Machine Learning Nanodegree
Capstone Proposal

SHEN, WEN-JIE

August 21, 2017

1 Domain Background

In many years, hand pose estimation is a difficult problem with many potential applications in human computer interaction, gesture understanding and augmented. Fortunately, since the widespread success of real-time human body pose estimation (Zhang, 2012), many method has appeared in the area of hand pose estimation.

Even with depth-sensors such as kinect, hand pose estimation is still problematic because the hand has self-similarity and self-occlusions in input as 2D image.

Considering real-time hand tracking, a popular type of approach is based on generative, model-based tracking methods.For example, particle swarm optimization is able to handle a lot of parameter for the estimating (Qian et al., 2014). This approach is fast but the approach needs the guarantee of converging therefore the initialization is very important to the approach. That is to say, the approach will lost tracking when the hand is moving fast.

Another type of approach directly predicts the locations of the joints from RGB or RBGD images. Random Forests might be the famous algorithm in this approach, it is used to perform a classification of depth images to label which pixel is joint (Keskin et al., 2012; Shotton et al., 2013). In order to infer the location of hidden joints, hierarchical model of hand can classifies viewpoint, then individual joints and to finally predict 3D locations of joints (Tang et al., 2013, 2014).

Applying Deep Learning to solve the task of Computer Vision is current trend, and a Convolutional Neural Network with a standard architecture performs well to be applied to this problem. One method used a CNN for feature extraction and infer the hand pose using inverse kinematics to finally generate a heat-map for the locations of joints (Tompson et al., 2014). Another method directly predict and refine the 2D joint locations in RGB images (Toshev and Szegedy, 2013).

More recently, scientists observe that the structure of the network is very important therefore they investigate different architectures to find the most appropriate one for the hand pose estimation problem (Oberweger et al., 2015; Sinha et al., 2016).

2 Problem Statement

This project will estimate the 3D hand model therefore the locations of joints is the output form a single 2D RGBD image.

$$j_{i=1} \in J, j_i = (x_i, y_i, z_i) \tag{1}$$

Therefore, a training set of depth images should be labeled with the 3D joint locations. The input is a small region that centered on the center of mass of the object as hand in depth images for simplifying this task.

3 Datasets and Inputs

NYU Hand Pose Dataset (Tompson et al., 2014): The dataset contains over 72k training and 8k test frames of RGBD data captured using the Primesense Carmine 1.09. As mentioned above, this task doesn't consider segregating the hand from the rest of the body. NYU Hand Pose Dataset provided The RDF(Random Decision Forest) that has the ground truth per-pixel labels to solve the segregating problem.

4 Solution Statement

The training data is labeled therefore this task cloud be seem a supervised learning problem. However, the loss function is most important. Minimizing the distance between the prediction and the expected output per joint is popular to train a model and on top of that, dimensionality reduction using deep-learning (Sinha et al., 2016) is interesting too.

5 Benchmark Model

The benchmark models are available in the NYU Hand Pose Dataset (Tompson et al., 2014) that contains 8252 frames of captured RGBD data with ground-truth hand-pose information for validation.

6 Evaluation Metrics

The average Euclidean distance between the predicted 3D joint location and the ground truth. This method is sample and effective. The hand pose estimation is usually as a real-time application therefore the performance of prediction is a term of evaluation metrics.

7 Project Design

7.1 Data preparation

First, The hand just fills a relatively small region in the entire depth image under normal usage. Hence, the input depth image only includes values that lie in a range under the ground truth per-pixel labels that provided by NYU Hand Pose Dataset, followed by median filtering for noise removal, depth normalization, and re-size the image while maintaining the aspect ratio to obtain a small depth image.

7.2 supervised learning

Next step is to implement a architecture to present the features in CNN and minimizing the corresponding loss function.

7.3 Model evaluation and tuning

Finally, the model could be tuned to reduce the over-fitting and increase the performance of prediction after evaluation.

References

Cem Keskin, Furkan Kiraç, Yunus Emre Kara, and Lale Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *ECCV (6)*, volume 7577 of

Lecture Notes in Computer Science, pages 852–863. Springer, 2012. ISBN 978-3-642-33782-6. URL http://dblp.uni-trier.de/db/conf/eccv/eccv2012-6.html#KeskinKKA12.

Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Hands deep in deep learning for hand pose estimation. *CoRR*, abs/1502.06807, 2015. URL http://arxiv.org/abs/1502.06807.

Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake. Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2821–2840, Dec 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.241.

Ayan Sinha, Chiho Choi, and Karthik Ramani. Deephand: Robust hand pose estimation by completing a matrix imputed with deep features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

D. Tang, T. H. Yu, and T. K. Kim. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *2013 IEEE International Conference on Computer Vision*, pages 3224–3231, Dec 2013. doi: 10.1109/ICCV.2013.400.

D. Tang, H. J. Chang, A. Tejani, and T. K. Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3786–3793, June 2014. doi: 10.1109/CVPR.2014.490.

Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Trans. Graph.*, 33(5):169:1–169:10, September 2014. ISSN 0730-0301. doi: 10.1145/2629500. URL http://doi.acm.org/10.1145/2629500.

Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. *CoRR*, abs/1312.4659, 2013. URL http://arxiv.org/abs/1312.4659.

Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE MultiMedia*, 19(2):4–10, April 2012. ISSN 1070-986X. doi: 10.1109/MMUL.2012.24. URL http://dx.doi.org/10.1109/MMUL.2012.24.