Multi-person 3D Pose Estimation in Crowded Scenes Based on Multi-View Geometry ——*Supplementary Material*——

He Chen*[1], Pengfei Guo1[1], Pengfei Li[1], Gim Hee Lee†[2], Gregory Chirikjian2[2,1]

[1] The Johns Hopkins University, USA
[2] National University of Singapore, Singapore
{hchen136, pguo4, pli32, gchirik1}@jhu.edu
{gimhee.lee@comp, mpegre@}.nus.edu.sg

1 Introduction

In this supplementary document, we first provide implementation details *i.e.* annotations and hyperparameters. Then, we demonstrate additional experimental results including convergence analysis and more qualitative results on Wildtrack dataset [3] and Chariot Mk I dataset [5].

2 Implementation Details

2.1 Details of Annotations

Popular human pose dataset (*e.g.* MSCOCO [9] and MPII [1]) contains limited body part types. For both of aforementioned datasets, feet annotations are represented by ankle only. However, accurate feet annotations are required by various graphics applications, such as avatar retargeting and 3D human shape reconstruction [2]. To evaluate proposed framework, we added 26 keypoints annotations of 2D joints (17 joints of body parts in MSCOCO style [9] and 6 additional labels of feet) to existing multi-camera datasets, including LOEWENPLATZ [4], Chariot Mk I [5], and Wildtrack [3]. 6 additional feet annotation are also merged to existed keypoints annotation of CMU Panoptic Dataset [6]. Figure 1 shows the position of 6 additional feet keypoints. Big toe (a), small toe (b), and heel (c) are identified in both of feet. All added annotations were first inferred by a state-of-the-art pose estimation algorithm [8], then were manually fine-tuned by human in Visipedia [10]. To propel the development of this area, those additional annotations of 4 multi-view datasets, which are recorded in various scenarios (*e.g.* indoor human activities, surveillance and autonomous driving), will be made available to the community.

---

* Equal first author contribution. † Jointly supervised this work.

Fig. 1. Visualization of the six joints on feet.

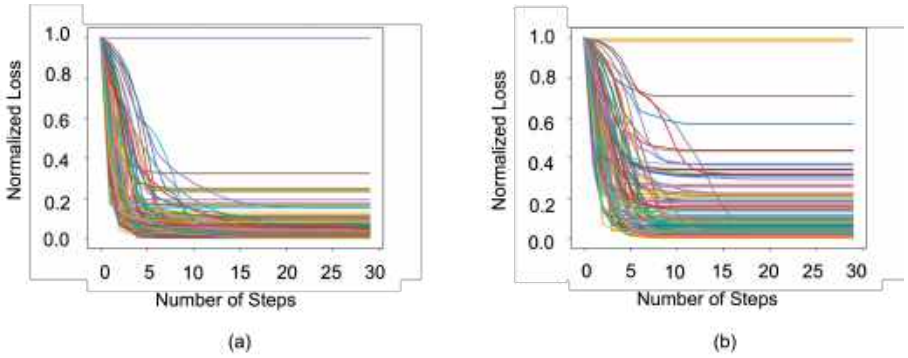

Fig. 2. Convergence Analysis through tracking of loss function values.(a) convergence analysis on Wildtrack dataset. (b) convergence analysis on LOEWENPLATZ dataset.

2.2 Details of Training and the other Hyperparameters

Our whole work is implemented in Pytorch with two Nvidia 2080 GPUs. We modified the Joint Candidate SPPE network by extending 6 additional channels for the output layer. The learning rate is 1 $\times 10^{-4}$ for the modified output layer and is 1 $\times 10^{-5}$ for all other layers during first 50 epochs, then learning rates are linearly decayed to 0 during last 50 epochs. The modified MSE loss applies double penalty on the error from 6 feet annotation joints to facilitate the improvement on feet detection accuracy. Data augmentation is deployed cross the holistic training phase, including random rotation ($\pm$ 15), scaling ($\pm$ 10), and random flip, and elastic deformable transformation with a small distortion factor (0.1). We adopt Adam optimizer [7] with exponential decay rates $\beta =$ (0.9, 0.999) for the moment estimates and batch size is set to 128. For stage two, feature matching is carried out based on ground warping, then 3D crowd reconstruction is deployed. For feature matching, due to limitation of view overlap, not every human body has a correspondence among all cameras. Because forcing nonexistent person to find a match would do harm to the final result, we set a threshold for the confidence of matching. If the cost for matching is larger than a threshold, we discard the match. This threshold is set 0.2.

