Evaluating Recent 2D Human Pose Estimators for 2D-3D Human Body Pose Lifting

Soroush Mehraban[1,2], Yiqian Qin[1]

[1]University of Toronto, [2]KITE Research institute – University Health Network

Motivation

- Monocular 3D human pose estimation is a critical computer vision task that mainly entails predicting 3D pixel coordinates of key body joints based on a 2D image or video.
- Current 3D human pose estimation process typically first estimate 2D positions of human body key joints using off-the-shelf 2D human pose estimation models. The resulting 2D estimations are then passed to the 3D human pose estimation model as input to estimate the corresponding 3D key joints positions.
- Various 2D human pose estimation models [4, 5, 6, 7] have been proposed lately, and their potential to serve as less noisy 2D input for 3D human pose estimation models remains unexplored.
- Our study aims to evaluate the performance of recently proposed 2D human pose estimation models on 2D-3D human pose lifting, and enhance the 2D-3D lifting results by merging outputs (2D positions of key joints) of these models.

Related Work

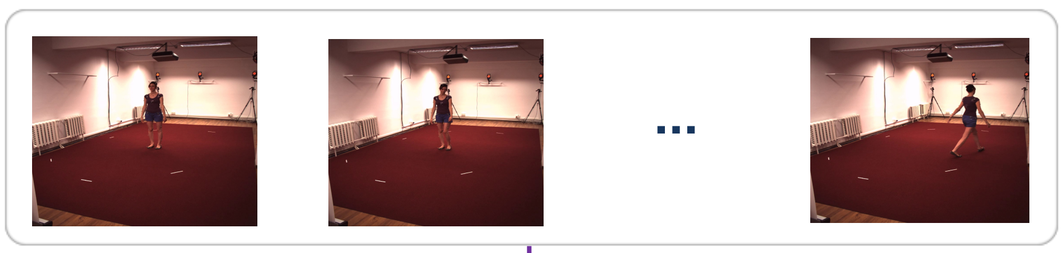
2D Human Pose Estimation

- These models receive a single RGB image as input, and output locations of key joints in 2D pixel coordinate.
- CPN [3] uses feature pyramid networks to identify features and then refines the occluded keypoints using a RefineNet module.
- TransPose [5] and ViTPose [6] use vision transformers to encode the image.
- MogaNet [4] proposes a new ConvNet structure, and PCT [7] proposes a structured representation to explore the joint dependency.

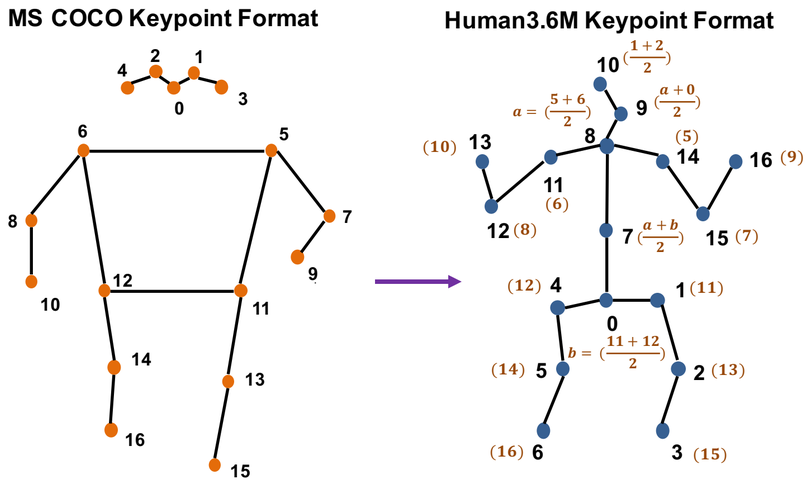
3D Human Pose Estimation

- These models lift a sequence of 2D poses to 3D pose sequences.
- PoseFormerV2 [1] leverages frequency-domain representation to infer 3D poses robust against sudden movements in noisy data.

Method



1 Use recent 2D human pose estimation models which are trained on MS COCO keypoint dataset to estimate 2D keypoints on Human3.6M dataset. Convert the resulting 2D keypoints estimations from MS COCO to Human3.6M format.



2 Apply 3 strategies to merge different 2D keypoints estimations:

- **Manual merging:** For each joint, select the best 2D estimation based on least mean distance compared to the 2D ground truth (acquired through camera projection).
- **Average merging:** For each frame, average the outputs from 2D estimators.
- **Weighted average merging:** Take weighted average based on the confidence score generated by 2D estimators.

3 Train PoseFormerV2 [1] using different 2D keypoints estimations and 3D ground truth. Evaluate models' performance for the task of 2D-3D human pose lifting.

Experimental Results

Quantitative Results

2D Estimator	finetuned	MPJPE (mm)↓
Detectron [2]	×	59.56
Detectron [2]	✓	55.91
CPN [3]	✓	49.65
MogaNet [4]	×	54.77
TransPose [5]	×	66.20
PCT [6]	×	53.26
ViTPose [7]	×	52.61
Merge (Manual)	×	51.96
Merge (Average)	×	52.53
Merge (Weighted Average)	×	52.50

Figure 1. The mean per-joint position error (mm) comparisons of estimated 3D keypoints on Human3.6M after training the PoseFormerV2 model using different 2D estimations.

