Motion Guided 3D Pose Estimation from Videos Supplementary Material

Jingbo Wang[1][0000−0001−9700−6262], Sijie Yan[1][0000−0003−4398−0590],
Yuanjun Xiong[2][0000−0002−6391−4921], and Dahua Lin[1][0000−0002−8865−7896]

[1] The Chinese University of Hong Kong {jbwang,ys016,dhlin}@ie.cuhk.edu.hk
[2] AWS/Amazon AI
yuanjx@amazon.com

1 Introduction

This supplementary material contain following details: (1) The formulation of spatial graph convolution operator in the main paper; (2)Comparison with other methods based on ground-truth 2D pose; (3) Impacts of 2D pose estimators; (4) Additional visualization results.

2 Details of the spatial temporal graph convolution

Formally, we have the spatial graph convolution f_s as:

$$f_s(v_{t,j}) = \sum_{v_{t,i} \in B(v_{t,j})} \frac{1}{Z_{t,j}(v_{t,i})} f_{in}(v_{t,i}) \cdot \mathbf{w}(l_{t,j}(v_{t,i})), \tag{1}$$

where $B(v_{t,j})$ is the neighbor set of node $v_{t,j}$, $l_{t,j}$ maps a node in the neighborhood to its subset label, $\mathbf{w}$ samples weights according to a subset label, and $Z_{t,j}(v_{t,i})$ is a normalization term equivalent the cardinality of the corresponding subset. Since the human torso and limbs act in very different ways, it inspires us to give the model spatial perception for distinguishing the central joints and limbic joints. To make spatial configuration explicit in the 3D pose estimation, we divide the neighborhood into three subsets:

$$l_{t,j}(v_{t,i}) = \begin{cases} 0 & \text{if } h_{t,j} = h_{t,i} \\ 1 & \text{if } h_{t,j} < h_{t,i} \\ 2 & \text{if } h_{t,j} > h_{t,i} \end{cases} \tag{2}$$

, where $h_{t,i}$ is the number of hops from $v_{t,i}$ to the root node (i.e. central hip in this work).

3 Comparison with other methods based on Ground-Truth 2D pose

Several state-of-the-arts report their results on 2D ground-truth to explore their upper bound in 3D pose estimation. The results are illustrated in the Table 1.

It can be seen that our method achieves the best performance (25.6 MPJPE) outperforming all other methods with the ground-truth input.

Table 1. To exclude the interference of 2D pose estimator, we compare our models and state-of-the-arts trained on ground truth 2D pose. Results showing the action-wise errors on Human3.6M under Protocol-1.

Protocol 1 (GT)	Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Somke	Wait	WalkD.	Walk	WalkT.	Ave.
Pavlakos [6]	47.5	50.5	48.3	49.3	50.7	55.2	46.1	48.0	61.1	78.1	51.05	48.3	52.9	41.5	46.4	51.9
Martinez [5]	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
Hossain [8]	35.7	39.3	44.6	43	47.2	54.0	38.3	37.5	51.6	61.3	46.5	41.4	47.3	34.2	39.4	44.1
Lee [3]	34.6	39.7	37.2	40.9	45.6	50.5	42.0	39.4	47.3	48.1	39.5	38.0	31.9	41.5	37.2	40.9
Pavllo [7]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37.2
Cai [1]	32.9	38.7	32.9	37.0	37.3	44.8	38.7	36.1	41.0	45.6	36.8	37.7	37.7	29.5	31.6	37.2
Lin [4]	30.1	33.7	28.7	31.0	33.7	40.1	33.8	28.5	38.6	40.8	32.4	31.7	33.8	25.3	24.3	32.8
UGCN	**23.0**	**25.7**	**22.8**	**22.6**	**24.1**	**30.6**	**24.9**	**24.5**	**31.1**	**35.0**	**25.6**	**24.3**	**25.1**	**19.8**	**18.4**	**25.6**

