O2-2-4

17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Occlusion-Robust 3D Hand Pose Estimation from a Single RGB Image

Asuka Ishii
NEC Corporations
1753, Shimonumabe, Nakahara-ku, Kawasaki, Kanagawa 211-8666, Japan
as-ishii@nec.com

Gaku Nakano
NEC Corporations
g-nakano@nec.com

Tetsuo Inoshita
NEC Corporations
tetsuo.inoshita@nec.com

Abstract

We propose an occlusion-robust network for 3D hand pose estimation from a single RGB image. Severe occlusions degrade the estimation accuracy of not only occluded keypoints but also visible keypoints. Since the existing methods based on a deep neural network perform convolutions on all keypoints regardless of visibility, inaccurate features from occluded keypoints affect the localization of visible keypoints. To suppress the influence of occluded keypoints, our proposed deep neural network consists of three modules: a 2D heatmap generator, parallel sub-joints network (PSJNet), and an ensemble network (EN). First, the 2D position of all keypoints in an input image is predicted as a 2D heatmap, similar to the existing methods. Then, PSJNet, which consists of several graph convolutional networks (GCN) in parallel, estimates multiple incomplete 3D poses in which some of the keypoints have been removed. Each GCN performs convolutions on a limited number of keypoints, therefore, features from occluded keypoints do not spread to the whole pose. Finally, EN merges the incomplete poses into a single 3D pose by selecting accurate positions from them. Experimental results on a public dataset RHD demonstrate that the proposed method outperforms the existing methods in the case of both small and severe occlusions.

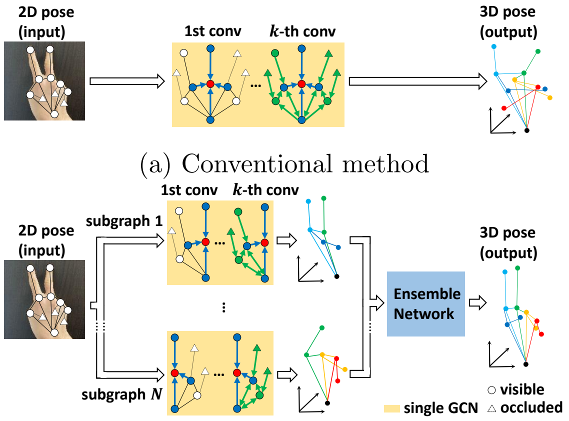
1 Introduction

3D hand pose estimation from images has been an important research topic for decades, as it can be widely used for many applications such as action recognition, sign language recognition, human-computer interaction, and virtual/augmented reality. An open issue is the occlusion handling in images. Hand motions or hand-object interactions lead to occlusions on keypoints due to movements of the palm or fingers. Since it is difficult to extract accurate features from occluded keypoints, severe occlusions significantly degrade the estimation accuracy.

A significant amount of reseach attention has been devoted to occlusions [1, 2, 3, 4], and large progress has been seen with recent advances in convolutional neural networks (CNN) [5, 6, 7, 8]. However, these works with CNNs require a depth image or video frames as the input, which are not always available in real applications. For example, a depth sensor can be used only if a hand is sufficiently close to it. Moreover, methods that require temporal information cannot be applied to a single image.

To avoid these limitations, 3D hand pose estimation from a single RGB image has been attracting interenst [9, 10, 11]. While these works improve the estimation accuracy, they do not address the occlusion issue.

