory information from tactile and force sensors to reduce geometric uncertainty from vision alone is not new [11]. Several recent approaches to integrate multi-modal data for robotic manipulation tasks and geometric reasoning have focused on the use of Gaussian Process Implicit Surfaces (GPIS) [15][3][16][2][5][8][12][10]. Unfortunately these approaches have difficulty making predictions in regions far from tactile observations. Others recent approaches based on heuristics [7][6][1] do not easily extend to more sophisticated geometries.

We present a data-driven approach utilizing a CNN as shown in Fig. 2. In our approach, a pointcloud of the visible portion of an object is captured, and used to provide an initial hypothesis of the object's geometry via [14]. This initial hypothesis is used to plan a grasp or exploratory action. The hand is then moved to the planned position via a guarded move, stopping when contact with the object occurs. At this point, the newly acquired tactile information is combined with the original partial view and sent through a CNN to create an updated object geometry hypothesis. This new hypothesis incorporates both the depth and tactile information. Fig. 1 shows how the completion quality is improved by the fusion of depth and tactile information over depth information alone.



(a) Input (b) Ground Truth (c) Depth Only Completion (d) Tactile + Depth Completion

Fig. 1: Egg completion from the YCB and grasp database holdout model set. It is hard to determine how far back the completion actually goes, and it is hard to differentiate what object this is as the dataset contains both eggs and bowls. The Tactile + Depth Completion is better as it uses the tactile information to alleviate both concerns.

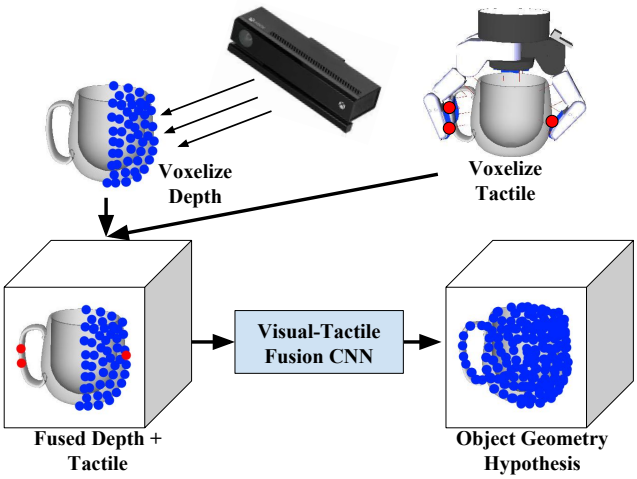


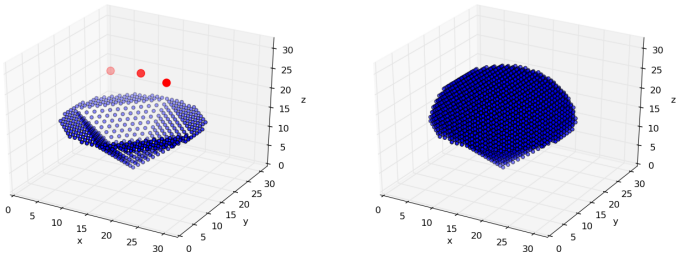
Fig. 2: Both Tactile and Depth information are independently captured and voxelized into 40^3 grids. These are merged into a shared occupancy map which is fed into a CNN to produce a hypothesis of the object's geometry.

II. Visual-Tactile Geometric Reasoning

A. Half-Shapes Dataset

In order to first evaluate our system, it was trained on a Half-Shapes dataset. This dataset consists of conjoined shapes. Both front and back halves of the objects were randomly chosen to be either a sphere, cube, or diamond. Next, synthetic sensory data was generated for these example shapes. Depth information was captured in addition to tactile information collected using both a tactile exploration (3 points from the back), and a tactile grasp (2 points from the back, 1 from the front). An example shape with tactile exploration sensory data is shown in Fig. 3.