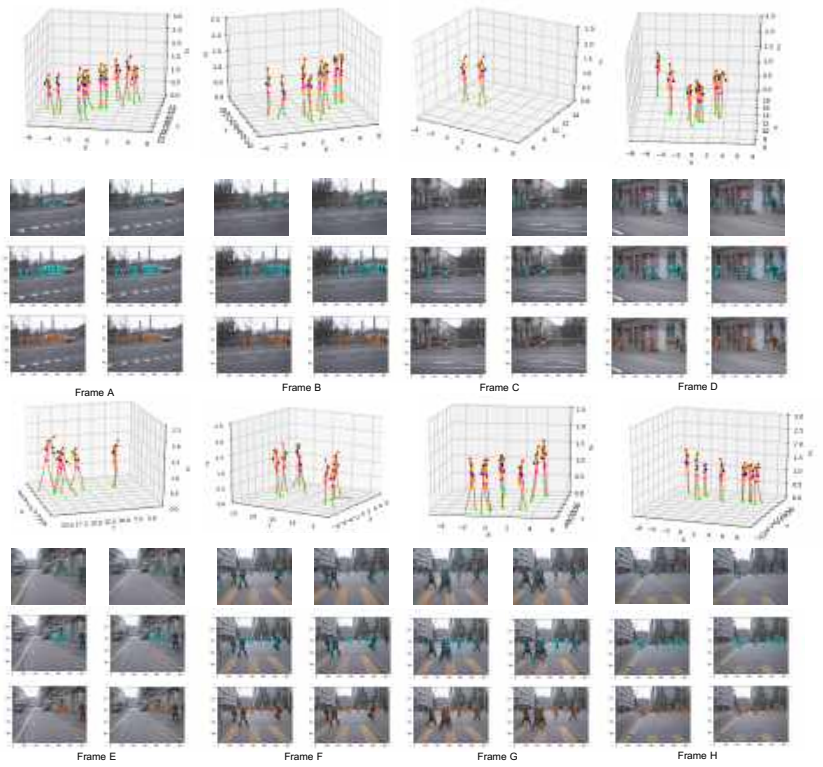

Fig. 3. Qualitative results on sampled frames collected from the video of LOEWENPLATZ dataset. For each frame (Frame A to Frame H). First row of each frame shows the estimated 3D crowd human pose. The second row of each frame shows 2D joints and skeletons detected by modified candidate joint SPPE with attention placed on feet. The third row shows visualization of 2D keypoint ground truth. The fourth row shows visualization of keypoints when ground truth and those points reprojected from estimated 3D pose are merged in the same images. (This figure is better to be observed on screen.)

3 Additional Experimental Results

Due to the limitation of page length in the original manuscript, we only showed a small proportion of the experimental results. Supplementary experiments are presented here. Convergence analysis is carried out. Evaluations on multiple datasets shows robustness of the proposed method.

3.1 Convergence Analysis

Since iterative approach is used in the proposed method to triangulate 3D joints, convergence analysis is necessary to be carried out. We run the code 100 times

