NETWORKS EFFECTIVELY UTILIZING 2D SPATIAL INFORMATION FOR ACCURATE 3D HAND POSE ESTIMATION

Baoen Liu, Shiliang Huang, Zhongfu Ye

National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China,
Hefei, 230026, China

ABSTRACT

In this work, we propose a new method for accurate 3D hand pose estimation from a single depth map using convolutional neural networks (CNN). Our method effectively makes use of 2D spatial information to improve the performance by two means. Firstly, we formulate 3D hand pose estimation as a two-task (2D joints detection and depth regression) problem so that we can directly utilize the ability of hourglass module on processing multi-scale information for estimating 2D joint coordinates. Secondly, 2D spatial information is used to help depth regression by introducing the spatial attention mechanism to our method. The experimental results demonstrate that our method achieves the state-of-the-art performance on ICVL hand posture dataset and a comparable performance with the state-of-the-arts on NYU dataset.

***Index Terms*—** CNN, hand pose estimation, multi-task learning, spatial attention mechanism

1. INTRODUCTION

Hand pose estimation is a challenging task in human-computer interaction. As the appearance of consumer depth cameras, such as Microsoft Kinect and Intel RealSense, accurate 3D hand pose estimation based on a singe depth image has attracted broad research interest. In recent years, a lot of advanced methods have been proposed and therefore the hand pose estimation has made a significant progress.

In previous hand pose estimation methods, the CNN-based methods represent an important branch. For better performance on estimating the 3D coordinates of hand joints, these methods usually focus on more powerful data representation [1, 2, 3], additional refinement [4, 5, 6, 7] and well-designed CNN architecture [1, 2, 7, 8, 9]. However, due to the complex physical structure and self-occlusions, accurate hand pose estimation is still a challenging problem.

In this work, We propose a simple but highly effective CNN-based method for accurate 3D hand pose estimation from a single depth map (see Fig.1). Our CNN model is based on the SE-ResNet [10] which has powerful representational capacity. The proposed method is advanced in utilizing 2D spatial information by two means. Firstly, we formulate the hand pose estimation as a two-task (2D joints detection and depth regression) problem. Then the 2D joint coordinates can be directly regressed with hourglass modules [11] which process features across all scales and fantastically capture the various spatial relationships associated with human hand. Secondly, in order to make better use of the 2D spatial information, the spatial attention mechanism is applied to our model to improve the features for depth regression. To promote the performance, an initial CNN model for hand location estimation is exploited. Our method is evaluated on two challenging hand posture datasets, ICVL [12] and NYU [13]. The experimental results demonstrate that our method achieves state-of-the-art performance on ICVL dataset. On NYU dataset, our method outperforms all the previous methods for a large margin except the V2V-PoseNet [3].

2. RELATED WORK

The conventional methods for hand pose estimation can be categorized into generative, discriminative, and hybrid methods. The review of these methods can be found in [7]. In this paper, we mainly talk about the CNN-based methods. Compared with other discriminative methods which utilize hand-craft features, the CNN-based methods combine feature extraction and hand pose estimation in one holistic procedure. In the past few years, more and more CNN-based methods have been proposed. Tompson et al. [13] use a CNN model to produce 2D heat maps and infer the 3D hand pose with inverse kinematics. Oberweger et al. [4] directly estimate the 3D coordinates of hand joints with multi-stage CNN using a linear layer as pose prior. In [6], an improved version of work in [4] is proposed with hand location refinement, data augmentation and more powerful network architecture. Ge et al. [1] employ 3 CNN models to separately regress 2D heat maps for each view with depth projections and fuse them to produce 3D hand pose. In [2], Ge et al. transform the

This work was supported by the Fundamental Research Funds for the Central Universities (Grant no. WK2350000002)

