A NOVEL FRAMEWORK OF HAND LOCALIZATION AND HAND POSE ESTIMATION

Yunlong Che[1], *Yuxiang Song*[1], *Yue Qi*[1,2,3]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
[2]Peng Cheng Laboratory, Shenzhen
[3]Qingdao Research Institute of Beihang University

ABSTRACT

In this paper, we propose a novel framework for hand localization and pose estimation from a single depth image. For hand localization, unlike most existing methods that using heuristic strategies, *e.g.* color segmentation, we propose Hierarchical Hand location Networks (HHLN) to estimate the hand location from coarse to fine in depth images, which is robust to the complex environment and efficient. It first applied at a low-resolution octree of the whole depth image and produced coarse hand region and then constructs the hand region into a high-resolution octree for fine location estimation. For pose estimation, we propose Wide Receptive-filed (WR-OCNN) which is able to capture meaningful hand structure in different scales and estimate the 3D hand pose accurately. Experiments on two widely-used hand datasets(NYU dataset and ICVL dataset) demonstrate the effectiveness and superiority of the proposed framework.

Index Terms— Hand Location, Hand Pose Estimation, Octree-based CNN,

1. INTRODUCTION

In recent years, hand localization and pose estimation are becoming increasingly important in the fields of human-computer interaction and computer vision [1]. With the advent of RGB-D cameras such as Kinect and Intel RealSense, many depth-based methods [2, 3] have been proposed to accomplish such challenging tasks.

Hand localization aims to estimate the center of hand from depth images with complicated environments., it plays a key role in the further analysis. For example, pose estimation largely benefits from an accurate hand location *et.al* [4]. Previous works rely either on skin color based detector or existing human skeleton trackers. Thompson *et.al* [5] use a random forest to classify each pixel into hand or background. Tagliasacchi *et.al* [6] localize and segment the hand based on a wirstband. However, the above methods are based on simple assumptions, *e.g.* the hand appears largest in front of the sensor or the wristband can be identified by color segmentation. These assumptions would be further from real scenarios with cluttered backgrounds. Recently, Some deep based methods have been presented for hand localization. Choi *et.al* [7] proposed a localization network for hand localization. However, it considers the whole scene of high-resolution input, thus suffers from the processing speed. Chen *et.al* [8] proposed a framework which integrates hand detection and pose estimation. But it uses the 2D object detection method, which is limited from depth data.

In this paper, we propose a Hierarchical Hand Localization Networks (HHLN) to estimate the hand location from coarse to fine in a single depth image. Concretely, we build a hierarchical structure that first applied at a low-resolution octree of the whole image to produce a coarse hand region and then constructs a high-resolution octree based on the region for fine location estimation based on DeconvNets. By using the hierarchical structure that processes the low resolution, the HHLN is robust to real scenarios and efficient for analyzing 3D hand shape.

Given the hand image region, pose estimation targets to evaluate the joints' locations of hand. There are two categories: **generative methods** and **discriminative methods**. The generative ones use the 3D hand model to approximate the point cloud, by optimizing a pre-defined energy function to estimate the hand pose. Oikonomidis *et.al* [9] fit a spheres/cylinders hand model to the depth image with particle swarm optimization(PSO). Qian *et.al* [10] used an ICP-PSO method to find the closest hand model parameters that match the observed data. Generative methods are more computation complexity than discriminative methods and easily prone to local minima due to the fast hand motion and complex structure. The discriminative ones tend to learn a regression function from training data, mapping the appearance in depth images to hand pose. Ge *et.al* [11] represented hand with a set of images rendered from three views and fed into a 2D CNN for pose estimation. Zhou *et.al* [12] embed joint rotation constraint into CNN to boost the accuracy.

Recently, many 3D CNN methods have been presented for 3D shape analysis. O-CNN [13] is one of the successful methods based on octree and efficient computation on high-resolution volumes. In this paper, we employ the O-CNN to estimate the hand pose. However, it fails to model complex structures of multiple scales which shows better performance as stated in [14]. To alleviate the problem and in-

