Supplemental Material: Video Based Reconstruction of 3D People Models

Thiemo Alldieck[1] Marcus Magnor[1] Weipeng Xu[2] Christian Theobalt[2] Gerard Pons-Moll[2]

[1]Computer Graphics Lab, TU Braunschweig, Germany
[2]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
{alldieck,magnor}@cg.cs.tu-bs.de {wxu,theobalt,gpons}@mpi-inf.mpg.de

1. Implementation Details

In this section, we present more implementation details of the presented method.

1.1. Optimization Parameters

The presented results are calculated using two empirically determined parameter sets: one for clothed subjects, one for subjects in minimal clothing. We found that the results are not very sensitive to optimization parameter weights and we select them so that the energy terms are balanced. The consensus objective function is defined as:

$$E_{\text{cons}} = E_{\text{data}} + w_{\text{lp}} E_{\text{lp}} + w_{\text{var}} E_{\text{var}} + w_{\text{sym}} E_{\text{sym}} \quad (1)$$

The method is initialized with $w_{\text{lp}} = 4.0$, $w_{\text{var}} = 0.6$ and $w_{\text{sym}} = 3.6$. For subjects in minimal clothing, we enforce a smoother surface with initializing $w_{\text{lp}} = 6.5$. We minimize E_{cons} with respect to model parameters and offsets. We update the point-to-line correspondences during optimization. An interesting direction to explore would be to extend [3] to continuously optimize line to surface correspondences, model parameters and offsets. In this work, we recompute correspondences during optimization. After each correspondence step, we re-initialize the three regularization terms E_{lp}, E_{var} and E_{sym}. To capture personal details, we gradually decrease the regularization weights.

1.2. Computation Time and Complexity

The results are calculated with Python code without highly parallel computation. No attempts for run-time optimization have been made. On an Intel Xeon E5-1630 v4 processor, the run-time for one frame of pose reconstruction is about 1 min including IO. Consensus shape estimation, meaning correspondence calculation and subsequent optimization on $F = 120$ frames, takes about 1:50 min.

Given, that the connectivity of the mesh is fixed and the maximum connectivity is bounded by constant k, the complexity of the regularization falls into $\mathcal{O}(N)$. As every new frame introduces more matches, the complexity of the optimization falls into $\mathcal{O}(FNP)$, with P being the number of pixels (upper bound for silhouette).

2. Scale Ambiguity

Scale is an intrinsic ambiguity in monocular methods when the distance of the person to the camera is not known. Multiple views of the person in different poses help to mitigate the problem but we have observed that the ambiguity remains. The reason is that pose differences induce additional 3D ambiguities which cannot be uniquely decoupled from global size, even on multiple frames. Therefore, we perform an evaluation that is not sensitive to scale. Before calculating the per-vertex point to surface error, we adjust the one-dimensional scale parameter to match the ground truth. This step is necessary to evaluate the quality of the shape reconstructions as otherwise, almost all error would come from the scale miss-alignment.

3. Comparison with the Depth Camera Based Approach [1]

We compare our method against state-of-the-art RGB-D based approach [1] on their dataset which we refer to as KinectCap in the main paper. To make a fair comparison we also adjust the scale of their result to match the ground truth. In the original paper, they performed an evaluation that was based on scan to reconstructed mesh distance. Since the scan contains noise they had to filter out noise by not considering scan points that are further away than a given threshold. We tried to make the fairest comparison possible so we report in the main paper their result using this method, which was 2.54cm. Since we did not know what threshold to use to filter out noise in the scan and since different scan point sampling/density can produce very different results we followed the strategy explained in the main paper which was also followed in [4]. We first perform non-rigid registration regularized by the body model to obtain a

ground truth registration (since registrations are regularized, they do not contain the noise in the scans). Then we compute a bi-directional surface to surface distance from the ground truth registration to the reconstructed shape. Following this strategy, their method achieves an accuracy of 3.2cm and ours 3.9cm. Our monocular approach is still not as accurate as approaches that use a depth camera [1] but produces comparable results despite using only a single RGB camera.