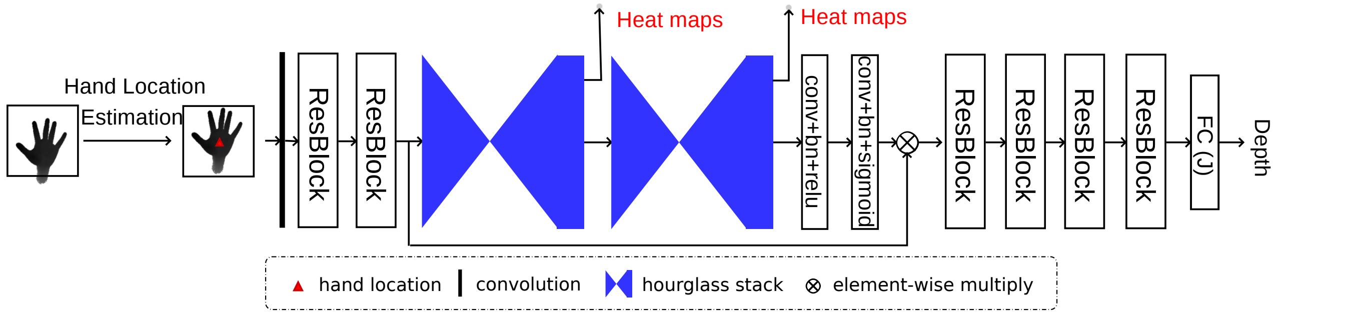


Fig. 1. Overview of our method. We choose 128*128 as size of input depth map and 64*64 as the input and output resolution of hourglass module with 128 feature channels in each layer. For consideration of the trade-off between performance and computation, we use 2 hourglass stacks for 2D joints detection. Figure is best viewed in color.

2D input depth map to the 3D form and exploit 3D CNN to estimate 3D coordinates of hand joints. Guo et al. [8, 14] propose a region ensemble network to accurately estimate the 3D coordinates of hand joints. Chen et al. [7] improve region ensemble network by cropping regions with joint locations guide and iteratively refining the estimated pose. Wan et al. [15] formulate 3D hand pose estimation as a dense regression and their method works on dense pixel-wise estimation. To overcome the weaknesses of directly regressing the 3D coordinates of hand joints with 2D depth map, Moon et al. [3] cast the 3D hand and human pose estimation problem into a voxel-to-voxel prediction that uses a 3D voxelized grid and estimates the per-voxel likelihood for each joint.

Most of the conventional CNN-based methods estimate 2D or 3D coordinates of hand joints in one-task setup. In contrast, some work on human pose estimation, a similar vision task, couple 2D joints detection and 3D regression in multi-task setup [16, 17]. Similar with [15], we adopt this idea for our model design and formulate the 3D hand pose as a two-task problem. Different from the previous sophisticated multi-task model for pose estimation, the multi-task design in our method is simple but highly effective.

3. OVERVIEW OF OUR METHOD

An overview of the architecture of our method is illustrated in Fig.1. We adopt the Squeeze-and-Excitation (SE) residual block [10], with dynamic channel-wise feature recalibration in residual branch, as our basic building block and simply denote it as *ResBlock* in the Figures. The structure of stacked hourglass networks used to detect 2D joints in our method is the same as illustrated in [11].

3.1. Hand Location Estimation

As the discussions in [6], the location of hand can greatly affect the performance of hand pose estimation. A poor hand location, such as center of mass (CoM) simply obtained from the rough hand region, will be harmful to the regularity of model input. In contrast, cropping the depth maps with more accurate hand locations will decrease the variance of data which will be fed into the pose estimation model, and significantly improve the performance.

The networks we used for hand location estimation is shown in Fig.2. The input depth map of this model is cropped according to a cube centered at the CoM of rough hand region in 3D space. In the following step, the estimated hand location will be used to help obtain better cropped depth maps for pose estimation.

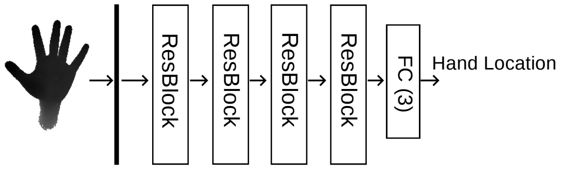


Fig. 2. Architecture of the hand location estimation networks.

3.2. Hand Pose Estimation

Different from most of conventional hand pose estimation methods, we consider the 3D hand pose estimation as a two-task (2D hand joints detection and depth regression) problem and couple these two tasks in one model. Since the hourglass module was proposed, it has been widely used and proven to be highly effective in processing 2D spatial information. Therefore, with the multi-task setup, we can easily use the powerful ability of hourglass module to process all-scale 2D spatial information for 2D joints detection with heat maps. Additionally, the 2D spatial information is applied to help depth regression.