4 Impact of 2D Pose Estimators

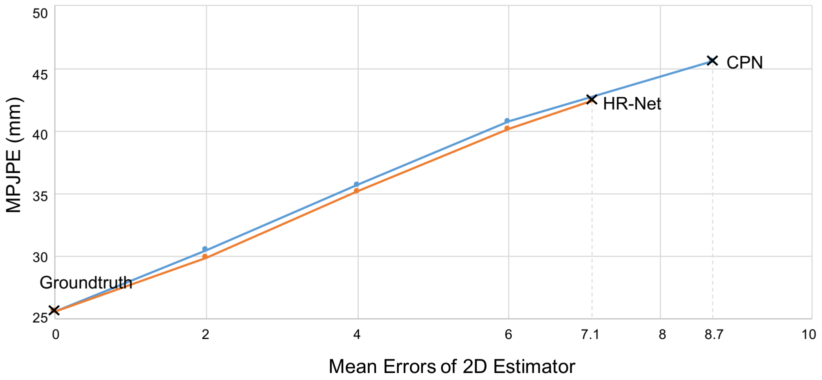

Fig. 1. Relationship between the performace of 3D pose estimation and the accuracy of input 2D poses.

As shwon in Table 6 of the manuscript, we achieved a lower MPJPE when using HR-Net [9] as the 2D pose estimator than using CPN [2]. To explore the impacts of the 2D pose estimator on the final performance, we combined the predicted 2D pose and the groudtruth by weighted addition for simulating a series of new 2D pose estimators. UGCN was trained taking as input these synchronized 2D pose. The results are shwon in the Figure 1. We can observer a near linear relationship between MPJPE of 3D poses and two-norm errors of 2D poses. Curves from two estimators have very similar tendency.

5 Visual Results

Visual results of estimated 3D pose by our UGCN are shwon in the Figure 2. More visualized results can be find in the **supplementary video**, including the following aspects: impacts of motion loss, the comparison with previous works, and the estimation results on noisy 2D poses.

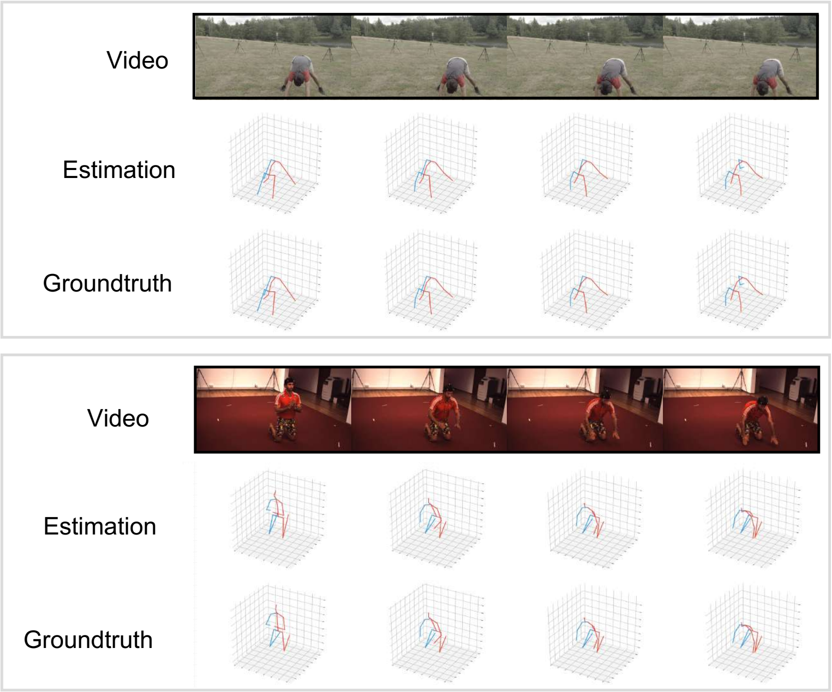

Fig. 2. 3D pose sequences estimated by UGCN on two datasets: MPI-INF-3DHP **(top)** and Human3.6M **(bottom)**.

References

1. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.J., Yuan, J., Thalmann, N.M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2272–2281 (2019)
2. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7103–7112 (2018)
3. Lee, K., Lee, I., Lee, S.: Propagating lstm: 3d pose estimation based on joint interdependency. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 119–135 (2018)

4. Lin, J., Lee, G.H.: Trajectory space factorization for deep video-based 3d human pose estimation. arXiv preprint arXiv:1908.08289 (2019)
5. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2640–2649 (2017)
6. Pavlakos, G., Zhou, X., Daniilidis, K.: Ordinal depth supervision for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7307–7316 (2018)
7. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7753–7762 (2019)
8. Rayat Imtiaz Hossain, M., Little, J.J.: Exploiting temporal information for 3d human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 68–84 (2018)
9. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)