4. More results

We show all 9 reconstruction results on image sequences rendered from the DynamicFAUST dataset in Fig. 1, and all 9 results from the BUFF scans in Fig. 2. It is worth noticing that the segmentation masks obtained from the scans in the BUFF dataset contain noise and missing data, which degrades the reconstruction quality of our method, especially for head, hands and feet. In addition, the pose reconstruction for the hip motion is less accurate than for people turning around. Note that the hip motion (in DynamicFAUST and BUFF) is probably not the most suitable motion pattern to reconstruct a static 3D person model but it allowed us to evaluate our approach numerically. Thus, the results using the rendered images of BUFF and DFAUST are slightly worse than results obtained with a real RGB camera. All the 24 reconstructed models in the People-Snapshot dataset are shown in Fig. 3.

References

[1] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *IEEE International Conf. on Computer Vision*, pages 2300–2308, 2015. 1, 2

[2] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 3

[3] J. Taylor, L. Bordeaux, T. Cashman, B. Corish, C. Keskin, T. Sharp, E. Soto, D. Sweeney, J. Valentin, B. Luff, et al. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Transactions on Graphics*, 35(4):143, 2016. 1

[4] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 1, 4

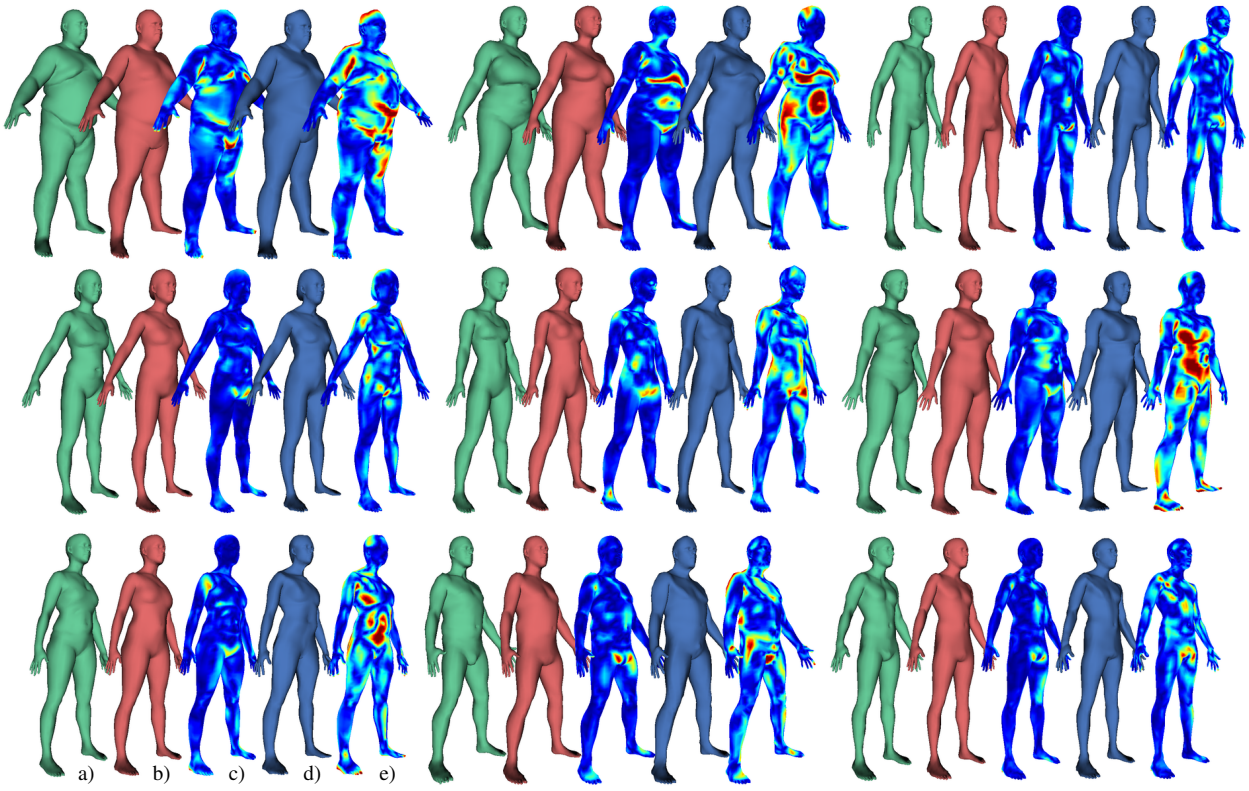

Figure 1. Our results on image sequences from D-FAUST [2]. (a) ground truth 3D scan, (b) consensus shape with ground truth poses (consensus-p), (c) consensus-p heatmap, (d) consensus shape (consensus), (e) consensus heat-map (blue means 0mm, red means $\geq$ 2cm).