Our model starts with a few layers which are used to extract the shallow features. After that, the extracted feature maps will be fed into the 2D joints detection modules (hourglass stacks) and the joint depth regression modules. The loss function of the whole model can be written as:

$$L = \sum_{i}^{N} L_{det}^{i} + L_{reg} \tag{1}$$

where L_{det}^{i} denotes the loss of 2D joints detection in i-th hourglass stack, L_{reg} denotes the loss of depth regression, and N is the number of hourglass stacks.

3.2.1. 2D Joints Detection

In 2D joints detection, we adopt the hourglass modules to capture the spatial relations of the parts of human hand. The motivation for this module design is to utilize all-scale information. Similar with [11], we stack the 2D joint estimation modules and apply intermediate-supervision. The output of 2D joints detection module is J (number of joints) heat maps. The loss function L_{det}^{i} of i-th hourglass stack is:

$$L_{det}^{i} = \sum_{j}^{J} L^2(H_o^j, H_g^j) \tag{2}$$

where H_o^j and H_g^j denote the output heat map and ground truth heat map for j-th joint. L^2 means L^2 distance.

3.2.2. Joint Depth Regression

For depth regression, a simple network is employed. Here we pay our attention to improving features for depth regression using 2D spatial information rather than sophisticated architecture design. Due to the fact that hand pose can be represented by the hand joints coordinates, we can just use the spatial information gained from 2D joints detection to enhance the information around the corresponding location in shallow feature maps. Based on this assertion, we introduce spatial attention mechanism to our work by innovatively using 2D spatial information to weight the element-wise values of shallow feature maps. Thus, the i-th input feature map of depth regression is:

$$F_{reg,i}(x) = M_i(x) * F_{raw,i}(x) \tag{3}$$

where $F_{raw,i}(x)$ denotes the i-th shallow feature map and $M_i(x)$ denotes the corresponding mask. $M_i(x)$ is obtained by normalizing the output feature maps of 2D joints detection modules (see Fig.1). The self-comparison experiments in the following section will demonstrate the effectiveness of this operation.

4. EXPERIMENTS

We apply our method on two public challenging hand posture datasets: ICVL [12] and NYU [13]. The augmented ICVL hand posture dataset we used contains 300K training images and 1.6K testing images with 16 joints. The later dataset NYU has 72K frames (each frame contains three depth images for three views) for training and 8K for testing. The annotation of NYU hand pose contains 36 joints. Following most of previous work, we only used frames from the frontal view and 14 out of 36 joints in evaluation. The performance is evaluated by two metrics [12, 18]: 3D Euclidean distance error (in millimeters) and percentage of frames in which all errors of joints are below a threshold. For consideration of the trade-off between performance and computation, we use 2 hourglass stacks in our model and each Resblock contains 2 convolutional layers and 1 SE block.

