INVESTIGATING REPRESENTATIONS FOR VISION AND TOUCH IN CONTACT-RICH ROBOT SCOOPING TASKS

Jung Ji-Eun
Department of Computer Science
Kyungpook National University
Daegu, South Korea

ABSTRACT

Contact-rich robotic manipulation in unstructured environments remains an open challenge in robotics with no established universal architectures or representations to handle the involved modalities. This paper analyzes different approaches for combining vision and touch to improve robotic scooping, using an open-source scooping dataset. We compare different architectures and modalities and analyze the their impact on in-distribution and out-of-distribution performance. We find that the best performing model on in-distribution terrains is the one which uses both vision and touch data and is trained end-to-end. However, the best performing model on out-of-distribution terrains is the one that uses only vision data.

1 INTRODUCTION

Advances in robotic manipulation have the potential to revolutionize many domains including manufacturing (Matheson et al., 2019), space exploration (Gao & Chien, 2017; Thangeda & Ornik, 2022), and even our everday lives (Kaplan, 2005; Wu et al., 2023). One of the primary sources of challenges hindering seamless application of existing robotic manipulation stack to real-world, unstructured environments comes from the complexity of handling contact mechanics (Cui & Trinkle, 2021). While everyday tasks such as picking a bottle, inserting a key into a lock, or scooping a spoonful of sugar from a sugar jar are trivial for humans, they are still challenging for robots. One of the primary reasons humans are able to complete these tasks effortlessy is because of our inherent ability to combine vision and touch information. Indeed, many recent works (Pinto et al., 2016; Calandra et al., 2017; 2018; Lee et al., 2019; Li et al., 2022) proposed several approaches to exploit multimodal information to increase performance on a variety of contact-rich robotic tasks.

However, despite the recent progress, there is no established universal architecture or representation to handle the involved multiple modalities on a variety of tasks. While some approaches (Lee et al., 2019; Sutanto et al., 2019) use self-supervised learning to learn a joint latent space representation across modalities as an intermediate step, some other approaches (Kumar et al., 2019; Jin et al., 2023) use end-to-end learning paradigm.

In this paper, we attempt to answer some of these questions by performing a detailed analysis of vision and tactile modalities along with different approaches to utilize them. Specifically, we focus on the problem of robot scooping (Schenck et al., 2017; Tuomainen et al., 2022) - a challenging task that is particularly contact-rich and is a good proxy for many other manipulation tasks that require careful handling of contact. Our experiments use the UIUC scooping dataset, an open-source robotic scooping dataset introduced by Zhu et al. (2023). The problem is challenging with real-world applications (Thangeda et al., 2024), and we extend it by utilizing force-torque information.

2 PROBLEM FORMULATION AND SOLUTION APPROACHES

We use a modified version of the few-shot adaptation problem studied in Zhu et al. (2023) and focus on the problem of estimating the scooped volume for our analysis. Formally, given an observation o that consists of RGB-D data and F/T sensor data from end-effector terrain interaction, and an action a that represents the parameters of scooping trajectory, we try to estimate the scooped volume v. We refer the reader to Zhu et al. (2023) for more details about the problem formulation and the dataset.

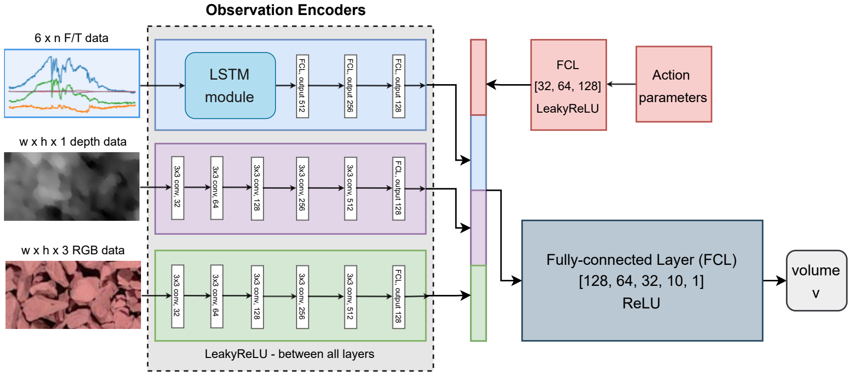

Figure 1: Overview of our architecture for estimating scooped volume from RGB-D and F/T data. All three observation encoders (in dotted box) are pretrained using reconstruction loss in SSL setting.

Figure 1 shows an overview of our architecture that includes both RGB-D and F/T observations as input along with the action parameters. In the case of end-to-end training, we train the entire model using MSE loss. In the case of two stage training, we first train the three observation encoders using self-supervised learning under reconstruction loss and freeze them. We then train the prediction model using MSE loss.