Four networks with the exact same architecture were trained on this dataset using different sensory data as input. The results are shown in Fig. 4. One network was only provided the tactile grasp information during training, and performed poorly. A second network was given only the depth information during training, and performed better than the first network, but still encountered many situations where it did not have enough



(a) (b)

Fig. 3: A Half-Shape dataset example. (a) The red dots represent the tactile exploration readings. The blue dots on represent to occupancy map from the depth image. (b) The ground truth 3d geometry.

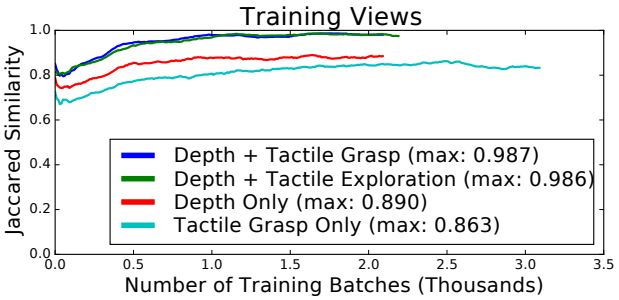


Fig. 4: Different runs of the shape completion system trained on the Half-Shape dataset. Input was provided from: Depth, Depth + Tactile Exploration, Depth + Tactile from Grasp, and only from a Tactile Grasp. When using both tactile and depth the system is able to complete the object almost 100% of the time. While depth or tactile alone are not sufficient to successfully reason about object geometry in all cases.

information to accurately complete the back half of the object. The other two networks were given the depth and tactile information. One in the form of a tactile grasp and the other from a tactile exploration. These networks were able to learn the task to completion. They successfully utilized the tactile information to differentiate between plausible geometries of occluded regions.

B. YCB and Grasp Dataset

After demonstrating on the Half-Shape dataset, we trained two additional models using 486 of the grasp[9] and YCB[4] dataset objects. These models come from the open-source dataset of voxelized training data provided by [14]. This dataset consists of approximately half a million pairs of oriented voxel grids. Where one grid's voxels are marked as occupied if visible to a camera, and the second grid's voxels are marked as occupied if the object intersects a given voxel, independent of perspective. This dataset was augmented with information about up to three additional occupied voxels marking where each finger intersects the object from a tactile exploration action done in the same manner as with the Half-Shape dataset. For training the network, we used a similar architecture to [14], but with three dense layers rather than

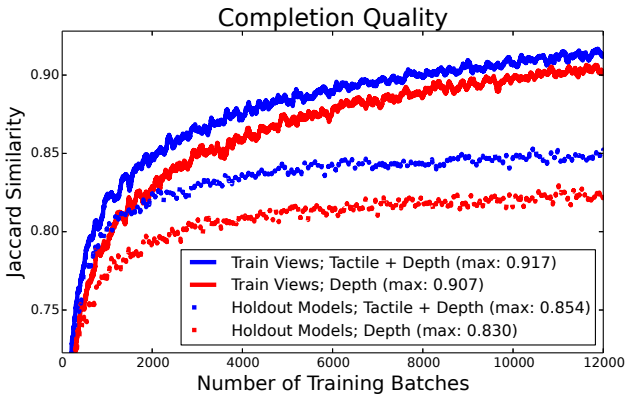


Fig. 5: Jaccard similarity for two CNNs, one (Red: Depth) trained with depth alone, the second (Blue: Tactile + Depth) trained with tactile and depth information. While training, the CNNs were evaluated on inputs they were being trained on (Train Views) and novel inputs from meshes they have never seen before (Holdout Models). In both evaluations the network provided with both depth and tactile is able to do a better job, this is especially true for Holdout Models demonstrated by the widened performance gap between the two networks.

two.

The completions were created using the post-processing code from [14] which merges the CNN output with the observed pointcloud of the object. An example completion from this dataset using our CNN is shown in Fig. 1. The tactile information allows the system to correctly predict how far back the completed object should extend and disambiguate between objects used in training that have similar depth maps but very different completions. Fig. 5 shows how completion quality improves as training progresses for two networks one trained using depth alone, and the second trained using depth and tactile information. It is interesting to note that difference in performance between the two networks is much larger on Holdout Models than on Train Views. This can be interpreted to mean that the additional tactile information is more useful on novel objects, while depth alone maybe sufficient for good completions if the object was used during training.