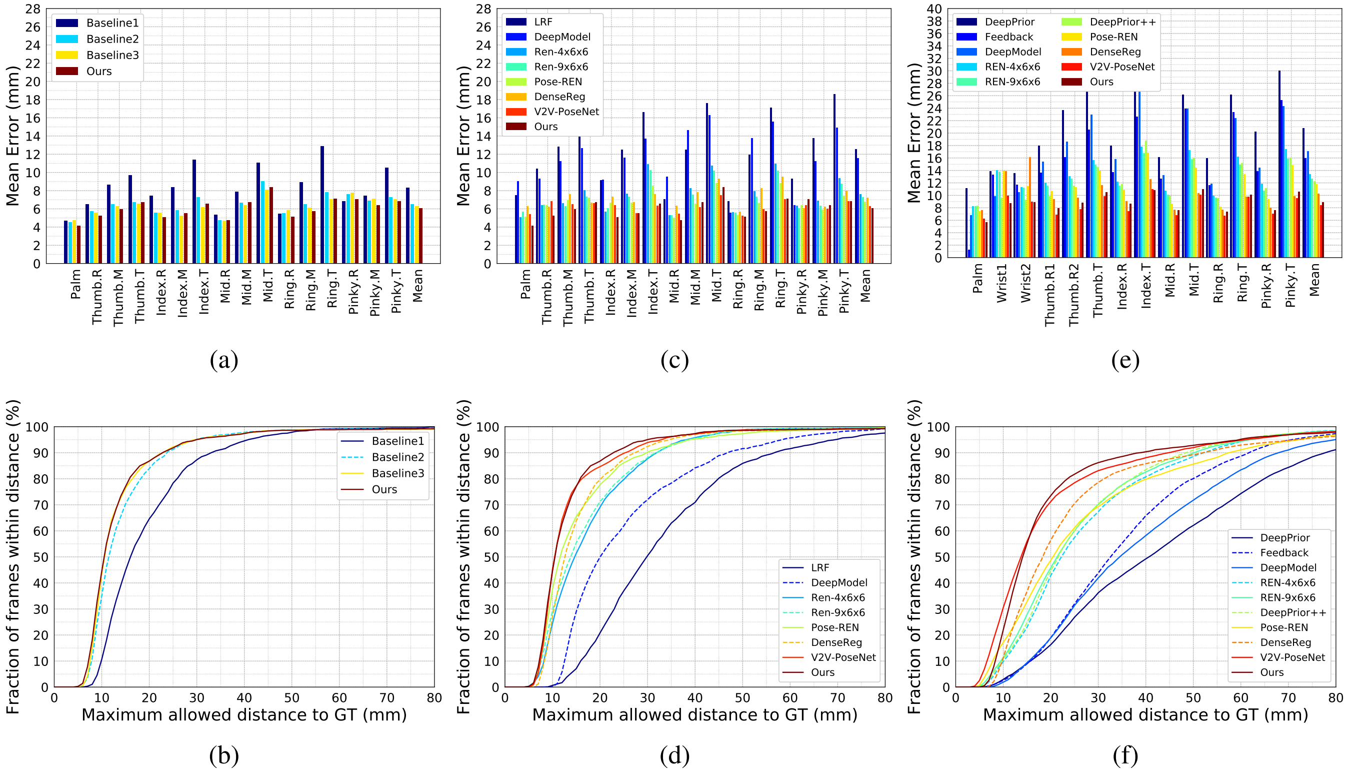
Our method is implemented with Pytorch. We use SGD with a mini-batch size of 32 to train our model in one NVIDIA GeForce GTX 1070 GPU with CUDA 8.0 and cuDNN 5.1. The weight decay is 0.00005 and the momentum is 0.9. We start from a learning rate of 0.1, and divide it by 10 for two times in the whole training procedure.

Table 1. Average 3D distance error (mm) of different methods.

Dataset	Comparison Methods	Average Joint Error(mm)
ICVL	DeepModel [19]	11.547
	Ren-4x6x6 [14]	7.618
	Ren-9x6x6 [8]	7.296
	Pose-REN [7]	6.783
	DenseReg [15]	7.232
	V2V-PoseNet [3]	6.276
	Ours	**6.092**
NYU	DeepPrior [4]	20.750
	Feedback [5]	15.973
	DeepModel [19]	17.036
	REN-4x6x6 [14]	13.393
	REN-9x6x6 [8]	12.694
	DeepPrior++ [6]	12.238
	Pose-REN [8]	11.811
	DenseReg [15]	10.214
	V2V-PoseNet [3]	**8.419**
	Ours	8.862

4.1. Self-comparisons

To demonstrate the advantages of our model design, we compare our method with three baselines: 1) The plain networks which estimate 3D hand pose in one-task setup. 2) Separate two-branch networks (depth regression with raw shallow feature). 3) Networks which directly use the output feature maps



(a)

(c)

(e)

(b)

(d)

(f)

Fig. 3. Comparisons for distance error (upper) and percentage of success frames (lower). (a) and (b) are the self-comparison results. (b) and (c) are the comparisons with state-of-the-arts on ICVL dataset. (e) and (f) are the comparisons with state-of-the-arts on NYU dataset.

of 2D joints detection modules for depth regression. All of these comparison experiments are on ICVL dataset. The Results of comparisons for the evaluation metrics mentioned before are shown in Fig 3. (a) and (b).

The results of baseline 1) and baseline 2) demonstrate that the multi-task setup which exploits the hourglass modules for 2D joints detection can significantly improve the performance. The results of baseline 2), baseline 3) and our method show the effectiveness of 2D spatial information in helping depth regression and the superiority of the attention mechanism used in our model.

