Human Shape and Pose Tracking Using Keyframes: Supplementary Material

Chun-Hao Huang[§], Edmond Boyer[†], Nassir Navab[§], Slobodan Ilic[§]

[§]Department of Computer Science, Technische Universität München
[†]LJK-INRIA Grenoble Rhône-Alpes
{huangc,slobodan.ilic,navab}@in.tum.de, edmond.boyer@inria.fr

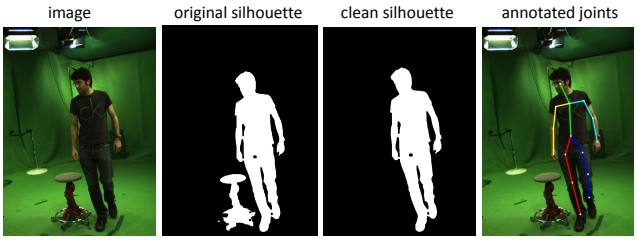


Figure 1. Example images, generated clean silhouettes, and annotated joint positions of *WalkChair*.

The supplementary material for the paper *Human Shape and Pose Tracking Using Keyframes* consists of this document and the accompanying video. It provides more details on the newly recorded sequences and more analysis on the experiment results.

1. New recorded sequences

In Fig. 1, we show one example frame of the newly recorded sequences. The occluding object, *i.e.* the chair, is kept after background subtraction, and therefore remains in the subsequent reconstructed point cloud. The reference surfaces at $t = 0$ is the smoothed reconstructed visual hulls. There is no need to register the surface to the point cloud with a rigid transformation to initialize the tracking.

We produce two different types of ground truth for evaluating shapes and poses, respectively. For shape evaluation, we remove the silhouettes of irrelevant objects manually, if they are not connected to the subjects, as shown in Fig. 1. The associated metric is the standard *silhouette overlap error* which measures the discrepancies between the contour of the projected surface and the contour in the observed silhouettes. To evaluate the estimated poses, we annotate the positions of joints in five cameras, and see how close to them the estimated joints are (*2D joint error*). The sequences and the associated ground truths will be publicly available upon publication.

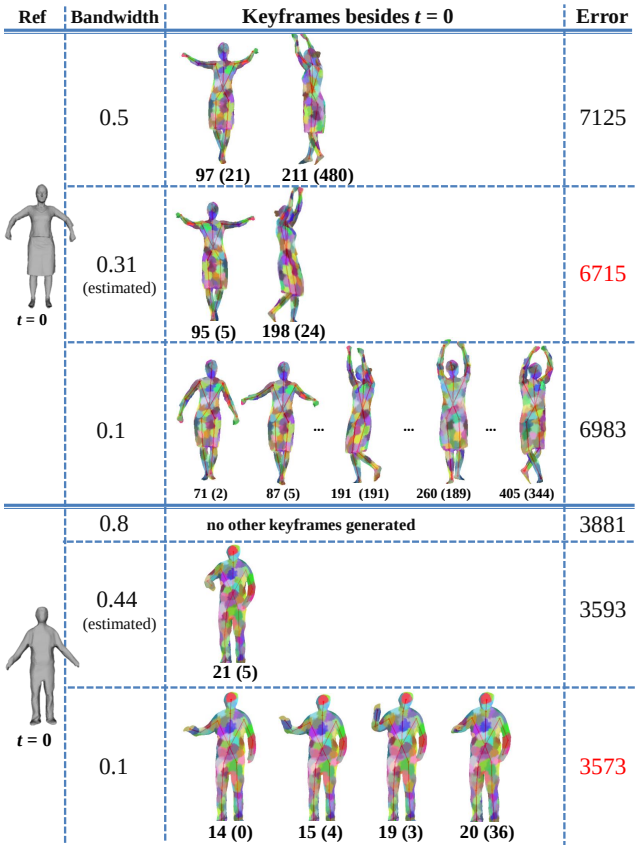
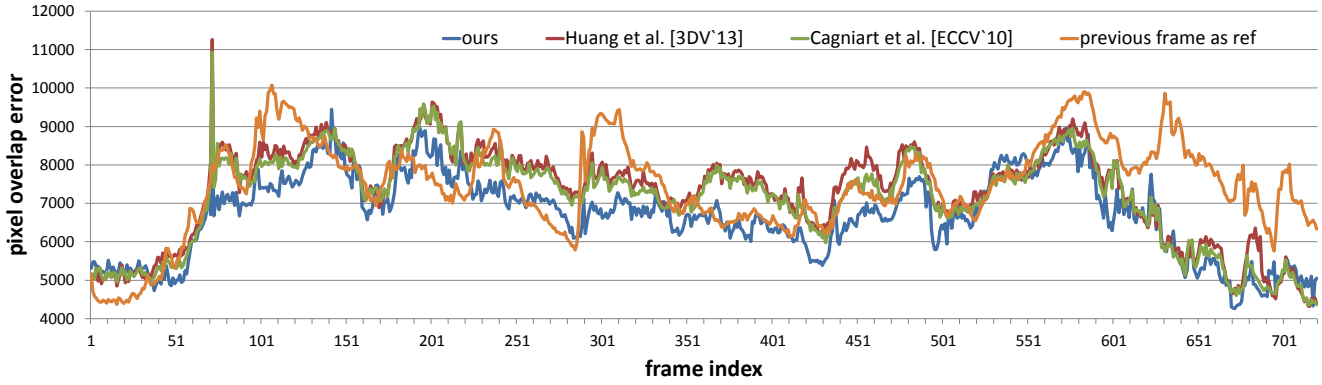