III. Conclusion

This work demonstrates an architecture for utilizing depth and tactile information to reason about object geometry. This is done via a CNN trained with both depth and tactile information. The CNN predicts object geometry, filling in the occluded regions of an object. At runtime, an initial object hypothesis can be generated using depth alone. Then the framework shown here can provide an improved understanding of the object's geometry by using newly collected tactile information. We have demonstrated that the system is able to produce better predictions of object geometry when utilizing both tactile and depth information as opposed to using depth information alone. This improved object geometry understanding can then be utilized for other robotic manipulation tasks.

References

[1] Alexander Bierbaum, Ilya Gubarev, and Rüdiger Dillmann. Robust shape recovery for sparse contact location and normal data from haptic exploration. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 3200–3205. IEEE, 2008.

[2] Marten Bjorkman, Yasemin Bekiroglu, Virgile Hogman, and Danica Kragic. Enhancing visual perception of shape through tactile glances. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 3180–3186. IEEE, 2013.

[3] Sergio Caccamo, Yasemin Bekiroglu, Carl Henrik Ek, and Danica Kragic. Active exploration using gaussian random fields and gaussian process implicit surfaces. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 582–589. IEEE, 2016.

[4] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *Advanced Robotics (ICAR), 2015 International Conference on*, pages 510–517. IEEE, 2015.

[5] Stanimir Dragiev, Marc Toussaint, and Michael Gienger. Gaussian process implicit surfaces for shape estimation and grasping. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 2845–2850. IEEE, 2011.

[6] Jarmo Ilonen, Jeannette Bohg, and Ville Kyrki. Fusing visual and tactile sensing for 3-d object reconstruction while grasping. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 3547–3554. IEEE, 2013.

[7] Jarmo Ilonen, Jeannette Bohg, and Ville Kyrki. Three-dimensional object reconstruction of symmetric objects by fusing visual and tactile sensing. *The International Journal of Robotics Research*, 33(2):321–341, 2014.

[8] Nawid Jamali, Carlo Ciliberto, Lorenzo Rosasco, and Lorenzo Natale. Active perception: Building objects' models using tactile exploration. In *Humanoid Robots (Humanoids), 2016 IEEE-RAS 16th International Conference on*, pages 179–185. IEEE, 2016.

[9] Daniel Kappler, Jeannette Bohg, and Stefan Schaal. Leveraging big data for grasp planning. In *ICRA*, pages 4304–4311. IEEE, 2015.

[10] Jeffrey Mahler, Sachin Patil, Ben Kehoe, Jur van den Berg, Matei Ciocarlie, Pieter Abbeel, and Ken Goldberg. GP-GPIS-OPT: Grasp planning with shape uncertainty using gaussian process implicit surfaces and sequential convex programming. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[11] Peter K Allen Andrew T Miller and Paul Y Oh Brian S Leibowitz. Integration of vision, force and tactile sensing for grasping. 1999.

[12] Nicolas Sommer, Miao Li, and Aude Billard. Bimanual compliant tactile exploration for grasping unknown objects. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 6400–6407. IEEE, 2014.

[13] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *arXiv preprint arXiv:1611.08974*, 2016.

[14] Jacob Varley, Chad DeChant, Avinash Nair, Joaquin Ruales, and Peter Allen. Shape completion enabled robotic grasping. In *IROS*, 2017.

[15] Oliver Williams and Andrew Fitzgibbon. Gaussian process implicit surfaces. *Gaussian Proc. in Practice*, pages 1–4, 2007.

[16] Zhengkun Yi, Roberto Calandra, Filipe Veiga, Herke van Hoof, Tucker Hermans, Yilei Zhang, and Jan Peters. Active tactile object exploration with gaussian processes. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 4925–4930. IEEE, 2016.