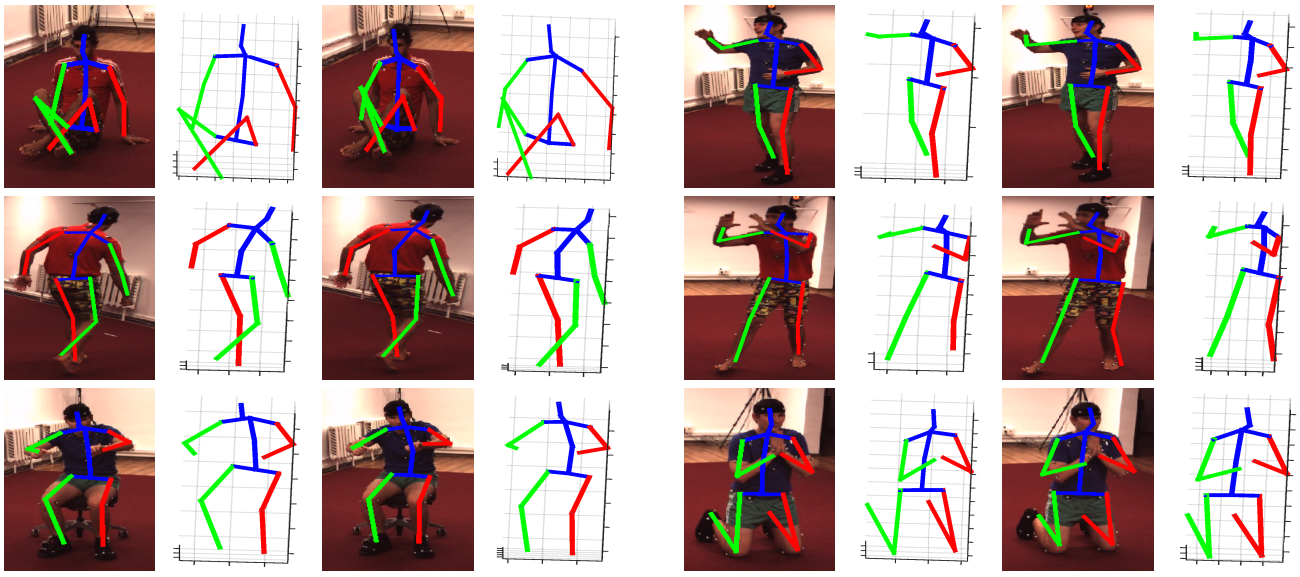
# Supplementary Material for
# *"Direct Prediction of 3D Body Poses from Motion Compensated Sequences"*

Bugra Tekin[1] Artem Rozantsev[1] Vincent Lepetit[1,2] Pascal Fua[1]
[1]CVLab, EPFL, Lausanne, Switzerland, {firstname.lastname}@epfl.ch
[2]TU Graz, Graz, Austria, lepetit@icg.tugraz.at



(a) Ground Truth (b) Our method (c) Ground Truth (d) Our method

Figure 1. 3D human pose estimation on Human3.6m for several different action categories **(a,c)** Ground-truth 3D poses and their projection in the images. **(b,d)** The skeletons recovered by our method and their projection on the image plane. We can reliably recover the 3D pose of the body in case of ambiguities, such as self-occlusions and mirroring. Best viewed in color.

In Section 1, we describe implementation details of our motion compensation algorithm. Then, in Section 2 we provide further visualizations and analysis that did not fit in the main submission because of space limitations. More visualizations can be found in the accompanying videos.

## 1. Implementation Details for Motion Compensation with CNNs

**Centering the body.** As explained in Section 3.3 of our main submission, we train Convolutional Neural Networks to predict the shifts of the person from the center of the bounding box. In order to obtain training images centered on the subject, we use the foreground masks that are part of the datasets to compute and center the bounding box at the root position of the person.

**Scale of the person.** We rely on masks' height to compute the scale of the person. While it gives only a rough estimate, it is sufficient to handle scale changes when they occur. In the future, we plan to train a single regressor to compensate for both shift and scale changes.

**Initialization.** In our approach, DPM (trained on VOC 2010) is used in the first image of a sequence to provide an initial estimate of the bounding box. The initial person detector provides rough location estimates of the person and our motion

compensation algorithm naturally compensates even for relatively large positional inaccuracies using the regressor, $\psi_{coarse}$, as can be seen in Fig. 2 and the accompanying videos.

