

# Efficient ConvNet-based Marker-less Motion Capture in General Scenes with a Low Number of Cameras

A. Elhayek  
MPI Informatics

E. de Aguiar  
MPI Informatics

A. Jain  
New York University

J. Tompson  
New York University

L. Pishchulin  
MPI Informatics

M. Andriluka  
Stanford University

C. Bregler  
New York University

B. Schiele  
MPI Informatics

C. Theobalt  
MPI Informatics

As shown in the paper, we evaluated our approach using the Box and Walk sequences from the HumanEva benchmark [3] and compared the results against Sigal et al. [4], Amin et al. [2] and Belagiannis et al. [1]. Table 4 in the paper, reproduced here as Table 1, summarizes the comparison results. As seen in the table, Amin et al. [2] shows very low average error but we also achieve similar results using our hybrid approach, outperforming the other methods. However, it is also important to consider the motion reconstruction quality over time and not only the average 3D joint position error. In the accompanying video, it is shown that our method presents a better temporal reconstruction when compared to Amin et al. [2]. Our results are more stable, presenting a good temporal coherent reconstruction over time. In contrast, [2] shows a considerable amount of jittering and wrong detections (i.e. jumps) over time.

Figure 1 plots the average 3D joint position error for the Box sequence for both approaches. Note that our approach presents a constant error level, with a lower variance. We believe that part of this error is coming mostly from a different initial joint configuration in our approach, i.e. bone lengths and joint locations. In contrast to Amin et al. [2], we do not train our model on the HumanEva dataset. Figure 2 shows our skeleton and the ground truth joint positions overlaid in the input images for all three camera views for the Box and Walk sequences. Note that for the error calculation we only use the skeleton joints that exist in HumanEva, which is less than the total number of joints our standard skeleton has. In the figure, the red and green joints are our reconstructed joints and the blue and pink joints are the ground truth information. Note that although our joint positions are matching the real underlying human skeleton better, our skeleton configuration (i.e. bone lengths and joint locations) is not the same as in the HumanEva skeleton.

We argue that our increased 3D joint position error value is partly due to the dimensions of our skeleton not matching exactly the dimensions of the HumanEva skeleton, so the error contains a constant offset. Also, please note that the

marker positions in HumanEva (on the surface of the actor) are not identical to joint positions (inside the body) which causes an offset anyways. We believe that with this observation and the high temporal stability of our approach, our results are of high quality.

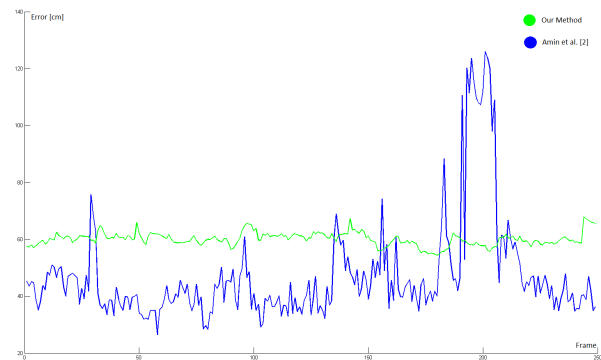


Figure 1. Plot showing the average 3D joint position error for the Box sequence using our approach (green curve) and Amin et al. [2] (blue curve).

Table 1. Comparison between the average 3D joint position error for the HumanEva Walk and Box sequences.

| Sequence               | Walk [cm] | Box [cm] |
|------------------------|-----------|----------|
| Amin et al. [2]        | 5.45      | 4.77     |
| Sigal et al. [4]       | 8.97      | -        |
| Belagiannis et al. [1] | 6.83      | 6.27     |
| Our approach           | 6.65      | 6.00     |

