Rule Of Thumb: Deep derotation for improved fingertip detection

Aaron Wetzler[1]
www.cs.technion.ac.il/~twerd
Ron Slossberg[1]
www.cs.technion.ac.il/~ronslos
Ron Kimmel[2]
www.cs.technion.ac.il/~ron

[1] Faculty of Electrical Engineering,
Technion Israel Institute of Technology
[2] Faculty of Computer Science,
Technion Israel Institute of Technology

In this paper we propose DeROT, a method for in-plane derotation of depth images using a deep convolutional neural network. The method is aimed at normalizing out the effects of rotation on highly articulated motion of deforming geometric surfaces such as hands. To support our approach we also describe a new pipeline for building a very large training database using high accuracy magnetic annotation and labeling of objects imaged by a depth camera. he proposed method reduces the complexity of learning in the space of articulated poses which is demonstrated by using two different state-of-the-art learning based hand pose estimation methods applied to fingertip detection. Significant classification improvements are shown over the baseline implementation. Our framework involves no tracking, kinematic constraints or explicit prior model of the articulated object.

DeROT: removing in-plane rotation Changing the global rotation of an object directly increases the variation in appearance of the object parts. For markerless situations, removing variability through partial canonization can significantly reduce the space of possible images used for pose learning instead of trying to explicitly learn the rotational variability through data augmentation. We therefore remove the variability as a preprocessing step during *both* a training phase *and* at run-time. To this end we propose to learn the rotation using a deep convolutional neural network (CNN) in a regression context based on a network similar to that of [4]. We show how this can be used to predict full three degrees of freedom (3 DOF) orientation information by training on a large database of hand images captured by a depth sensor. This is then combined with a useful insight which we call "Rule of thumb": there is almost always an in-plane rotation which can be applied to an image of the hand which forces the base of the thumb to be on the right side of the image. Synthetic and real examples of the results of applying DeROT to images of a hand can be seen in Figure 1.

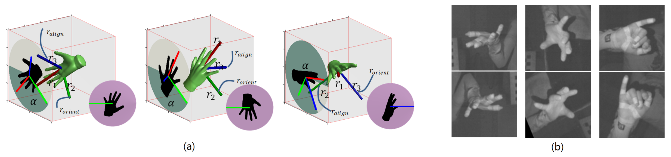

Figure 1: Synthetic and real examples of DeROT. a) The depth projection of the virtual hand before applying DeROT can be seen on the left wall of the cube representing the camera plane. The axis marked r_{orient} is projected onto the camera plane and used in DeROT to define the angle α. The purple circle contains the resulting image of the hand after applying derotation by angle α. b) The top row of images are un-derotated. The bottom row have been derotated by α obtained by DeROT. Note that the thumb is consistently on the right of the image.

Fingertip detection. In this work we specifically focus on per frame fingertip detection in depth images without either tracking or kinematic modeling. We propose useful modifications to the popular machine learning based methods of Keskin *et al*. [3] and Tompson *et al*. [4]. Our preprocessing step then involves cropping input images of hands and rotating them about their center of mass using the predicted angle of derotation produced by DeROT.

The HandNet database No currently available hand data-sets (*e.g.* [5],[2],[4]) include accurate full 3 DOF ground truth hand orientations on a large database of real depth images. A significant contribution of this paper is therefore the creation of a new, large-scale database of fully annotated depth images of hands that we call HandNet[1]. To build and annotate this database we use a RealSense camera combined with 2*mm* TrakStar [1] DC magnetic trackers. Sensor slippage is prevented by affixing the sensors inside tight elastic seams. The close fitting pockets also prevent the hand depth profile from being affected by the attached sensors. To calibrate between the camera and sensor frames we position the magnetic sensors on the corners of a checkerboard pattern to create physical correspondence between the detected corner locations and the actual sensors. The setup can be seen in Figure 2. Sensors are modeled as 3D oriented ellipsoids and ray-cast into the camera frame. Discrete fingertip labels as well as heat-maps and orientation information are then trivially associated with each input image. The database is created from 10 participants in total who perform random hand motions with extensive pose variation. In total, after quality filtering, HandNet contains 212928 unique poses which is to the best of our knowledge the largest annotated, front facing database of real hand images currently available.

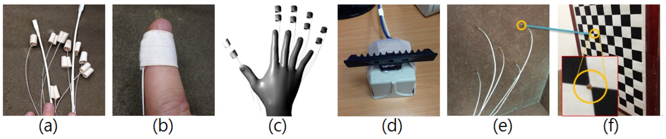

Figure 2: The data capture setup. a) 2mm magnetic sensors. The larger rectangular sensors are not used. b) A fingertip sensor inside the inner seam. c) Virtual model used for planning a multi-sensor setup. We only use 5 sensors. d) The RealSense camera rigidly fixed to the TrakStar transmitter. e) The back of the wooden calibration board where the glass sensor housings are firmly pushed through. f) The front of the calibration board where the glass sensor housings are visible on the corners as seen in the inset.

Experiments We perform our experiments using our HandNet database and the publicly available database NYUHands [4]. All experiments are performed separately on the two data sets. Using our deep derotation method we show up to 20.5% improvement in mean average precision (mAP) over our baseline results for both fingertip detection methods. We also compare our results to a non-learning based method similar to PCA and show that it produces inferior results. This shows that using derotation, specifically DeROT, significantly improves the localization ability of machine-learning based per-frame fingertip detectors by reducing the variance of the observed image space. Furthermore we find that this procedure works despite the extremely high range of potential poses. We see this approach as an alternative to data augmentation and as a useful preprocessing step in pipelines dedicated to articulated object pose extraction such as hands. For our experiments a single random decision tree mostly outperforms a convolutional neural network. Although they are trained with different data and objectives it hints that there is no silver bullet to determining which machine learning approach is more appropriate.

Acknowledgments This research was supported by European Community's FP7- ERC program, grant agreement no. 267414.

[1] Ascension TrakStar. `http://www.ascension-tech.com/`, 2015.

[2] Q. Chen, S. Xiao, W. Yichen, T. Xiaoou, and S. Jian. Realtime and robust hand tracking from depth. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1106–1113. IEEE, 2014.

[3] C. Keskin, F. Kiraç, Y. Emre Kara, and L. Akarun. Real time hand pose estimation using depth sensors. In *Consumer Depth Cameras for Computer Vision*, pages 119–137. Springer, 2013.

[4] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (TOC)*, 33, 2014.

[5] W. Zhao, J. Chai, and Y. Xu. Combining marker-based mocap and rgb-d camera for acquiring high-fidelity hand motion data. In *Symposium on Computer Animation*, pages 33–42, 2012.

[1] To advance research in the field this database and relevant code is available at `www.cs.technion.ac.il/~twerd/HandNet/`.