Figure 2. Examples of our motion compensation algorithm. For each pair of images, the left one depicts the initial bounding box, and the right one depicts the aligned bounding box using our motion compensation algorithm.

**Motion Compensation vs. Centering the Detections at Each Frame.** For the alignment of the body across time, it is also possible to compute the center of the root part of a DPM-based pose estimator. However, it is time-consuming and computationally heavy to detect body parts at each frame. Therefore, instead of computing DPM in all the frames of the sequence, we use it *only in the first frame* and use motion compensation to iteratively center the body in subsequent frames. This is a more efficient and elegant solution than detecting body parts at each time instant. In order to justify the efficiency of our CNN-based motion compensation approach (CNN-MC), we compare its timing to that of DPM detections in Table 1. Additionally, we provide comparisons to the timings of conventional optical-flow techniques [4, 9, 10]. We show that our approach to image alignment is substantially faster than these approaches.

| Method: | Time (sec) |
|---|---|
| DPM [6], run sequentially | 9.645 |
| Large Displacement Optical Flow [4] | 0.967 |
| Lucas-Kanade Optical Flow [9, 10] | 0.140 |
| CNN-MC | 0.006 |

Table 1. Timings (in seconds per image) of our motion compensation algorithm (CNN-MC) in comparison to DPM [6] and optical-flow [4, 9, 10]. Our motion compensation algorithm aligns the body in subsequent frames by shifting the body to the center of the bounding box. DPM takes the center of the root part of the part detector at each frame. Our approach to aligning the body is orders of magnitude faster.

**Temporal Heuristic.** Although we treat each frame independently, we exploit additional temporal heuristic in our approach by initializing the motion compensation algorithm using the bounding box from the previous frame. Simultaneous motion compensation in multiple frames not only increases model complexity but could yield incorrect estimates when the motion direction changes fast.

## 2. Further Analysis and Visualizations

In this section, we provide additional analysis of our experimental results and further visualizations for our 3D body pose recovery method [1].

**Evaluation.** The parameters of Deep Network regressor are cross-validated on a validation set and used for all the actions in the dataset. We consider the average error excluding the first and last $T/2$ frames (0.24 seconds for $T = 24$ at 50 fps) to evaluate the performance.

**Additional Comparisons on HumanEva-I.** On HumanEva [11], we trained our regressors on training sequences of subjects S1, S2 and S3 and evaluated on the "validation" sequences as in [2, 3, 5, 12] as explained in Section 4.2. [1, 14] followed a different experimental procedure where they use the same subject for training and testing purposes. In order to compare our results to these baselines as well, we employ the same subject-specific experimental setup and provide analysis in Table 2. The results demonstrate that our method yields state-of-the-art 3D human pose estimation accuracy, also with this experimental setting.

**Additional Comparisons on Human3.6m.** [8] is a recently published structured deep learning method and uses the correlations among joint points for 3D human pose estimation. As shown in the main submission, we outperform all pose estimation methods on Human3.6m, HumanEva or KTH Multiview Football II that do not use structural dependencies. In

[1] More visualizations can be found in the accompanying videos.

| Method: | S1 | S2 | S3 | Average |
|---|---|---|---|---|
| Yao et al. [14] | 41.6 | 64.0 | 46.5 | 50.7 |
| Amin et al. [1] | 56.7 | 52.1 | 62.4 | 57.1 |
| Ours | 38.4 | 27.9 | 52.1 | **39.5** |

Table 2. 3D joint position errors (in mm) on the *Walking* sequences of HumanEva-I. We compare our approach against [1, 14].

Table 3, we further show that we also outperform [8] on average over the action classes for which the authors reported accuracy numbers even though our algorithm do not rely on using the dependencies among the human body parts.

| Method: | Discussion | Eating | Greeting | Taking Photo | Walking | Walking Dog | Average |
|---|---|---|---|---|---|---|---|
| Li et al. [8] | **136.88** | 96.94 | **124.74** | **168.68** | 69.97 | 132.17 | 121.56 |
| RSTV+DN (Ours) | 147.72 | **88.83** | 125.28 | 182.73 | **55.07** | **126.29** | **120.98** |

Table 3. 3D joint position errors (in mm) on Human3.6m. We compare our approach against [8].