## References

- [1] *3D Pictorial Structures for Multiple Human Pose Estimation*, June 2014. 1

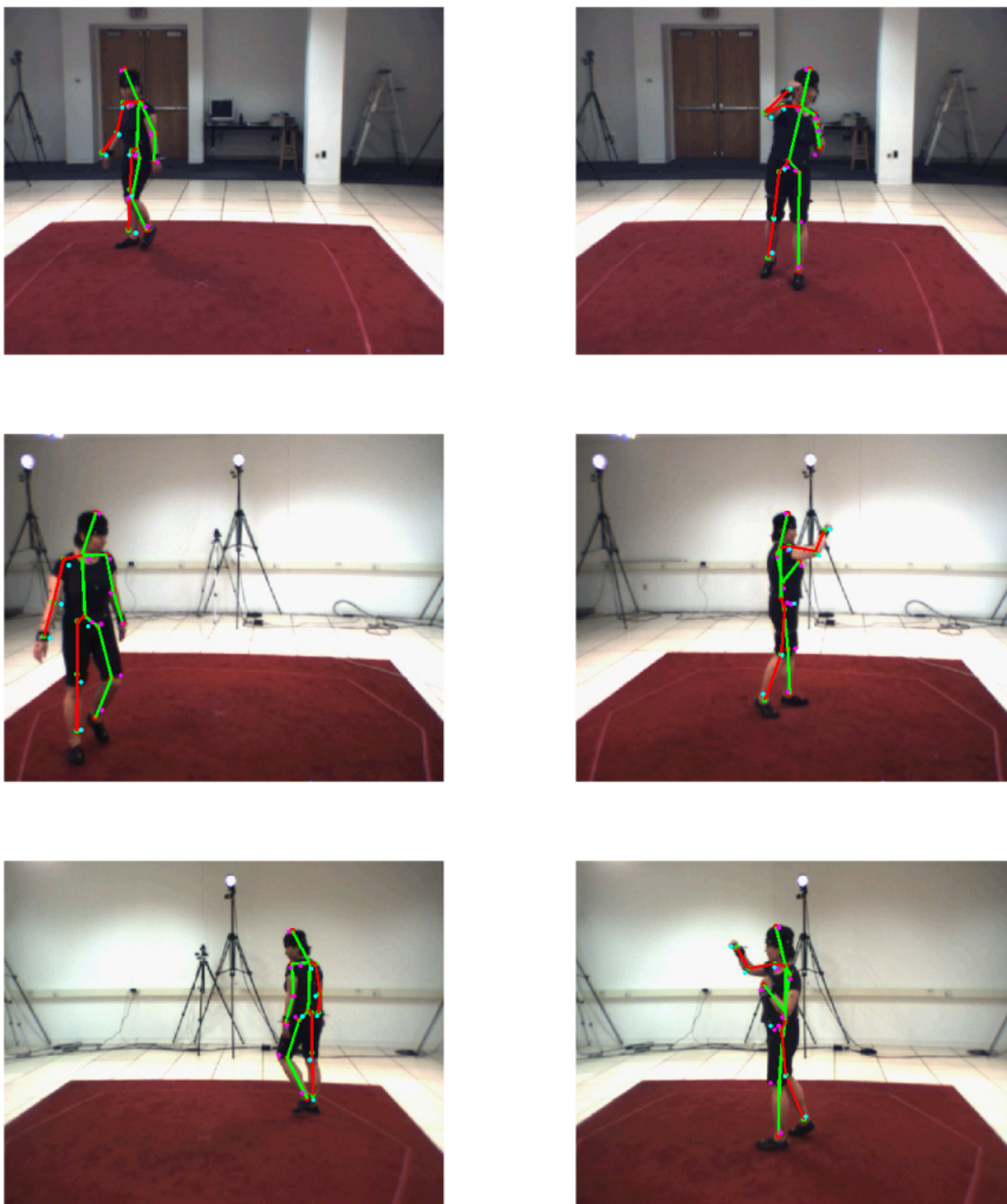


Figure 2. Differences between our initial skeleton configuration and the HumanEva skeleton configuration (our ground truth) - Box (left column) and Walk (right column) sequences - can cause an increase in our 3D average joint position error. In the figures, the red and green joints are our reconstructions and the blue and pink joints are the ground truth positions.

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

[3] L. Sigal, A. Balan, and M. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87:4–27, 2010.

1

- [4] L. Sigal, M. Isard, H. Haussecker, and M. Black. Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. *International Journal of Computer Vision*, 98(1):15–48, 2012. 1