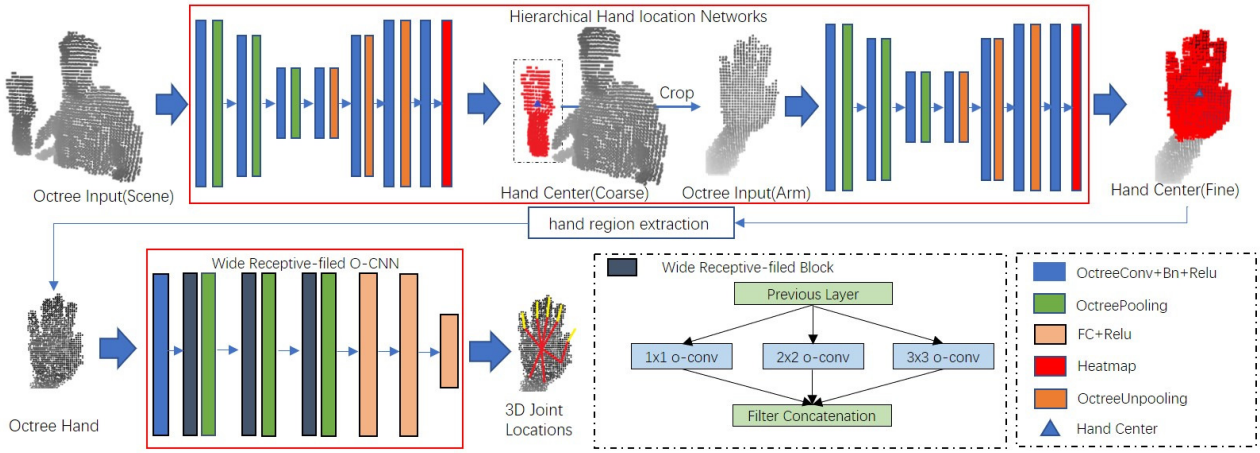

Fig. 1. Overview of the our framework. Given a depth image, we first convert it to an octree structure of low resolution and fed into the Hierarchical Hand Location Network to produce the hand region from coarse to fine. Then, we use a multi-views extremes points based method to extract the hand region. Finally, the Wide Receptive-field Octree-based CNN takes the octree hand region as input and output the 3D hand pose.

spired by the capacity of Inception structure [14], we extend O-CNN by broadening the receptive field to capture multi-scales hand structures. More specifically, we add additional convolution streams with different kernels in each convolution layer, thus build the Wide Receptive-field Octree-based CNN (WR-OCNN).

As shown in Fig.1, we combine HHLN and WR-OCNN to construct a joint framework. We use a low-resolution octree to represents the depth image and fed into the HHLN to estimate the position of hand center, which represents by a likelihood map. To robustly extract hand region, a multi-views extreme points based method is proposed to extract hand region. Subsequently, the WR-OCNN takes hand octree as input and regresses the hand joint locations. Note that with the help of the HHLN, our framework could process the whole depth image, thus improves the speed. We conduct a set of experiments on two public datasets to evaluate our framework and the results demonstrate its effectiveness.

2. HAND LOCALIZATION AND POSE ESTIMATION

Our framework consists of hand localization and pose estimation. In this section, we will describe the components of this framework, *i.e.* the Hierarchy Hand Location Network (HHLN) to locate the hand region and the Wide Receptive-filed OCNN (WR-OCNN) to estimate the hand pose.

2.1. Hand Localization

Our HHLN is able to locate the hand from coarse to fine in a cluttered background. In particular, the HHLN processes the depth image in a hierarchical way, which consists of two stages. In the first stage, we constructed a low-resolution octree from the whole depth image and fed into the DeconvNets [1] to produce the coarse hand region. In the second stage, based on the coarse hand region, we constructed a high-resolution octree and fed into another DeconvNets to estimate the fine hand region. Different from the DeconvNets that is used for part shape segmentation in [13], we aim to estimate the per-voxel likelihood heatmap for hand center. In each stage, the hand region is cropped from depth image according to the heatmap where the value is larger than a pre-defined threshold. Like [15], the per-voxel likelihood heatmap is computed as follows:

$$H_c^*(i,j,k) = exp(-\frac{(i-i_c)^2+(j-j_c)^2+(k-k_c)^2}{2\sigma^2}) \quad (1)$$

where (i,j,k) is the center coordinate of each voxel and (i_c,j_c,k_c) is the center coordinate of hand.

