Harvesting Multiple Views for Marker-less 3D Human Pose Annotations Supplementary Material

Georgios Pavlakos[1], Xiaowei Zhou[1], Konstantinos G. Derpanis[2], Kostas Daniilidis[1]
[1] University of Pennsylvania [2] Ryerson University

In this supplementary, we provide material that could not be included in the main manuscript due to space constraints. Section 1 provides additional quantitative evaluation of our approach for multi-view pose estimation, and comparison with the state-of-the-art for HumanEva-I [4]. Section 2 provides full results of the multi-view optimization on Human3.6M after refining the generic 2D pose ConvNet with different annotation sets.

1. HumanEva-I evaluation

We provide additional empirical evaluation on HumanEva-I [4]. While the focus of our work is not explicitly on multi-view 3D human pose estimation, it is interesting to compare the performance on this task with prior work that reports results on this benchmark. We use the same protocol as prior work [5, 1, 2], evaluating on the Boxing and Walking validation sequences of Subject 1. The results are presented in Table 1. We achieve the lowest error for Walking, while we are competitive with the other approaches for Boxing. It is interesting to note that Sigal *et al*. [5], Amin *et al*. [1], and Belagiannis *et al*. [2] train their models on HumanEva-I, while we use a generic ConvNet for joint prediction. Furthermore, Elhayek *et al*. [3] make use of temporal information over the video sequence, while our predictions are based on a single frame.

	Sigal *et al*. [5]	Amin *et al*. [1]	Belagiannis *et al*. [2]	Elhayek *et al*. [3]	Ours
Walking	89.7	54.5	68.3	66.5	**50.2**
Boxing	-	**47.7**	62.7	60.0	61.9

Table 1: Quantitative evaluation on HumanEva-I. Reported results are average 3D joint position errors in mm. Sigal *et al*. [5], Amin *et al*. [1], and Belagiannis *et al*. [2] train their appearance models on HumanEva-I, while we use a generic ConvNet for 2D joint prediction. Elhayek *et al*. [3] make use of temporal information over the video sequence.

2. "Personalization" results on Human3.6M

Here we present extensive quantitative results of multi-view optimization after refining the generic 2D pose ConvNet with different annotations sets of Human3.6M. These results extend Table 2 of the main manuscript which focuses on three actions and the average error due to space constraints. The performance for all actions is presented in Table 2. Using the annotations provided from the proposed multi-view optimization ("PS" and "PS+sel") outperforms other refinement strategies.

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
Generic	41.18	49.19	42.79	43.44	55.62	46.91	40.33	63.68
HM	41.90	49.50	44.05	43.85	57.27	46.94	40.75	57.57
HM+sel	42.00	47.99	43.06	43.44	59.15	45.91	40.71	52.50
PS	39.62	43.32	**40.62**	40.08	**52.77**	42.73	**38.37**	51.32
PS+sel	**39.32**	**43.04**	41.31	**39.76**	53.46	**41.84**	38.39	**45.98**
	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	Average
Generic	97.56	119.90	52.12	42.68	51.93	41.79	39.37	56.89
HM	86.37	100.39	55.86	43.08	49.61	43.07	40.53	55.13
HM+sel	91.49	110.30	52.80	42.90	48.18	43.81	40.32	55.62
PS	79.39	97.26	**48.59**	40.76	52.32	41.54	39.85	51.18
PS+sel	**68.09**	**73.91**	50.63	**39.65**	**42.78**	**39.09**	**37.05**	**47.83**

Table 2: Quantitative comparison of multi-view optimization after refining the ConvNet with different annotation sets and evaluating on Human3.6M dataset. We present results for all actions and overall, extending Table 2 of the main manuscript. The numbers are the average 3D joint error (mm).

References

[1] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view pictorial structures for 3D human pose estimation. In *BMVC*, 2013. 1

[2] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3D pictorial structures for multiple human pose estimation. In *CVPR*, 2014. 1

[3] A. Elhayek, E. Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In *CVPR*, 2015. 1

[4] L. Sigal, A. O. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2):4–27, 2010. 1

[5] L. Sigal, M. Isard, H. W. Haussecker, and M. J. Black. Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. *IJCV*, 98(1):15–48, 2012. 1