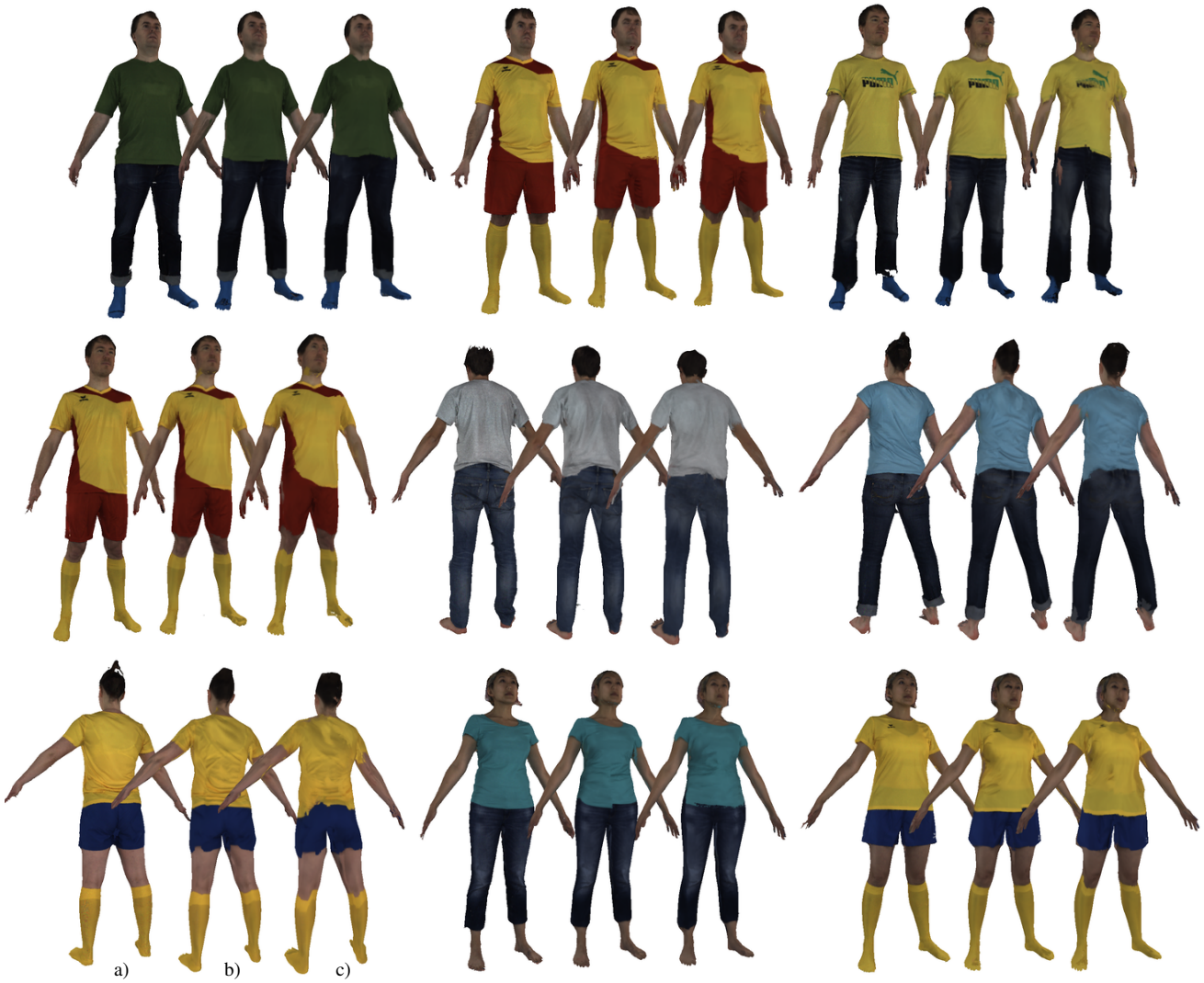

Figure 2. Our results on image sequences from BUFF [4]. (a) ground truth scan, (b) consensus shape with ground truth poses and texture, (c) consensus shape with texture.

Figure 3. Results on our People-Snapshot dataset. We blurred the faces for the subjects that did not give consent.