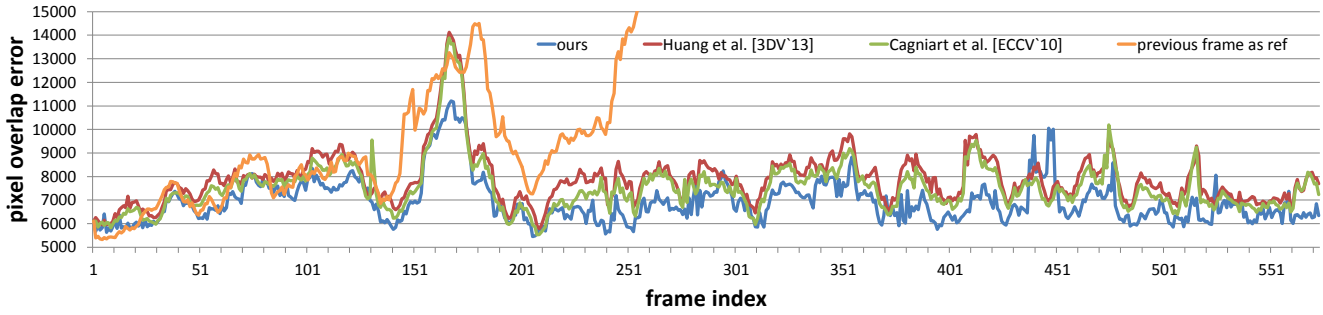
Figure 2. Generated keyframe pool of *Skirt* [2] (top) and *HammerTable* (bottom) in varying mean-shift bandwidths.

2. Supplementary results

Influence of mean-shift bandwidth. In Fig. 2 we visualize the generated keyframe pools of *Skirt* and *HammerTable* in different bandwidths. Two sequences are chosen because the subjects repeat the actions. With small bandwidths, we observe many similar key poses, which however does not guarantee smaller error. With the estimated bandwidths we



(a) *Skirt*



(b) *Dance*

Figure 3. Pixel overlap error of *Dance* and *Skirt* [2] in each frame, averaged over 8 cameras. Image resolution: 1004×1004. **Blue: ours**. Green: Cagniart *et al*. [1]. Red: Huang *et al*. [3]. Orange: using the previous frame as the reference model.