**Stability.** 3D pose predictions of our approach are stable as can be seen in accompanying videos, as they are obtained for each overlapping temporal window with 1 frame shift. The comparison in Fig. 3 further demonstrates that RSTV+DN obtains the most stable and accurate predictions.

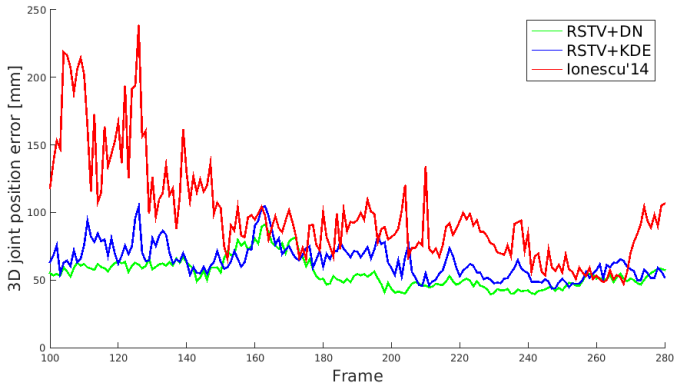
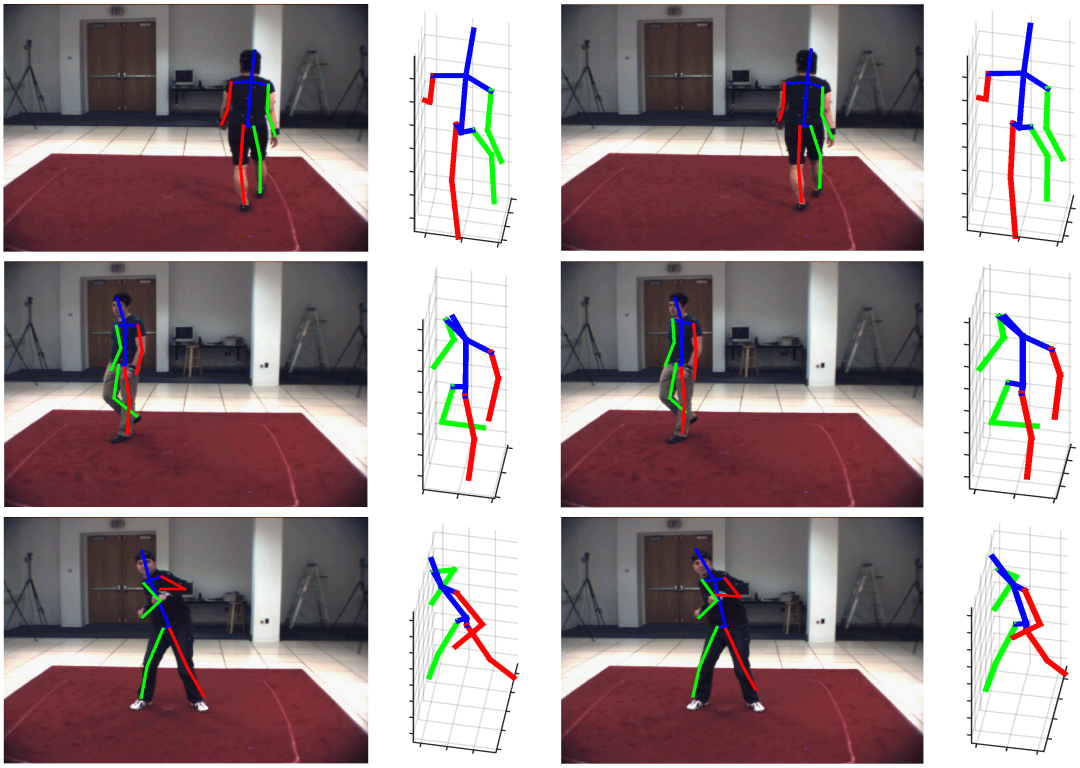



Figure 3. 3D joint position errors across frames for the *Walking* sequence corresponding to Subject 9, Trial 1, Camera 1 in the Human3.6m dataset. The results of [7] compared to our methods RSTV+KDE and RSTV+DN. RSTV+DN yields the best accuracy on average with the added advantage of temporal consistency. Best viewed in color.