If occlusions are not considerd, degradation of the accuracy of the occluded keypoints is inevitable. The existing methods aim to improve the estimation accuracy of occluded keypoints. However, we experimentally found that severe occlusions also degrades the accuracy of *visible* keypoints. More details are presented in Sec. 3.2. We hypothesize that this is because the existing methods based on a deep neural network (DNN) perform convolutions on all keypoints regardless of the visibility. Let us consider an example where a graph convolutional network (GCN) reconstructs a 3D hand pose from a 2D pose estimated by a CNN. Figure 1(a) illustrates how the 3D position of a visible keypoint (denoted by a red circle) is determined by other visible and occluded keypoints (shown in circles and triangles, respectively). The existing networks extract features from nodes at a distance of k via the neighboring nodes by performing convolutions k times, the flows of which are indicated by green arrows. Thus, inaccurate features from occluded keypoints are used to localize the 3D position of visible keypoints. As a result, the estimation of visible keypoints is corrupted. There are two approaches to overcoming this issue: (i) improving the CNN for 2D pose estimation on occluded keypoints, and (ii) suppressing the influence of occluded keypoints when reconstructing a 3D pose from a 2D pose. The existing methods discussed above try to solve the occlusion issue by taking the first approach. In contrast,



(a) Conventional method

(b) Proposed method

Figure 1: Visual comparison of the proposed and conventional methods. Blue and green arrows indicate feature extractions by the first and k-th convolutions on neighboring nodes, respectively. (a) Since the conventional method performs convolutions on all keypoints regardless of visibility, inaccurate features from occluded keypoints affect the localization of visible keypoints, which collapses the final 3D pose. (b) N subnetworks in the proposed method perform convolutions on a limited number of keypoints, so features from occluded keypoints do not spread to the whole pose. EN merges N incomplete 3D poses into a single 3D pose.

Table 1: Architecture of the feature encoder. S1 or S2: strides for a 3×3 convolution. J: the number of keypoints. m: the number of removed keypoints. D: the number of dimensions of a GCN, *i.e.* $D = 2$ in PSJNet and $D = 3$ in EN. See Fig. 2 for J, m, and D.

Layer	Dims (C × H × W)
Feature map	$256 \times 64 \times 64$
Conv3 × 3(S2)–BN–ReLU	$256 \times 32 \times 32$
Conv3 × 3(S1)–BN–ReLU	$256 \times 32 \times 32$
Conv3 × 3(S1)–BN–ReLU	$D(J - m) \times 32 \times 32$
Global average pooling	$D(J - m)$

we examine the second approach in this paper.

To this end, we propose a unique DNN architecture consisting of multiple GCNs in parallel followed by another GCN to merge the outputs of all GCNs. Figure 1(b) shows the concept of the proposed method. Each GCN performs convolutions on a limited number of keypoints (*i.e.*, some of the keypoints have been removed). The combination of removal differs in each GCN. Each GCN predicts an incomplete 3D pose, and then the final GCN merges the incomplete poses into a single 3D pose by selecting accurate positions from them. We demonstrate that the proposed method outperforms the existing methods on RHD [12].

2 Proposed Method

The proposed network consists of three modules: a 2D heatmap generator, a parallel sub-joints network (PSJNet), and an ensemble network (EN). First, from an input image, the 2D heatmap generator predicts a 2D pose, the shape of which is $J \times 2$, where $J = 21$ is the number of keypoints. Then, each branch of PSJNet estimates an incomplete 3D pose from the 2D pose in which some of the keypoints have been removed. Specifically, each GCN reconstructs 3D pose $(J - m) \times 3$ from two $(J - m) \times 2$ vectors, one from a partial 2D pose and the other from image features of the Feature Encoder (FE), where m is the number of removed keypoints. In the experiment, the sets of removed keypoints are {TIP}, {DIP}, {PIP}, {MCP} in every finger, *i.e.* $N = 4$ and $m = 5$. Finally, EN receives N incomplete 3D poses from PSJNet, an $N(J - m) \times 3$ vector, and an image feature from FE, a $(J - m) \times 3$ vector. EN merges the incomplete poses into a single 3D pose $J \times 3$ by selecting accurate positions from them. Table 1 summarizes the network architecture of FE.

The 3D hand pose reconstruction from a single 2D pose is ill-posed due to the depth and scale ambiguity. Therefore, similarly to Zimmermann *et al.* [12] and Cai *et al.* [9], our network outputs a 3D pose of i-th keypoint $\boldsymbol{x}_i = (x_i, y_i, d_i)$ as follows:

$$d_i = \frac{z_i - z_{\text{root}}}{s}, \tag{1}$$

where z_i and z_{root} denote the absolute depth of the i-th point and the wrist, respectively, and $s = \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2$ is the length of a keybone. We choose k and $k + 1$ for the first bone of the middle finger in this paper.

