Faster VoxelPose: Real-time 3D Human Pose Estimation by Orthographic Projection

Supplementary Material

Hang Ye[1*], Wentao Zhu[2,3*], Chunyu Wang[4+],
Rujie Wu[2,3], and Yizhou Wang[2,3,5]

[1] Yuanpei College, Peking University
[2] Center on Frontiers of Computing Studies, Peking University
[3] School of Computer Science, Peking University
[4] Microsoft Research Asia
[5] Inst. for Artificial Intelligence, Peking University
{yehang, wtzhu, wu_rujie, yizhou.wang}@pku.edu.cn, chnuwa@microsoft.com

Abstract. This document provides supplementary information which is not elaborated in our manuscript due to space limits. Section 1 gives details about the implementation of our method. Section 2 describes the datasets and evaluation metrics we use in our experiments. Section 3 presents additional experiment results. We also present a video demo that includes additional results.

1 Implementation details

1.1 Human Detection Networks

Following [7], we discretize the overall motion space into $L \times W \times H$ voxels. In our experiments, we set $L = W = 80$ and $H = 20$.

Inspired by [7], we adopt a similar Encoder-Decoder architecture in the Human Detection Networks. The key difference is that we replace all expensive 3D convolutions with 2D and 1D convolutions. The basic components of our fully-convolutional networks include vanilla convolutional block and residual convolution block. The former is comprised of one convolutional layer, one batch-norm layer and ReLU while the latter consists of two consecutive basic convolutional blocks with residual connection. At the initial stage, the feature volume is fed into a 7×7 convolutional layer. In the subsequent Encoder structure, the feature representation is downsampled through three 3× 3 residual convolutional blocks with maxpooling. The Decoder adopts a symmetric design, but with deconvolution operations. Finally, the network generates the results through a 1×1 convolutional layer. Following [9], the outputs of 2D networks are fed into three

[*] Equal contribution.
[+] Corresponding author.

branches to estimate the feature map, the local offset and the size of the bounding box respectively. They share an identical design, which consists of a 3 × 3 convolution, ReLU and another 1 × 1 convolution.

The 1D convolutional network shares the same architecture with its 2D counterpart except for two aspects: 1) all convolutional operations are replaced with 1D convolutions 2) we just maintain the branch for estimating feature maps along the z-axis.

1.2 Joint Localization Networks

The architecture of Joint Localization Networks is essentially the same as one 2D CNN branch of HDN. The outputs of 2D estimators are further fed into a shared confidence network, which consists of one convolutional layer, one global average pooling layer plus a fully-connected layer.

1.3 Training

We train HDN and JLN jointly to convergence. On the CMU Panoptic dataset, our model is trained 10 epochs with batch size 8. On the Shelf and Campus datasets, we train our model for 30 epochs with the same batch size. The learning rate is set to be $\alpha = 0.0001$ using Adam [5] optimizer. The parameters above are empirically determined.

In the bounding box regression branch of HDN, we add a safety margin $\delta = 200$mm to GT, as missing information of body joints will be fatal to the subsequent prediction.

2 Experimental Details

2.1 Dataset

CMU Panoptic [4] This dataset captures multiple people engaging in social activities in an indoor setting. It contains massive sequences in various scenarios. We use the sequences captured by five HD cameras (3, 6, 12, 13, 23). The training and testing split is identical with [6, 7].

Shelf [1] This dataset captures four people disassembling a shelf using five cameras. We follow previous works [2, 3, 6, 7] in evaluating only three of the four persons on the test set frames 300-600 since one person is severely occluded. Due to the lack of complete annotations of ground-truth poses, we train with synthetic heatmaps following previous works [6–8].

Campus [1] This dataset captures multiple people interacting with each other in an outdoor environment by three cameras. We follow previous works [2, 3, 6, 7] and perform evaluation on the test set frames 350-470, 650-750. Similar to the Shelf dataset, we also conduct training on synthetic heatmaps.

2.2 Evaluation Metrics

PCP For the Percentage of Correct Parts, we pair each GT pose with the closest estimation and calculate the percentage of correct parts. Specifically, the match is counted as correct if their distance is within a threshold T. Following [2,3,6,7], we set T to be half of the corresponding limb length. Note that PCP does not penalize false positive results.

AP$_{\boldsymbol{K}}$ In order to evaluate the results more comprehensively, we follow [6, 7] to measure the Average Precision (within Kmm). Specifically, a predicted joint is considered as correct if there is a corresponding GT joint within distance threshold K.

MPJPE We first pair the nearest GT for each predicted joint, then calculate the corresponding Mean Per Joint Position Error in millimeters.

