Deep Learning Methods for Tool-in-Hand Tracking

Master Thesis Proposal by Jonas Hein
Supervisors: Dr. Philipp Fürnstahl, Matthias Seibold, Dr. Federica Bogo
March 26, 2020

I. Introduction

Tool-in-hand tracking has many potential applications in various fields: In educational or industrial scenarios, tool-in-hand tracking could be used to teach the correct usage of a tool, or to guide through a series of steps e.g. in the context of a repair or assembly process. Mixed-reality devices need to understand their surroundings, and tracking the tool in the users' hand is a first step to recognizing the users' actions and intentions.

In clinical applications, tool-in-hand tracking has great potential for automated surgery phase detection, in a surgical guidance system, or as a safety system via human error detection. However, surgical scenarios are very challenging, especially since the tools are often occluded by the hands and the acceptable error margins are very small.

II. Related Work

The pose estimation of hands [11–13, 16] or objects [1, 6, 8, 9, 15] has been discussed extensively in the literature. However, for hand-object interactions, jointly estimating the poses for both hand and object might be superior to an isolated pose estimation, since the joint model will learn to take both hand and object into account, thus having an advantage in situations where either the hand or object is occluded by the other (which is in fact the case for almost all real-world hand-tool interactions). Similar to many approaches for individual pose estimation of hands and objects, most approaches for the joint pose estimation of hand and object are based on RGB input [4, 5, 7]. [14] proposes a model which predicts not only a 3D hand pose and a 6D object pose, but also classifies the object as well as the executed action, which results in more accurate pose estimates. [5] reconstructs both hand and object as meshes, enabling it not only to estimate the objects' pose but also to reconstruct novel instances of a known object category.

While recent deep learning based methods produce state-of-the-art results, they entail the problem of gathering large amounts of training data. The collection of training data is especially cumbersome for 6D pose estimation and hand-object interactions, also due to the naturally occurring self-occlusions of both object and hand. However, several datasets have been published recently [4, 5], with some of them focusing on the egocentric perspective [3]. In [2], thermal imaging was used to capture the contact area between hand and object, thus giving more insights on typical grasps.

All related work mentioned above focused on everyday objects and scenarios. In contrast, the surgical environment differs greatly from usual environments and a surgeon will use specialized tools and other medical equipment in an operating room. This domain gap can potentially reduce the performance of models trained on general datasets. However, there are no public datasets for surgical scenarios available yet.

III. Methodology

The goal of this thesis is to develop a tool-in-hand tracking solution for surgical scenarios. In these scenarios, it is assumed that the surgeon wears a Microsoft HoloLens and uses a set of known tools, e.g. a drill. Ideally, the tracking solution only relies on the HoloLens's embedded sensors, which limits the possible inputs to (combinations of) egocentric RGB-D or stereo-RGB images, as well as any postprocessed data based on these sensors. The use of additional external sensors is not planned. Furthermore, it estimates the 3D pose of the surgeons hand, the hand configuration, as well as the 6D pose of the tool in use. The tracking of the hand pose is motivated by the idea of improving the pose estimation of the tool by jointly tracking both tool and hand, thus creating a system which is robust towards occlusions. However, the tool is the main focus of the tracking solution.

Since there are no public datasets for surgical scenarios yet, the generation of synthetic training data is an important part of the thesis. Following the approach from [7], synthetic data samples can be generated by using

the GraspIt! simulator[10] to calculate reasonable grasps for the 3D models of each tool. Then, the hand and tool configuration can be rendered from various viewpoints, while making use of data augmentation techniques to diversify the dataset. The exact methodology for generating the data is still open for discussion. A small set of real samples can be gathered and manually labelled at a later stage of the thesis (for model refinement or testing).

The final model will be evaluated thoroughly and with respect to reliability, accuracy and other established metrics.