2.1 Loss Function

Our method has two types of loss function: 2D loss and 3D loss. We calculate the mean square error (MSE) for training the 2D heatmap generator as:

$$L_{\text{2D}} = \sum_{i \in J} \left\| \boldsymbol{h}_i - \hat{\boldsymbol{h}}_i \right\|_2, \tag{2}$$

where $\boldsymbol{h}_i$ and $\hat{\boldsymbol{h}}_i$ represent the predicted heatmap of the i-th keypoint and the corresponding ground-truth, respectively. The 3D loss L_{3D} for training the GCNs and the FEs is defined by

$$L_{\text{3D}} = \alpha \sum_{i \in J} \text{smooth}_{\text{L1}}(\boldsymbol{p}_i, \hat{\boldsymbol{p}}_i) + \beta \sum_{i \in J} \text{smooth}_{\text{L1}}(d_i, \hat{d}_i). \tag{3}$$

The variables with a hat, $\hat{\boldsymbol{p}}_i = (\hat{x}_i, \hat{y}_i)$ and $\hat{d}_i$, represent the ground-truth of the i-th keypoint, and $\boldsymbol{p}_i = (x_i, y_i)$ and d_i represent predictions. The coefficients α and β

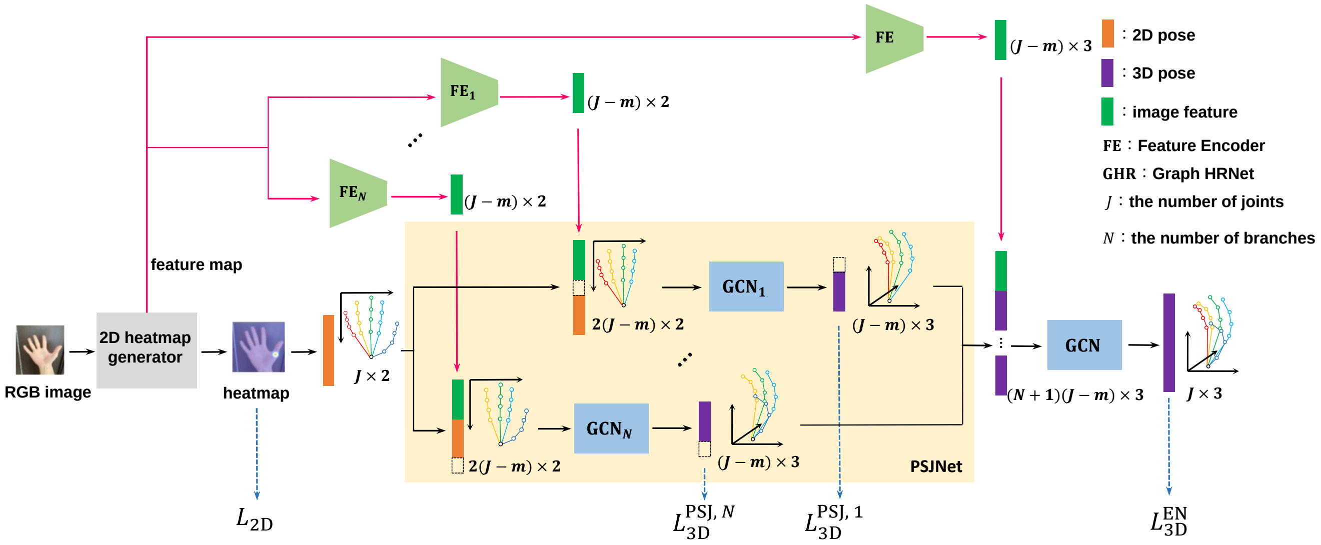


Figure 2: Overview of proposed network. First, the 2D heatmap generator predicts the 2D position of all keypoints in an input image. Then, PSJNet, which is composed of N GCNs in parallel, estimates multiple incomplete 3D poses in which part of keypoints have been removed. FEs extract image features from an intermediate feature map in the 2D heatmap generator. Finally, EN merges the incomplete N poses into a single 3D pose.