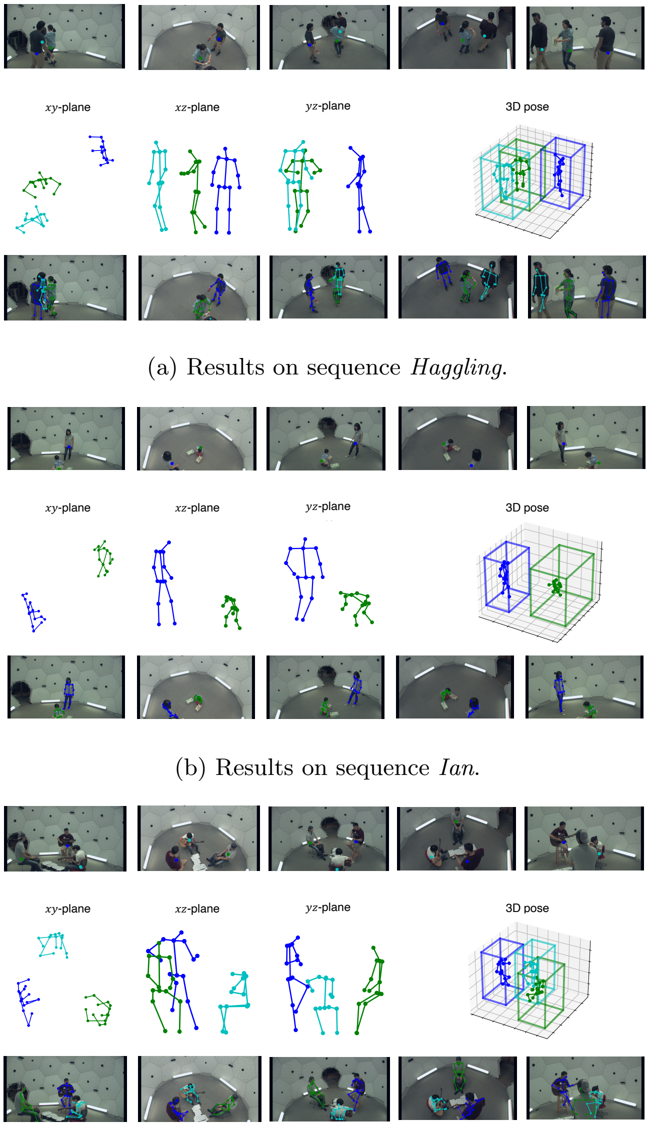
3 Additional Results

Num.	1	2	3	4	5	6	7	8	9	10
HDN	18.27	17.90	17.71	18.22	18.28	18.50	17.37	17.86	18.30	18.45
JLN	13.16	13.67	14.22	15.03	16.72	18.40	20.70	21.22	24.01	25.88
Total	31.43	31.57	31.93	33.25	35.00	36.90	38.07	39.08	42.31	44.33

Table 1: **Experiment of scalability.** We measure the average inference time cost of each module in milliseconds (ms) while varying the number of persons present in the synthetic scene.

We study the influence of the number of persons on inference time. The results are shown in Table. 1. The time increase is mainly on the feature construction phase of JLN.

We present additional qualitative results in Fig. 1. Please refer to the attached video for more results.



(a) Results on sequence *Haggling*.

(b) Results on sequence *Ian*.

(c) Results on sequence *Band*.

Fig. 1: **Additional Results on the CMU Panoptic Dataset.** We present the results on three different action sequences of the test set. For each figure, the first row illustrates the estimated root joints in HDN. The second row shows the estimated 2D poses on the three orthogonal re-projection planes and the fused 3D pose in JLN. The last row shows the 2D back-projection of the estimated 3D pose to each camera view.

References

1. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures for multiple human pose estimation. In: CVPR (2014)
2. Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X.: Fast and robust multi-person 3d pose estimation from multiple views (2019)
3. Huang, C., Jiang, S., Li, Y., Zhang, Z., Traish, J.M., Deng, C., Ferguson, S., Xu, R.Y.D.: End-to-end dynamic matching network for multi-view multi-person 3d pose estimation. In: ECCV (2020)
4. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: ICCV (2015)
5. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint **arXiv:1412.6980** (2014)
6. Lin, J., Lee, G.H.: Multi-view multi-person 3d pose estimation with plane sweep stereo. In: CVPR (2021)
7. Tu, H., Wang, C., Zeng, W.: Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In: ECCV (2020)
8. Wu, S., et al: Graph-Based 3D Multi-Person pose estimation using Multi-View images. In: ICCV (2021)
9. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. In: CVPR (2019)