As shown in Fig.1 (top row). The HHLN consists of a pair of DeconvNets that applied at low-resolution octee of the whole depth image and high-resolution octree of coarse hand region. Both of them have the same structure, which cascades a deconvolution network after a convolution network. The convolution network has four octree convolution layers with the kernel sizes are 3,3,3 and 1. The deconvolution network is the mirror of convolution network where the convolution and pooling operators are replaced by deconvolution and unpooling operators. After the last octree deconvolution layer, we adopt the mean square error as a loss function L as follows:

$$L = \sum_{n=1}^{N} \sum_{i,j,k} ||H_c^*(i,j,k) - H_c(i,j,k)|| \quad (2)$$

where H_c^* and H_c are the ground-truth and estimated heatmaps for hand center.

[1]The DeconvNets mentioned in this paper refers to the version in [13], aiming to 3D part shape segmentation.

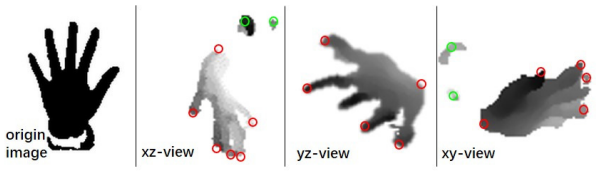

Fig. 2. The illusion of hand region extraction. The red circles are extreme points and the green ones are the outliers

In addition, most deep learning based approaches extract a fixed-size metric cube from images around the hand. However, directly resizing the input image will change the original topology. Inspire by XY-fingers [10], we also propose a multi-views extreme points based method to find an optimal hand region in 3D space.

We build a bounding box which centered at max likelihood position and contains nearby points. Then we project these points to xy, xz, yz plane to produce three images based on projective directional Truncated Signed Distance Function(TSDF) [16]. For each image, we start from the center position and compute its 3D geodesic distance map to all pixels using distance transform. Then add the maximal voxel in the geodesic distance map as a new extreme point and update the distance map in an incremental manner. After repeated the process for N times and removed the outliers (usually the "flying pixels"), we finally obtain K extreme points. Fig.2 shows the extreme points of the hand. The hand region is then represented by points within the 3D contour composed by extreme points.

2.2. Hand Pose estimation

After extracted the hand region using HHLN, we build WR-OCNN (shown in the bottom row of Fig.1) that takes octree of hand region as input to estimate the joints' locations of hand.

Since the Inception block [14] has been widely used for its competence of learning various scale features. We then design a wide receptive-field block (shown in Fig.1) and apply it into O-CNN, enabling WR-OCNN to capture more complex structure of hand. The receptive-field block is a combination of three convolution streams with specific kernels, *i.e.* 1×1, 3×3 and 2×2. All those layers with their output filter banks are concatenated into a single output vector. The WR-OCNN starts with three wide receptive-field blocks, each block follows a max pooling layer. After that, the extracted feature maps are fed into three fully-connected layers to regress 3D hand joint locations. We use the Minimum Square Error loss to train the WR-OCNN.

We also employ a data augmentation method to increase generalization of the network. Specifically, we apply random rotation along z axis with the range of $[-15°, 15°]$ and scaling displacement [0.8, 1.2] to the constructed octree.

3. EXPERIMENTS

In this section, we first depict our implementation details and then show the comparison of ours and other state-of-the-art ones.

3.1. Implementation details

We train and evaluate our networks on a PC with Intel Core i7 6700K, 32GB of RAM and a Nvidia 1080-Ti GPU. Models are implemented within the caffe framework [17]. When training the HHLN, we use Adam optimizer with learning rate 0.005, batch size 8, weight decay 0.0005. For WR-OCNN, we set the learning rate to 0.01, batch size 32, the other parameters are the same as HHLN.

We evaluate the hand location performance using 3D distance error between the hand center and ground truth. Following [7], we use middle finger' root as hand center. To evaluate the hand pose estimation, we employ two popular metrics, *i.e.* per-joint mean error distance overall test frames and the proportion of test frames whose maximum error falls below a threshold.

Baselines: To validate the effective of the counterparts of the proposed framework, we create several baselines:

1)B1: We use a single DeconvNets instead of the hierarchical structure that takes the whole scene of high resolution as input to estimate the hand region.

2)B2: We replace the HHLN in our framework and follow the hand localization in [13] to estimate the hand pose.

3)B3: We use the same convolution structure in [13] rather than Wide Receptive-field block in our framework to estimate the hand pose.