are weight factors to balance the value range of the two losses. We set $\alpha = 0.01$ and $\beta = 1$ in the experiment. Similar to [9], we used the smooth L1 loss.The total 3D loss function L_{3D}^{total} is defined by the sum of the 3D loss (Eq. (3)) for each GCN of PSJNet $L_{3D}^{\mathrm{PSJ},k}$ and EN L_{3D}^{EN}, *i.e.*

$$L_{3D}^{\mathrm{total}} = L_{3D}^{\mathrm{EN}} + \frac{1}{N}\sum_{k\in N} L_{3D}^{\mathrm{PSJ},k}. \tag{4}$$

2.2 Training

For stable convergence, we train the proposed network by the following four steps:

1. Train the 2D heatmap generator by minimizing L_{2D}.
2. Train PSJNet and the connected FEs by minimizing $\sum L_{3D}^{\mathrm{PSJ},k}$.
3. Train EN and its FE by minimizing L_{3D}^{EN}.
4. Finetune PSJNet and EN by minimizing L_{3D}^{total}.

After the first step, the 2D heatmap generator is frozen. PSJNet and the FEs are also frozen at the third step.

3 Experiments

The proposed method aims to reduce the accuracy degradation of visible keypoints caused by severe occlusions. In this experiment, we define a severe occlusion as the situation where more than half of the 21 keypoints, *i.e.* > 10 keypoints, are occluded. We evaluated the performance of the proposed and existing methods on images under small occlusions and severe occlusions in RHD [12].

3.1 Settings

We compared the following three methods in this experiment.

Cai *et al.* [9]
CNN-based method with RGB+D images. Trained full-supervised setting.

HopeNet [11]
GCN-based method with RGB images. We replaced the 2D keypoint predictor with HRNet-W32 [13], and customized it to output only 21 keypoints for fair comparisons.

Proposed
GCN-based method with RGB images. We used HRNet-W32 for the 2D heatmap generator, and Adaptive Graph U-Net [11] for GCNs.

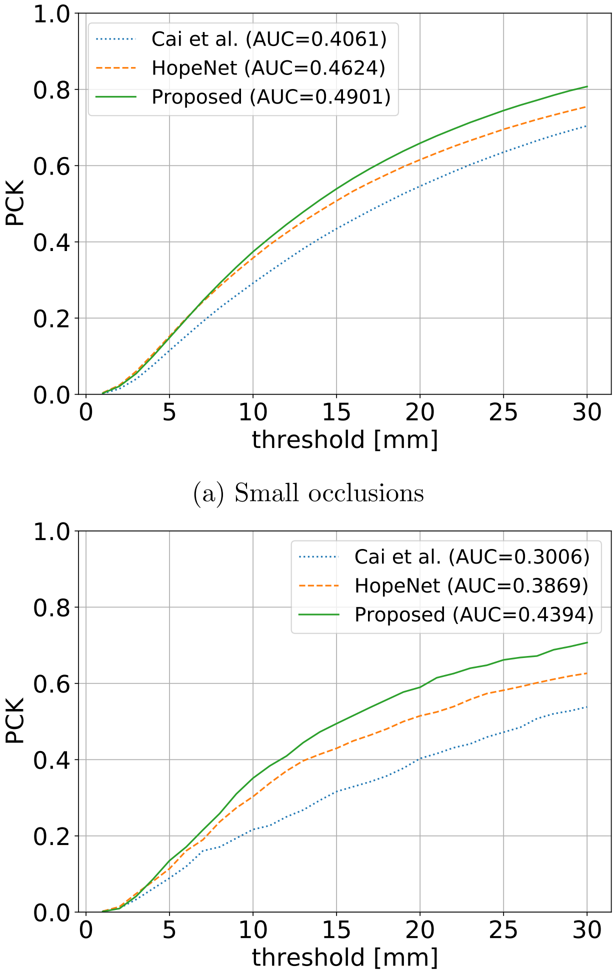
We measured the percentage of correct keypoints (PCK) and its area under the curve (AUC) for evaluating the pose estimation accuracy. To calculate PCK in the camera coordinate systems, we assume that the global scale of the keybone, the root depth, and the intrinsic camera parameters are known in the experiments. This is the same assumption as the previous work by [9, 12].