4.2. Comparisons with State-of-the-arts

The comparisons for 3D average joint error on both ICVL ans NYU dataset are illustrated in Table.1. Fig.3 (c)-(f) show the comparisons for per-joint distance error and percentage of success frames. As the experimental results, our method achieves state-of-the-art performance on ICVL dataset. On NYU dataset, although not as good as the performance of [3] in average 3D distance error, our method outperforms various other proposed methods for a large margin. Additionally, our method achieves the best performance when the error threshold is larger than 15mm.

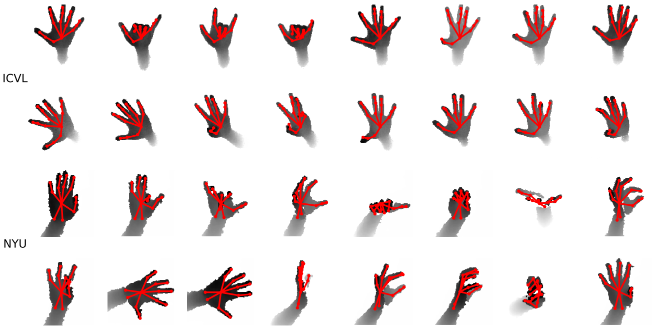


Fig. 4. Examples of our estimated hand pose.

5. CONCLUSIONS

In this paper, we propose a method effectively utilizing 2D spatial information for accurate 3D hand pose estimation. Our method adopts multi-task setup and directly utilizes stacked hourglass networks for 2D hand joints detection. In addition, 2D spatial information is applied to promote the performance of depth regression by introducing spatial attention mechanism to our method. In the future, we will continue to improve our method and extend it to related vision tasks.

6. REFERENCES

[1] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann, "Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3593–3601.

[2] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, vol. 1, p. 5.

[3] Gyeongsik Moon, Ju Yong 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," *arXiv preprint arXiv:1711.07399*, 2017.

[4] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit, "Hands deep in deep learning for hand pose estimation," *arXiv preprint arXiv:1502.06807*, 2015.

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

[6] Markus Oberweger and Vincent Lepetit, "Deepprior++: Improving fast and accurate 3d hand pose estimation," in *ICCV workshop*, 2017, vol. 840, p. 2.

[7] Xinghao Chen, Guijin Wang, Hengkai Guo, and Cairong Zhang, "Pose guided structured region ensemble network for cascaded hand pose estimation," *arXiv preprint arXiv:1708.03416*, 2017.

[8] Hengkai Guo, Guijin Wang, Xinghao Chen, Cairong Zhang, Fei Qiao, and Huazhong Yang, "Region ensemble network: Improving convolutional network for hand pose estimation," *arXiv preprint arXiv:1702.02447*, 2017.

[9] Xiaoming Deng, Shuo Yang, Yinda Zhang, Ping Tan, Liang Chang, and Hongan Wang, "Hand3d: Hand pose estimation using 3d neural network," *arXiv preprint arXiv:1704.02224*, 2017.

[10] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, 2017.

[11] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.

[12] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim, "Latent regression forest: Structured estimation of 3d articulated hand posture," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3786–3793.

[13] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *ACM Transactions on Graphics (ToG)*, vol. 33, no. 5, pp. 169, 2014.

[14] Hengkai Guo, Guijin Wang, Xinghao Chen, and Cairong Zhang, "Towards good practices for deep 3d hand pose estimation," *arXiv preprint arXiv:1707.07248*, 2017.

[15] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao, "Dense 3d regression for hand pose estimation," *arXiv preprint arXiv:1711.08996*, 2017.

[16] Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu, "Deep multitask architecture for integrated 2d and 3d human sensing," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid, "Lcr-net: Localization-classification-regression for human pose," in *CVPR 2017-IEEE Conference on Computer Vision & Pattern Recognition*, 2017.

[18] Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun, "Cascaded hand pose regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 824–832.

[19] Xingyi Zhou, Qingfu Wan, Wei Zhang, Xiangyang Xue, and Yichen Wei, "Model-based deep hand pose estimation," *arXiv preprint arXiv:1606.06854*, 2016.