Object Pose Estimation with Geometric Tactile Rendering and Tactile Image Matching

Maria Bauza, Eric Valls, Bryan Lim, Theo Sechopoulos, Alberto Rodriguez
Mechanical Engineering Department — Massachusetts Institute of Technology

I. Introduction

Robotics history sends a clear lesson: accurate and reliable perception is an enabler of progress in robotics. From depth cameras to convolutional neural networks, we have seen how advances in perception foster the development of new techniques and applications. For instance, the invention of high-resolution LIDAR fueled self-driving cars, and the generalization capacity of deep neural networks has dominated progress in perception and grasp planning in warehouse automation [1, 2, 3]. The long term goal of our research is to understand the key role that tactile sensing plays in that progress. In particular we are interested in robotic manipulation applications where occlusions difficult accurate object pose estimation, and where behavior is dominated by contact interactions.

In this work we propose a framework to estimate the pose of a touched object, as illustrated in Fig. 1. Given a 3D model of the object, the framework builds in simulation an object-specific perception model, tailored at estimating the pose of the object from one–or possibly multiple–tactile images of the object. As a result, the approach localizes objects from the first touch, i.e. without requiring any previous interaction. The perception model is based on two key ideas:

- **Geometric tactile rendering** in simulation of the local shapes that the tactile sensor would observe from a dense set of contact poses with the object.
- **Tactile image matching** of the real observed local shape vs. the dense simulated set. A key contribution is to do this comparison in an object-specific embedding learnt in simulation, which provides robustness and speed compared to methods based on pixel comparisons.

The proposed approach is motivated by scenarios where the key requirement is estimation accuracy, and where object models will be available beforehand. Many industrial scenarios fit this category.

Several previous solutions to developing accurate object-specific models for tactile pose estimation require tactile exploration of the object [4, 5]. Acquiring this tactile experience can be expensive, and in many cases unrealistic. In this paper, instead, we learn the perception model directly from the 3D model of the object. The results in Sec. II for four objects show that the model learned in simulation directly transfers to the real world. We attribute this both to the object-specific nature of the learned model, and due to the high-resolution nature of the tactile sensors used.

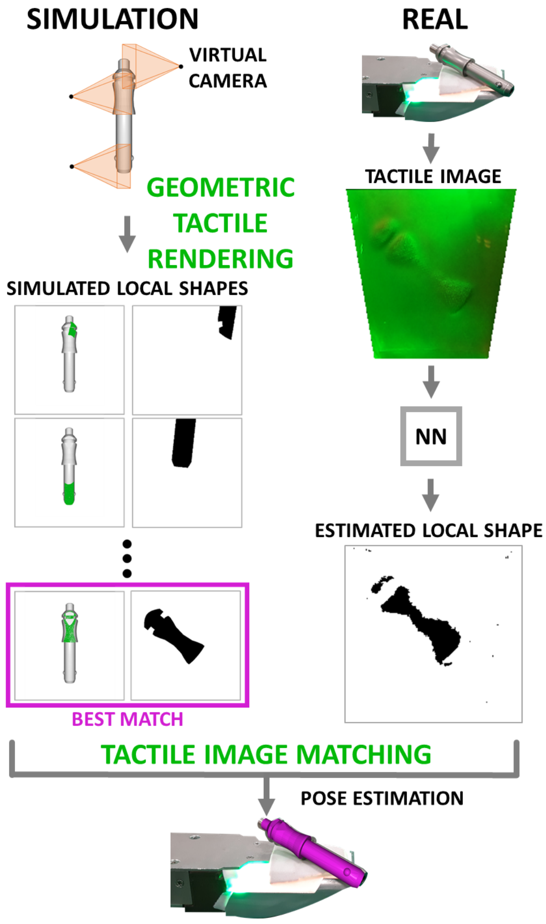

Fig. 1. **Tactile pose estimation.** (left column) In simulation, we render geometric tactile images of the object from a dense set of possible contacts between the object and the tactile sensor. (right column) The real sensor generates a tactile image from where we estimate its geometric local shape. We then match it against the simulated set of local shapes to find the distribution of contact poses that are more likely to have generated it. For efficiency and robustness, we do the local shape matching in an embedding learnt for the particular object.

Also key to the approach is that by simulating a dense set of tactile imprints, the algorithm can reason over pose distributions, not only best estimates. The learned embedding allows to efficiently compute [cosine] distances between a new tactile shape and the simulated dense set. This results in a weighted distribution over object poses rather than just a single pose prediction. This is key to dealing with the fact that tactile provides local observations which in many cases might not be sufficiently discriminative of a single pose.

Finally, by maintaining distributions in pose space, we can incorporate extra constraints over the likelihood of each pose. We have shown this in the case of multi-contact, where information from multiple tactile readings must be combined simultaneously, and in filtering, where pose constraints, in the form of new tactile observations, come over time (results omitted due to space constraints). By operating in a discretization of the pose space, our framework can potentially handle many other pose constrains including constraints from other perception systems (e.g., vision) or kinematics (e.g., non-penetration and grasp stability).

In summary, the main contribution of this work is a framework for tactile pose estimation for objects with known geometry, with the following primary strengths:

1. Provides accurate pose estimation from the first touch, without requiring any previous interaction with the object.
2. Reasons over pose distributions by efficiently computing distances of a real tactile shape vs. a dense set of simulated tactile shapes.
3. Integrates other types of pose constraints such as those arising from spatial distributions in multi-contact or from temporal distributions in filtering.

II. Results

A. Real data collection

While most computations of the algorithm are done in simulation, the end goal of our approach is to provide accurate pose estimation in the real world. To that aim, we specially designed a system that reliably collects tactile imprints and their associated poses:

Tactile sensor. We consider the tactile sensor GelSlim [6] which provides high-resolution tactile readings.

