Fast and Robust Multi-Person 3D Pose Estimation from Multiple Views Supplementary material

Junting Dong[1]* Wen Jiang[1] Qixing Huang[2] Hujun Bao[1] Xiaowei Zhou[1]†
[1]Zhejiang University [2]University of Texas at Austin

1. Proposed Baseline Method

Here, we describe the details of the proposed baseline method in Section 4.2 ablation analysis of the paper. In order to compare the performance on the Campus and Shelf datasets, we re-implement previous works [1, 4] with the state-of-the-art 2D pose detector [3]. First, we perform 2D pose estimation on each view with the 2D pose detector. The resulting 2D keypoints and part affinity fields (PAFs) are used for the subsequent steps. For the same type of keypoints, we triangulate each pair of 2D keypoints from different views to construct the state space, which is composed of all possible candidates of this type. Then, we define the unary terms and pairwise terms of 3D candidates as follows. Given the calibrated cameras, we can project each 3D candidate to each view and acquire the confidence value from the heatmap. The unary terms are obtained with the product of corresponding confidence values. To obtain the pairwise terms of adjacent joints, we project the 3D candidates to each view and calculate the affinity between them with PAFs. We use the max-product algorithm [2] to do multiple pose inferences for final 3D poses.

2. Additional Results

We evaluate the proposed approach on the CMU Panoptic dataset [4] quantitatively. The authors of the dataset also proposed an algorithm that could accurately reconstruct 3D poses using hundreds of cameras. These reconstructions were used as ground truth to evaluate the results with fewer cameras. We test on the '160422 ultimatum1' sequence with images from five cameras uniformly sampled among all cameras. Figure 1 shows the resulting PCKs. The result of [4] is copied from the original paper. Note that the cameras used in [4] were not specified in the original paper and might be different from the ones in our experiment. So the result from [4] is not rigorously comparable. But it still provides a meaningful baseline. When the threshold is small, their method acquires higher accuracy because the 'ground truth' is obtained from the same algorithm but with more cameras. As the threshold increases, our approach exceeds the compared method with a large margin. The results demonstrate that the proposed approach is able to robustly recover the multiple 3D poses in a crowded scene with a few cameras and outperforms the baseline approach.

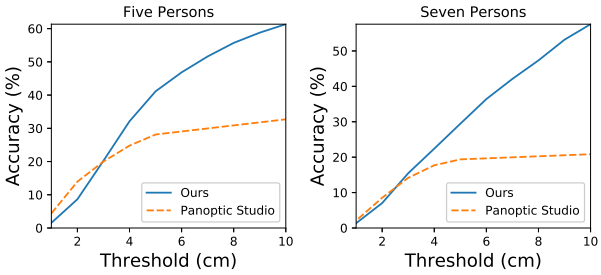

Figure 1: Performance comparison on the CMU Panoptic dataset with five VGA cameras. The x and y axes represent the threshold of Probability of Correct Keypoint and accuracy respectively. The result of the baseline is taken from original paper [4]. The ground truth is also obtained from their algorithm but with more cameras.

In addition, we provide a video to show our result on the Shelf dataset. In the video, we fit the SMPL [5] model to the resulting skeleton for better visualization.

References

[1] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures revisited: Multiple human pose estimation. *T-PAMI*, 38(10):1929–1942, 2016.

[2] M. Burenius, J. Sullivan, and S. Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *CVPR*, 2013.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[4] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. S. Godisart, B. Nabbe, I. Matthews, et al. Panoptic studio:

*The authors from Zhejiang University are affiliated with the State Key Lab of CAD&CG and the ZJU-SenseTime Joint Lab of 3D Vision.
†Corresponding author.

A massively multiview system for social interaction capture. *T-PAMI*, 2017.

[5] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015.