3.2. Results on NYU dataset

NYU dataset [5] contains 72,757 training frames and 8,252 testing frames, which are original depth images with complicated environment. On NYU dataset, we follow [11] and use a subset of 14 hand joints. We compare our method to three state-of-the-art methods, including DeepPrior[18], Feedback[3] and DeepModel[12]. As can be seen in Fig.3, our method outperforms those methods in terms of the error thresholds. The mean error distance for all joints of our method is 15.62mm, which is 2mm smaller than the results of DeepModel and 5mm smaller than the results of DeepPrior. Compared with B2, our framework shows better performance, indicating the effectiveness of the combination of hand localization and pose estimation. The performance gain is more obvious w.r.t. B3, showing the capacity of out WR-OCNN for capturing more complex hand structure.

In Table 1, we compare our HHLN with Choi *et.al* [7] and B1 in terms of the mean distance error overall test frames. The performance gain is more obvious w.r.t. B1 and [7], indicating the key role of our hierarchical localization structure. Meanwhile, HHLN shows faster process speed.

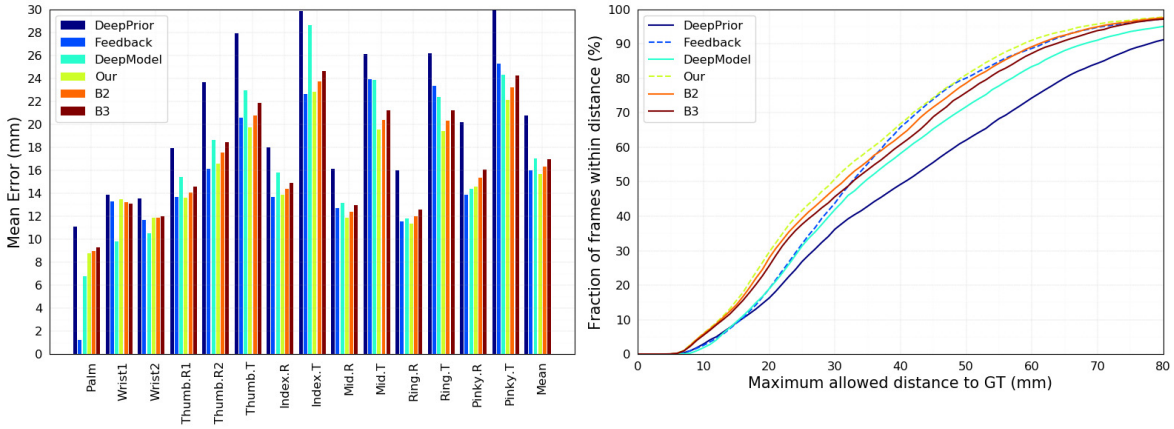

Fig. 3. Comparison of our method and others on NYU dataset.

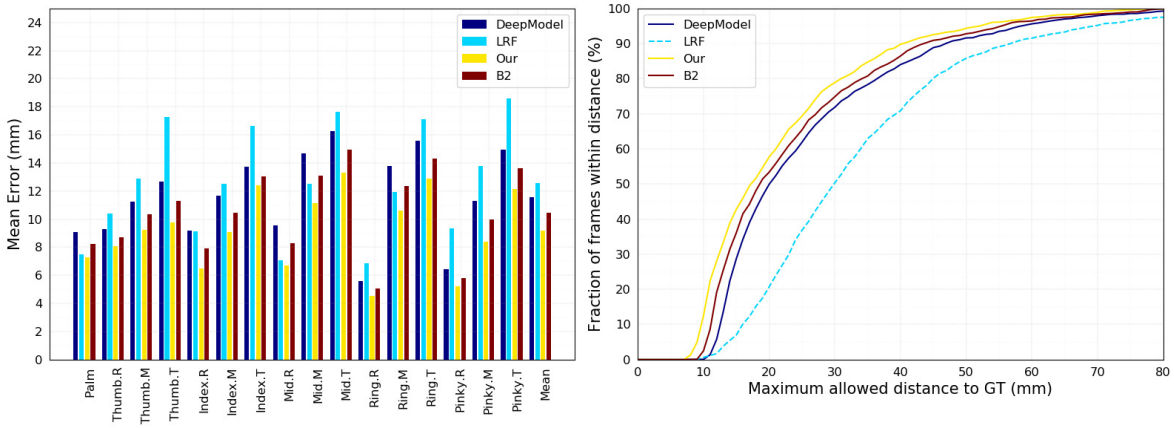

Fig. 4. Comparison of our method and others on ICVL dataset.

Table 1. Comparison of different methods for hand localization.