3.2 Results

Figure 3(a) and (b) show the PCK curves on RHD, which are the performance on visible keypoints under

Table 2: AUC under each condition and the relative values between small and severe occlusions. All: all visible keypoints regardless of occlusions. Small occ.: more than 10 visible keypoints under small occlusions. Severe occ.: less than 11 visible keypoints under severe occlusions. Rel. value: the relative performance of AUC for severe occlusions with respect to AUC for small occlusions.

Method	All	Small occ. (A)	Severe occ. (B)	Rel. value (= B/A)
Cai *et al.*	0.4055	0.4061	0.3006	0.7402
HopeNet	0.4606	0.4624	0.3869	0.8366
Proposed	**0.4889**	**0.4901**	**0.4394**	**0.8965**



(a) Small occlusions

(b) Severe occlusions

Figure 3: PCK curves on visible keypoints under small and severe occlusions. Severe occlusions degrades the performance of all methods. The proposed method outperforms the existing methods by around 0.1 points at the 30 mm threshold.

small and severe occlusions, respectively. As we can see in the figure, severe occlusions degrades the performance of all methods. The PCK curves also clearly indicate that the proposed method outperforms the existing methods for both occlusion scenarios. The AUC results are summarized in Table 2, where AUC for all visible keypoints are additionally shown. We also calculated the relative performance (shown in the last column) between AUC for small and severe occlusions to check the degradation caused by severe occlusions. A relative value approaching 1 indicates stable performance regardless of the amount of occlusions, while a relative value approaching 0 indicates a degraded performance due to the effect of occlusions. Our comparison of the relative values shows that the proposed method is more robust to the amount of occlusions than the existing methods.

4 Conclusion

In this paper, we have proposed an occlusion-robust network for 3D hand pose estimation from a single RGB image. Working from a hypothesis that convolutions on all keypoints cause inaccurate feature propagation from occluded keypoints, we configured multiple GCNs to predict incomplete 3D hand poses in which some of the keypoints have been removed. The incomplete poses are merged into a single pose by an ensemble network. Our hypothesis was validated by experiments on RHD, where the proposed method significantly improved the estimation accuracy and was more robust to occlusions than the existing methods.

References

[1] Sigal Leonid and Michael J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vol. 2. pp. 2041–2048, 2006.

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

[3] Ghiasi Golnaz, Yang Yi, Ramanan Deva, and Fowlkes Charless C. Parsing occluded people. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2401–2408, 2014.

[4] Rafi Umer, Juergen Gall, and Bastian Leibe. A semantic occlusion model for human pose estimation from a single depth image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 67-74, 2015.

[5] Albert Haque, Boya Peng, Zelun Luo, Alexandre Alahi, Serena Yeung, and Li Fei-Fei. Towards viewpoint invariant 3d human pose estimation. In European Conference on Computer Vision, pp. 160–177. Springer, 2016.

[6] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In Proceedings of the IEEE International Conference on Computer Vision, pp. 1154–1163, 2017.

[7] Qi Ye and Tae-Kyun Kim. Occlusion-aware hand pose estimation using hierarchical mixture density network. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–817, 2018.

[8] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. Occlusion-aware networks for 3d human pose estimation in video. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 723–732, 2019.

[9] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly supervised 3d hand pose estimation from monocular rgb images. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 666–682, 2018.

[10] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 10833–10842, 2019.

[11] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6608–6617, 2020.

[12] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In Proceedings of the IEEE international conference on computer vision, pp. 4903–4911, 2017.

[13] Ke Sun, Bin Xiao, Dong Liu, and JingdongWang. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5693–5703, 2019.