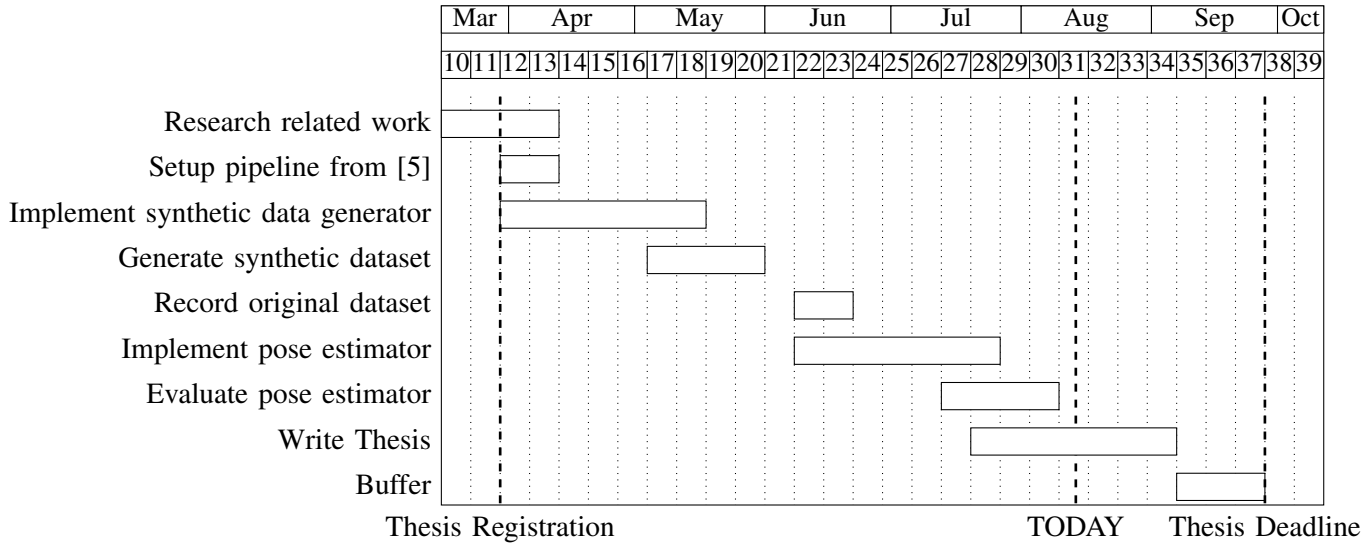

Fig. 1. Proposed time plan on the basis of calendar weeks.

REFERENCES

[1] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *European conference on computer vision*, pages 536–551. Springer, 2014.

[2] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8709–8719, 2019.

[3] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018.

[4] Shreyas Hampali, Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Ho-3d: A multi-user, multi-object dataset for joint 3d hand-object pose estimation. *arXiv preprint arXiv:1907.01481*, 2019.

[5] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11807–11816, 2019.

[6] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1521–1529, 2017.

[7] Mia Kokic, Danica Kragic, and Jeannette Bohg. Learning to estimate pose and shape of hand-held objects from rgb images. *arXiv preprint arXiv:1903.03340*, 2019.

[8] Alexander Krull, Eric Brachmann, Frank Michel, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Learning analysis-by-synthesis for 6d pose estimation in rgb-d images. In *Proceedings of the IEEE international conference on computer vision*, pages 954–962, 2015.

[9] Frank Michel, Alexander Kirillov, Eric Brachmann, Alexander Krull, Stefan Gumhold, Bogdan Savchynskyy, and Carsten Rother. Global hypothesis generation for 6d object pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 462–471, 2017.

[10] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004.

[11] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–59, 2018.

[12] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *2011 International Conference on Computer Vision*, pages 2088–2095. IEEE, 2011.

[13] Paschalis Panteleris and Antonis Argyros. Back to rgb: 3d tracking of hands and hand-object interactions based on short-baseline stereo. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 575–584, 2017.

[14] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2019.

[15] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[16] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. 3d hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214*, 2016.