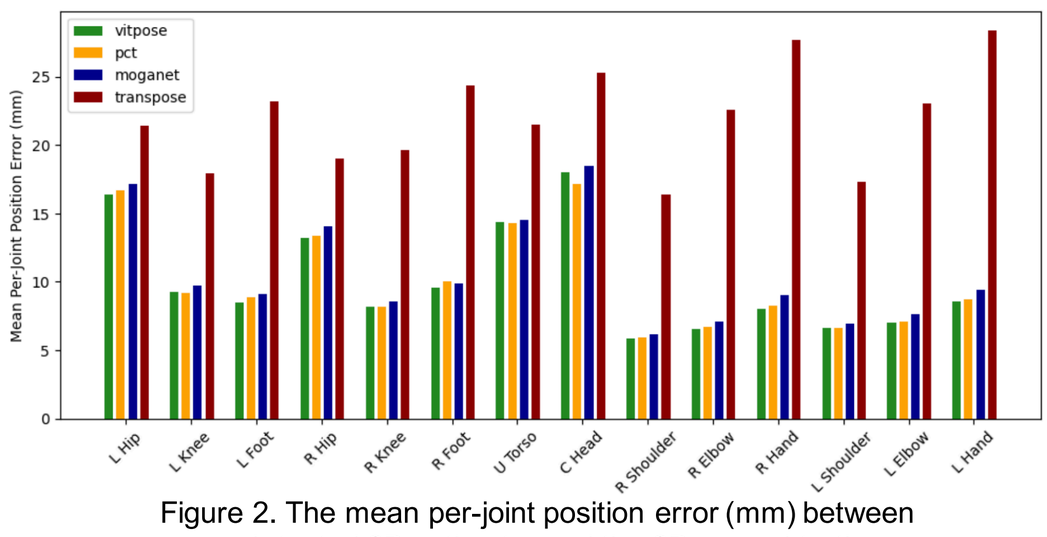


Figure 2. The mean per-joint position error (mm) between each tested 2D estimator and the 2D ground truth.

Qualitative Results

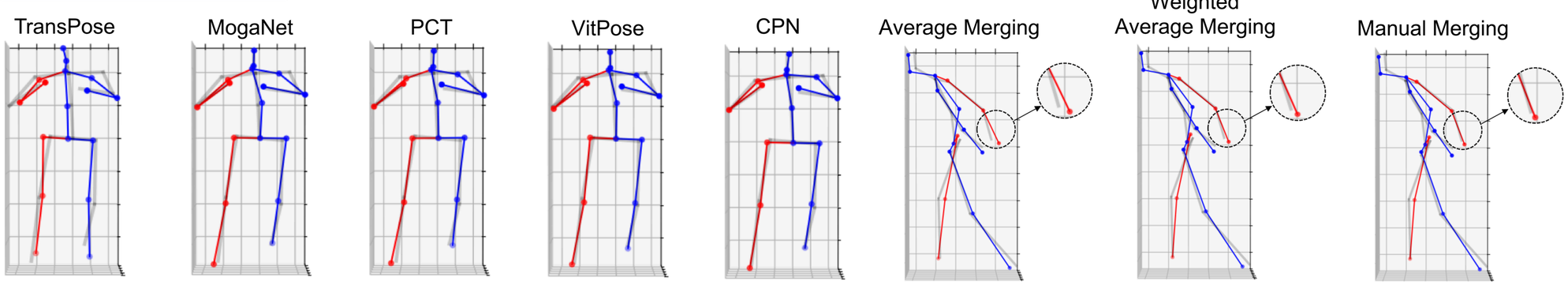


Figure 3. Qualitative comparisons of estimated 3D keypoints on Human3.6M after training the PoseFormerV2 model using different 2D estimations. The transparent gray skeleton is the ground-truth 3D pose.

References

[1] Zhao, Qitao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. "PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8877-8886. 2023.

[2] Güler, Rıza Alp, Natalia Neverova, and Iasonas Kokkinos. "Densepose: Dense human pose estimation in the wild." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7297-7306. 2018.

[3] Chen, Yilun, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. "Cascaded pyramid network for multi-person pose estimation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7103-7112. 2018.

[4] Li, Siyuan, Zedong Wang, Zicheng Liu, Cheng Tan, Haitao Lin, Di Wu, Zhiyuan Chen, Jiangbin Zheng, and Stan Z. Li. "Efficient multi-order gated aggregation network." *arXiv preprint arXiv:2211.03295* (2022).

[5] Yang, Sen, Zhibin Quan, Mu Nie, and Wankou Yang. "Transpose: Keypoint localization via transformer." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11802-11812. 2021.

[6] Geng, Zigang, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. "Human Pose as Compositional Tokens." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 660-671. 2023.

[7] Xu, Yufei, Jing Zhang, Qiming Zhang, and Dacheng Tao. "Vitpose: Simple vision transformer baselines for human pose estimation." *Advances in Neural Information Processing Systems* 35 (2022): 38571-38584.