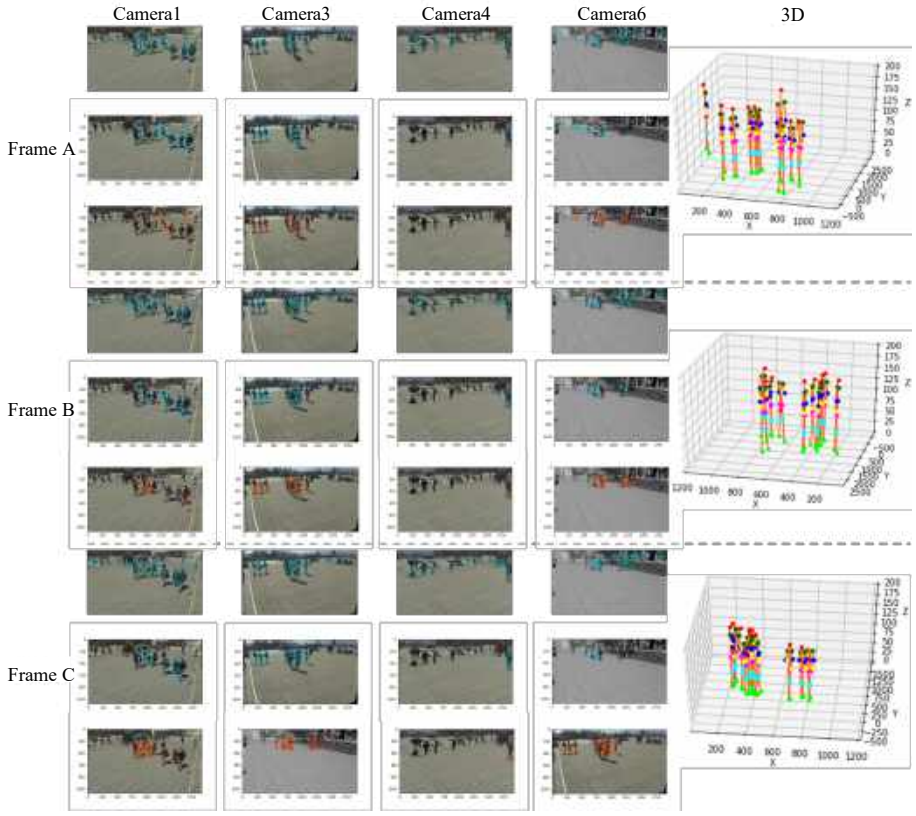

Fig. 4. Qualitative results on three consecutive frames from Wildtrack dataset. The four columns represent results of Camera1, Camera3, Camera4, and Camera6 from Wildtrack dataset. In the last column, 3D crowd human pose visualization are presented. For each frame, the first row shows results of modified candidate joint SPPE with attention on the feet; the second row shows visualization of key points when ground truth(blue dots); The third row shows those points reprojected points(orange dots) from estimated 3D pose are merged in the same images. (This figure is better to be observed on screen.)

and track the normalized value of loss function with respect to number of steps as shown in Figure 4. By normalization, for each of the 100 times, we divide the value of loss by its value at step 0. In this way, the decreasing slope could be visualized clearly. From Figure 4, it can be observed that the loss always converge. The reason that convergence value is different is because measurement error is different, so that more time is required for the loss to converge if the measurement error is larger. The intuition is similar to gradient descent, good initialization usually means fewer steps are required before convergence, vice versa.

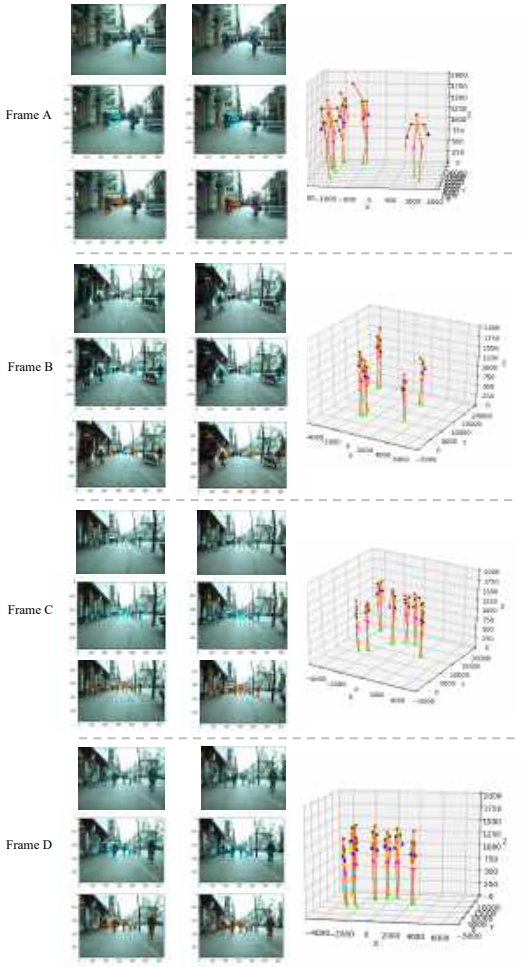

Fig. 5. Qualitative results on sampled frames collected from the video of Chariot Mk I dataset. For each frame (Frame A to Frame D): The first row shows estimated 3D crowd human pose. The second row shows result of 2D joints and skeletons detected by modified candidate joint SPPE with the attention stressed on feet. The third row shows visualization of 2D key point ground truth. The fourth row shows visualization of key points when ground truth and those points reprojected from estimated 3D pose are merged in the same images. (This figure is better to be observed on screen.)

3.2 Supplementary Qualitative Results

Figure 4 shows the additional qualitative results on Wildtrack Dataset by three consecutive frames. Figure 5 demonstrates the qualitative result on Chariot Mk I dataset. This dataset provides videos from pedestrian perspective, and the scenes are dynamic scenes captured by moving cameras. This video vividly simulate vieos in the wild. Besides translation together with human body, the dy-

namic scene also suffers from hand shaking. As can be observed from the result, our method is still robust in this scenario. The orange dots representing reprojected keypoints almost perfectly overlaps with the blue dots representing ground truth when they are drawn in the same image. Figure 3 shows more results on LOEWENPLATZ dataset.

References

1. Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele.
2. Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.
3. Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proc. CVPR*, pages 5030–5039, 2018.
4. Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. Robust multi-person tracking from a mobile platform. In *IEEE Trans. PAMI*, pages 1831–1846, 2009.
5. Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. A mobile vision system for robust multi-person tracking. In *Proc. CVPR*, pages 1–8. IEEE, 2008.
6. Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proc. ICCV*, 2015.
7. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
8. J. Li, C. Wang, H. Zhu, Y. Mao, H. Fang, and C. Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark, 2019.
9. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. ECCV*, pages 740–755. Springer, 2014.
10. Visipedia. Visipedia annotation toolkit, 2018.