Robot platform. To get controlled touches on the sensor, we fix the sensor to the environment and use a 4 axis robotic stage with translation and rotation in the horizontal plane and vertical motion.

Objects. We test our algorithm on 4 objects from the dataset [7] which contains more than 6k objects meshes from McMaster. For each object, we build a dense grid that contains the set of poses that would result in contact with the sensor. The distance between closest neighbours is no further than 2mm in average.

B. Pose estimation results

We test the accuracy of our approach by estimating object poses from single tactile imprints. For each object, we collected 150 pairs of tactile images and object poses. Given two poses we measure their distance by sampling 5000 points on the object 3D model and averaging the distance between these points when the object is at either of the two poses. To account for the different sizes of the object, we also compute a *normalized pose error* that divides the original pose error by the mean pose error obtained from predicting a random pose from the grid of that object.

Fig. 2 shows the accuracy results for tactile pose estimation for each of the four objects, in the form of error distributions. We include the error distributions for:

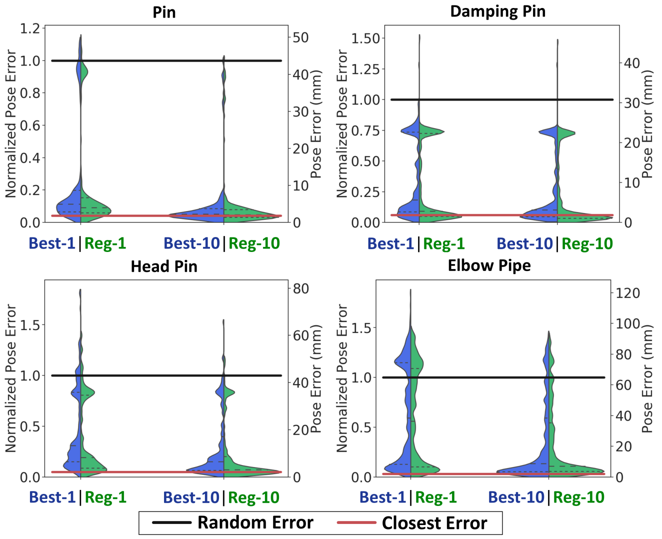

Fig. 2. **Pose estimation results.** For each object, we show in blue the error distributions for the best match and the best out of 10. The distributions in green refine the previous results using pointcloud registration between local shapes. We observe that most of the error distributions are far from the random error (black line) and close to the error obtained when selecting the closest element from the grid (red line). For some objects like *damping pin*, we see multimodality in the error distributions due to different contact poses resulting in similar local shapes.

1) *Best-1*: only considers the most likely pose of the grid.
2) *Reg-1*: refines the most likely pose using FilterReg[8].
3) *Best-10*: considers the 10 most likely poses of the grid and selects the one that leads to lowest pose error. This approach requires knowledge of the true pose and it is not applicable in practice.
4) *Reg-10*: takes the 10 most likely poses, refines them using FilterReg, and selects the one with lowest error.

For all objects, we observe that most of the error distributions are below the expected random error. Moreover most errors are small and gathered around the closest error from the grid (red line), specially when considering the best out of the 10. In some cases, the error distributions are multimodal. That happens specially when different poses of the object lead to very similar local shapes.

Selecting the best error out of the 10 best poses results in considerably lower errors and suggests that our approach can provide meaningful pose distributions. For the case of the *elbow pipe* we even observe that the error distributions become almost unimodal when selecting the best out of 10. Another important observation coming from these plots is that the median errors are considerably low for all objects even when only considering the most likely pose. For *Best-1* and *Reg-1*, we get median normalized errors of 0.11 and 0.09 for *pin* (4.9 and 4mm), 0.18 and 0.09 for *damping pin* (5.6 and 2.8mm), and 0.28 and 0.18 for *head* (11.8 and 7.4mm). The *elbow pipe* is more challenging and bigger, and we get median normalized errors of 0.59 and 0.56 (38.5 and 36.4mm). Finally, adding FilterReg clearly improves the results. This is because it can refine the poses estimates when the initial distance between local shapes is low enough and locally transforming them is possible.

References

[1] A. Zeng, S. Song, K. Yu *et al.*, “Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[2] A. Milan, T. Pham, K. Vijay, D. Morrison, A. W. Tow, L. Liu, J. Erskine, R. Grinover, A. Gurman, T. Hunn *et al.*, “Semantic segmentation from limited training data,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1908–1915.

[3] M. Schwarz, C. Lenz, G. M. García, S. Koo, A. S. Periyasamy, M. Schreiber, and S. Behnke, “Fast object learning and dual-arm coordination for cluttered stowing, picking, and packing,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3347–3354.

[4] R. Li, R. Platt, W. Yuan, A. ten Pas, N. Roscup, M. A. Srinivasan, and E. Adelson, “Localization and manipulation of small parts using gelsight tactile sensing,” in *Intelligent Robots and Systems (IROS), 2014 IEEE/RSJ International Conference on.*, 2014.

[5] M. Bauza, O. Canal, and A. Rodriguez, “Tactile mapping and localization from high-resolution tactile imprints,” *arXiv preprint arXiv:1904.10944*, 2019.

[6] E. Donlon, S. Dong, M. Liu, J. Li, E. Adelson, and A. Rodriguez, “Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger,” *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[7] E. Corona, K. Kundu, and S. Fidler, “Pose estimation for objects with rotational symmetry,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 7215–7222.

[8] W. Gao and R. Tedrake, “Filterreg: Robust and efficient probabilistic point-set registration using gaussian filter and twist parameterization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 095–11 104.