Method	run time(ms)	mean error(pixels)
B1	52	8.06
Choi [7]	48	13.62
HHLN	**29**	**6.29**

3.3. Results on ICVL dataset

ICVL dataset [19] contains 33K training frames and 1.6K testing frames, each of which contains 16 joints. In the ICVL dataset, hand region is centered in the frame, we thus do not need to perform the hand localization and only report the result of hand pose estimation. On ICVL dataset, we compare our method to two state-of-the-art methods, *i.e.* DeepModel[12], LRF [19]. As shown in Fig.4, when the error threshold is 10mm, the proportions of good frames of our method is higher than other methods. In terms of the mean error distances, our method outperforms state-of-the-art methods on most of the hand joints and achieves the smallest overall mean error distances.

4. CONCLUSIONS

In this paper, we propose a novel framework for hand localization and pose estimation. For the former, we proposed a hierarchy hand location network to estimate the hand region from coarse to fine. Thanks to the hierarchical structure, our HHLN is robust to the complex environment and efficient. For the latter, we build Wide Receptive-filed(WR-OCNN) which is able to capture meaningful hand structure in different scales. Experimental results on two public hand pose datasets show the effectiveness of the proposed framework.

5. ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 61572054, in part by the National Key R&D Program of China under Grant 2017YFB1002602, and in part by the Applied Basic Research Program of Qingdao under Grant 16-10-1-3-xx.

6. REFERENCES

[1] Ali Erol, George Bebis, Mircea Nicolescu, Richard D. Boyle, and Xander Twombly, "Vision based hand pose estimation: A review," *Computer Vision and Image Understanding*, vol. 108, no. 12, pp. 52–73, 2007.

[2] Stan Melax, Leonid Keselman, and Sterling Orsten, "Dynamics based 3d skeletal hand tracking," in *ACM SIGGRAPH Symposium on Interactive 3d Graphics and Games*, 2013, pp. 184–184.

[3] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit, "Training a feedback loop for hand pose estimation," in *IEEE International Conference on Computer Vision*, 2015, pp. 3316–3324.

[4] Markus Oberweger and Vincent Lepetit, "Deepprior++: Improving fast and accurate 3d hand pose estimation," in *IEEE International Conference on Computer Vision Workshops*, 2018, pp. 585–594.

[5] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *Acm Transactions on Graphics*, vol. 33, no. 5, pp. 1–10, 2014.

[6] Andrea Tagliasacchi, Matthias Schroder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly, "Robust articulated-icp for real-time hand tracking," *Computer Graphics Forum*, vol. 34, no. 5, pp. 101114, 2015.

[7] Chiho Choi, Sangpil Kim, and Karthik Ramani, "Learning hand articulations by hallucinating heat distribution," in *IEEE International Conference on Computer Vision*, 2017, pp. 3123–3132.

[8] Tzu Yang Chen, Min Yu Wu, Yu Hsun Hsieh, and Li Chen Fu, "Deep learning for integrated hand detection and pose estimation," in *International Conference on Pattern Recognition*, 2017, pp. 615–620.

[9] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis Argyros, "Efficient model-based 3d tracking of hand articulations using kinect. bmvc 2011," in *BMVC*, 2011.

[10] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun, "Realtime and robust hand tracking from depth," in *Computer Vision and Pattern Recognition*, 2014, pp. 1106–1113.

[11] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann, "3d convolutional neural networks for efficient and robust hand pose estimation from single depth images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5679–5688.

[12] Xingyi Zhou, Qingfu Wan, Wei Zhang, Xiangyang Xue, and Yichen Wei, "Model-based deep hand pose estimation," in *International Joint Conference on Artificial Intelligence*, 2016, pp. 2421–2427.

[13] Peng Shuai Wang, Yang Liu, Yu Xiao Guo, Chun Yu Sun, and Xin Tong, "O-cnn: octree-based convolutional neural networks for 3d shape analysis," *Acm Transactions on Graphics*, vol. 36, no. 4, pp. 72, 2017.

[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014.

[15] Gyeongsik Moon, Juyong Chang, and Kyoung Mu Lee, "V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] Shuran Song and Jianxiong Xiao, "Deep sliding shapes for amodal 3d object detection in rgb-d images," in *Computer Vision and Pattern Recognition*, 2016, pp. 808–816.

[17] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[18] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit, "Hands deep in deep learning for hand pose estimation," *Computer Science*, 2016.

[19] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae Kyun Kim, "Latent regression forest: Structured estimation of 3d articulated hand posture," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3786–3793.