**Generalization.** We demonstrate the generalization ability of our approach in the following ways.

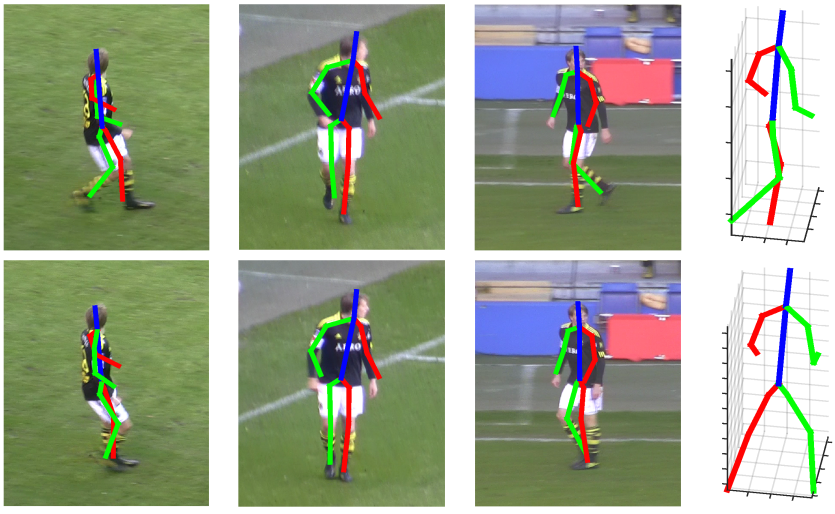
- HumanEva-II provides only a test dataset and no training data, therefore, we trained our regressors on HumanEva-I using videos captured from different camera views;
- Data from Human3.6m exhibit large variations in terms of body shapes, clothing, poses and viewing angles within and across training/test splits [7]. Also, different people appear in the training and test data.
- The size of the training set in HumanEva is too small to train a deep network. However, we tested on HumanEva-I using Deep Network regressors trained on Human3.6m and report a pose estimation accuracy of 75.4 mm. on Subject 1. As skeleton configurations for Human3.6m and HumanEva do not exactly match each other, the error contains a constant offset. However, we still obtain accurate pose estimates and outperform [13], which reports a 99.6 mm. accuracy, even though it is trained on HumanEva.

**Visualization.** We provide additional qualitative results for the Human3.6m, HumanEva and KTH Multiview Football II datasets in Fig. 1, Fig. 4 and Fig. 5.



(a) Ground Truth (b) Our method

Figure 4. 3D human pose estimation on HumanEva. The rows correspond to the *Walking* and *Box* actions. **(a)** Reprojection in the images and ground-truth 3D pose. **(b)** The skeletons recovered by our method and their projection on the image plane. Best viewed in color.



(a) Camera 1 (b) Camera 2 (c) Camera 3 (d) 3D Pose

Figure 5. Results on KTH Multiview Football II. The 3D skeletons are recovered from Camera 1 images and projected on those of Camera 2 and 3, which were not used to compute the poses. Best viewed in color.

## References


[1] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-View Pictorial Structures for 3D Human Pose Estimation. In *BMVC*, 2013.

[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.

[3] L. Bo and C. Sminchisescu. Twin Gaussian Processes for Structured Prediction. *IJCV*, 2010.

[4] T. Brox and J. Malik. Large Displacement Optical Flow: Descriptor Matching in Variational Motion Estimation. *PAMI*, 2011.

[5] 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.

[6] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. *PAMI*, 2010.

[7] C. Ionescu, I. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *PAMI*, 2014.

[8] S. Li, W. Zhang, and A. B. Chan. Maximum-Margin Structured Learning with Deep Networks for 3D Human Pose Estimation. In *ICCV*, 2015.

[9] B. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *IJCAI*, pages 674–679, 1981.

[10] D. Park, C. L. Zitnick, D. Ramanan, and P. Dollár. Exploring Weak Stabilization for Motion Feature Extraction. In *CVPR*, 2013.

[11] L. Sigal, A. Balan, and M. J. Black. Humaneva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. *IJCV*, 2010.

[12] 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*, 2012.

[13] E. Simo-Serra, A. Ramisa, G. Alenya, C. Torras, and F. Moreno-Noguer. Single Image 3D Human Pose Estimation from Noisy Observations. In *CVPR*, 2012.

[14] A. Yao, J. Gall, L. J. Van Gool, and R. Urtasun. Learning Probabilistic Non-Linear Latent Models for Tracking Complex Activities. In *NIPS*, 2011.