We analyze the following four different architectures: (i) *VisionTactileE2E*, that uses both RGB-D and F/T data as input and is trained end-to-end, (ii) *VisionTactileLatent*, that uses both RGB-D and F/T data as input and is trained in two stages, (iii) *VisionOnlyE2E*, that uses only RGB-D data as input and is trained end-to-end, and (iv) *VisionOnlyLatent*, that uses only RGB-D data as input and is trained in two stages using SSL.

3 Results and Discussion

We use the training and out-of-distribution testing terrains specified in UIUC dataset. We further split the training dataset to perform in-distribution testing. We used consistent choices across experiments to ensure fairness and also tuned hyperparameters each approach independently to maximize performance. Table 1 shows the results from our experiments. Surprisingly, the best performing model on out-of-distribution terrains is *VisionOnlyE2E*. One might expect that the SSL model would perform better on out-of-distribution terrains as they could learn training-data agnostic representations. One possible explanation for this is that the dataset was not large enough to train a good SSL model. Further, the F/T data in out of distribution cases was significantly different from in distribution cases there by hurting the performance.

Table 1: MAE of Predicted Volume Values for Different Approaches

Terrain	VisionTactileE2E	VisionTactileLatent	VisionOnlyE2E	VisionOnlyLatent
In-Dist	**8.2 ± 2.12**	12.1 ± 3.4	13.4 ± 2.8	18.21 ± 3.1
Out-of-Dist	29.4 ± 4.48	35.4 ± 4.1	**21.1 ± 3.5**	27.34 ± 2.44

4 Conclusion and Future Work

In this paper, we analyzed different approaches for robotic scooping using visual and tactile observations. In addition to extending existing work by using F/T data, we also analyzed the impact of using SSL and end-to-end training and the efficacy of introducing tactile modality. In the future, we plan to extend our analysis to other large scale datasets to draw confident, generalizable conclusions.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

References

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? *arXiv preprint arXiv:1710.05512*, 2017.

Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018.

Jinda Cui and Jeff Trinkle. Toward next-generation learned robot manipulation. *Science robotics*, 6 (54):eabd9461, 2021.

Yang Gao and Steve Chien. Review on space robotics: Toward top-level science through space exploration. *Science Robotics*, 2(7):eaan5074, 2017.

Piaopiao Jin, Yinjie Lin, Yaoxian Song, Tiefeng Li, and Wei Yang. Vision-force-fused curriculum learning for robotic contact-rich assembly tasks. *Frontiers in Neurorobotics*, 17, 2023.

Frédéric Kaplan. Everyday robotics: robots as everyday objects. In *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*, pp. 59–64, 2005.

Visak Kumar, Tucker Hermans, Dieter Fox, Stan Birchfield, and Jonathan Tremblay. Contextual reinforcement learning of visuo-tactile multi-fingered grasping policies. *arXiv preprint arXiv:1911.09233*, 2019.

Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8943–8950. IEEE, 2019.

Hao Li, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A Lee, Huazhe Xu, Edward Adelson, Li Fei-Fei, Ruohan Gao, and Jiajun Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation. *arXiv preprint arXiv:2212.03858*, 2022.

Eloise Matheson, Riccardo Minto, Emanuele GG Zampieri, Maurizio Faccio, and Giulio Rosati. Human–robot collaboration in manufacturing applications: A review. *Robotics*, 8(4):100, 2019.

Lerrel Pinto, Dhiraj Gandhi, Yuanfeng Han, Yong-Lae Park, and Abhinav Gupta. The curious robot: Learning visual representations via physical interactions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 3–18. Springer, 2016.

Connor Schenck, Jonathan Tompson, Sergey Levine, and Dieter Fox. Learning robotic manipulation of granular media. In *Conference on Robot Learning*, pp. 239–248. PMLR, 2017.

Giovanni Sutanto, Nathan Ratliff, Balakumar Sundaralingam, Yevgen Chebotar, Zhe Su, Ankur Handa, and Dieter Fox. Learning latent space dynamics for tactile servoing. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 3622–3628. IEEE, 2019.

Pranay Thangeda and Melkior Ornik. Adaptive sampling site selection for robotic exploration in unknown environments. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4120–4125. IEEE, 2022.

Pranay Thangeda, Ashish Goel, Erica L Tevere, Yifan Zhu, Erik Kramer, Adriana Daca, Hari D Nayar, Kris Hauser, and Melkior Ornik. Learning and autonomy for extraterrestrial terrain sampling: An experience report from owlat deployment. In *AIAA SCITECH 2024 Forum*, pp. 1962, 2024.

Neea Tuomainen, David Blanco-Mulero, and Ville Kyrki. Manipulation of granular materials by learning particle interactions. *IEEE Robotics and Automation Letters*, 7(2):5663–5670, 2022.

Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: Personalized robot assistance with large language models. *arXiv preprint arXiv:2305.05658*, 2023.

Yifan Zhu, Pranay Thangeda, Melkior Ornik, and Kris Hauser. Few-shot adaptation for manipulating granular materials under domain shift. In *Proceedings of Robotics: Science and Systems*, July 2023. doi: 10.15607/RSS.2023.XIX.048.