Matching and Recovering 3D People from Multiple Views
Supplementary material

Alejandro Perez-Yus Antonio Agudo
Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Spain

Matching	3DPS	PCP5	PCP2	MPJPE	PCK_{50}	PCK_{100}	PCK_{150}
[1]	[1]	81.64	44.11	93.5	37.16	68.45	80.65
[1]	Ours	86.00	47.80	86.2	33.78	76.25	86.84
Ours	[1]	88.23	49.30	75.2	**39.16**	75.14	88.40
Ours	Ours	**90.57**	**50.31**	**70.7**	34.76	**80.33**	**91.57**

Table 1. **Quantitative evaluation of our method in CMU Panoptic dataset**. Here we show all metrics used in this work (all in percentages except MPJPE, in mm) and a comparison of our results with [1].

1. Quantitative esults in CMU Panoptic

Apart from *Campus*, *Shelf*, and *KTH Football II*, we have also evaluated our method in the *CMU Panoptic* dataset [2]. This is a huge dataset that includes 480 VGA cameras (640 × 480), 31 HD cameras (1920 × 1080) and other types of sensors, arranged in a *dome* structure, where individuals are performing several actions. The ground truth 3D pose of all actors involved is provided.

Since there is no widespread protocol of evaluation like with the aforementioned datasets, we performed our own experiments to evaluate our method, and compared our results with [1], using their code. In particular, we used the scene *ultimatum1*, selecting 500 frames (18900-19400), where the scene is populated with up to 7 actors; and we chose five of the several first automatically downloaded VGA cameras (01_01, 03_12, 04_07, 06_15, 18_13) so that the whole scene could be mostly observed. However, since the 3D ground truth poses are not obtained from said cameras, not all human joints were observable by all cameras at all times. Thus, we projected ground truth points towards each camera in order to ignore ground truth poses where not all joints were inside the image in at least two camera views. Otherwise, it would be impossible to recover the depth and the pose would be corrupted. This approach makes the experiment similar to the previous datasets, considering the ground truth poses provided were always clearly observable by several views in full.

The results of all metrics described in this work are shown in Table 1. Due to discrepancies in the way joints are represented, we do not count with head and neck joints for MPJPE and PCK metrics (just the three joints from each arm and leg). Besides, unlike with Campus and Shelf, we provide PCP data as the global average of all poses, instead of the average of the PCP of each actor involved. Additionally, to provide a more complete insight of our contributions, we have cross-evaluated our matching algorithm and 3DPS with the proposed counterparts in [1], to see how much each part contributes to the end result. We can see that our 3DPS obtains much better results even with their matching, which shows that our 3DPS with physico-geometric constraints is more capable of addressing errors in the matching part. Moreover, using our robust matching, even with their 3DPS, we obtain even better results, showing again the importance of a good matching. In the end, our complete pipeline clearly outperforms all other combinations, showing the efficacy of our method against the competing method. It is important to note that we performed these experiments without fine-tuning or modifying anything specifically for this dataset, which shows that, overall, our method generalizes well to new data.

2. Video attachment

With this submission of supplementary material, we also provide a video attachment. The intention of this video is twofold: make the explanation more visual and easy to understand, and show our complete results in full sequences. Particularly, we show our matching and 3D pose results in Campus and Shelf, and extended sequences of both datasets in the end, as well as all sequences from KTH Football II, and the frames evaluated in CMU Panoptic as described in previous section. For all sequences, it is possible to see the results of the matching, the 3D pose recovered, and the projection of said pose back to the image.

References

[1] J. Dong, W. Jiang, Q. Huang, H. Bao, and X. Zhou. Fast and robust multi-person 3D pose estimation from multiple views. In *CVPR*, 2019.

[2] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *TPAMI*, 41(1):190–204, 2017.