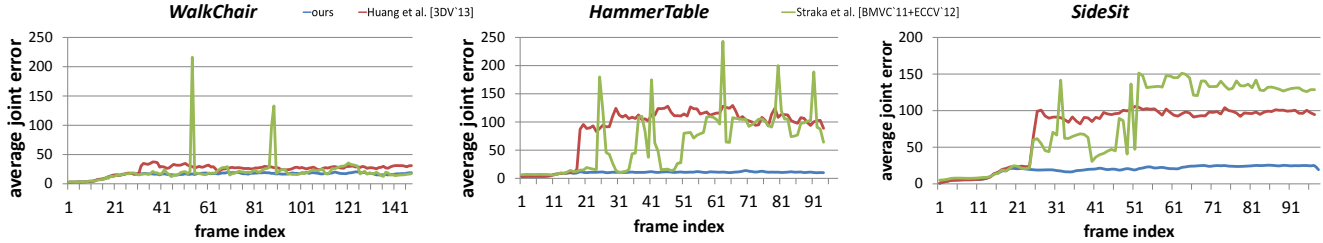


Figure 4. The curves of 2D joint error of three newly recorded sequences. Image resolution: 1000×1000. **Blue: ours**. Green: Straka *et al*. [4] + [5]. Red: Huang *et al*. [3].

not only obtain distinctive key poses but also provide comparable performance.

Further quantitative analysis. Table 2 in the main paper shows the overall average pixel overlap error of *Dance* and *Skirt*. In Fig. 3, we report the error in each frame. Broadly speaking, our approach attains smaller error over the whole sequences, compared with Cagniart *et al*. [1] and Huang *et al*. [3]. In Fig. 4, we further report the 2D joint error of *WalkChair*, *HammerTable*, and *SideSit*. We see that while [3] fails to track at a certain point, and Straka *et al*. [4] + [5] produces sporadic high errors, our approach obtains consistent low error over sequences.

To further justify the advantage of our keyframe-based framework, we make a comparison with following two strategies:

1. Adhering to $t = 0$ as the reference model.

2. Adhering to previous frame as the reference model.

The benefit of our approach over the first strategy (*i.e*. ref: $t = 0$) is already presented in Fig. 3, Fig. 6, and the corresponding text in the main paper. Here we concentrate on comparing with the 2nd strategy, which always uses the tracked result of previous frame as the reference model for the current frame. In Fig. 5(a-c), we overlay the corresponding results of $t = 102$ in *Skirt* sequence. For this frame only, using the previous frame result as reference actually

(a) ref: $t = 0$ (b) ref: prev. ($t = 101$) (c) ours (ref: $t = 95$) (d) prev. $t = 31$ (e) $t = 256$ (f) $t = 462$

Figure 5. Comparison of three different strategies (a-c), and the disadvantages of using always the previous frame result as the reference model (d-f). Better to be viewed in the pdf file.

yields smallest error. We demonstrate in Fig. 5(d-f) the potential drawback of this strategy: drifting. We see that the blue patch is supposed to be at the back side of the subject ($t = 31$), but it moves *along* the surface embedding during tracking, and ends up at the front side of the body ($t = 462$). In the very beginning of the tracking, drifting is difficult to be observed via overlap error because the silhouette does not differ too much (orange curves in Fig. 3). However, as the errors accumulates, drifting gradually deteriorate the results, and eventually leads to noticeably large errors (*Skirt*), or even a tracking failure (*Dance*).

Generated keyframe pool. We show the identified keyframes of all testing sequences and the associated estimated bandwidth (BW) in Fig. 6. Thanks to the way we create virtual samples, we do not observe duplicate keyframes in the same sequences, and the delay time are all within acceptable range.

Further qualitative results. In Fig. 7, we further demonstrate the effectiveness of our approach on taking care of outliers and missing data. In Fig. 7(b), we observe that the hand of the subject is connected to the table in both the silhouette and the point cloud. Such observations confuse methods like [4] which results in the high peak error in Fig. 4, whereas our method still estimates the pose and the shape successfully. In Fig. 7(c), despite that the ball observations have close interaction with the subject, we still obtain correct shape around the right leg. In Fig. 7(d), we see that our method properly handles merging body parts (the right hand), and excludes outliers, while [1] does not manage to do so.

References

[1] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic deformable surface tracking from multiple videos. In *ECCV*, 2010.

[2] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *CVPR*, 2009.

[3] C.-H. Huang, E. Boyer, and S. Ilic. Robust human body shape and pose tracking. In *3DV*, 2013.

[4] M. Straka, S. Hauswiesner, M. Ruether, and H. Bischof. Skeletal graph based human pose estimation in real-time. In *BMVC*, 2011.

[5] M. Straka, S. Hauswiesner, M. Rüther, and H. Bischof. Simultaneous shape and pose adaption of articulated models using linear optimization. In *ECCV*, 2012.

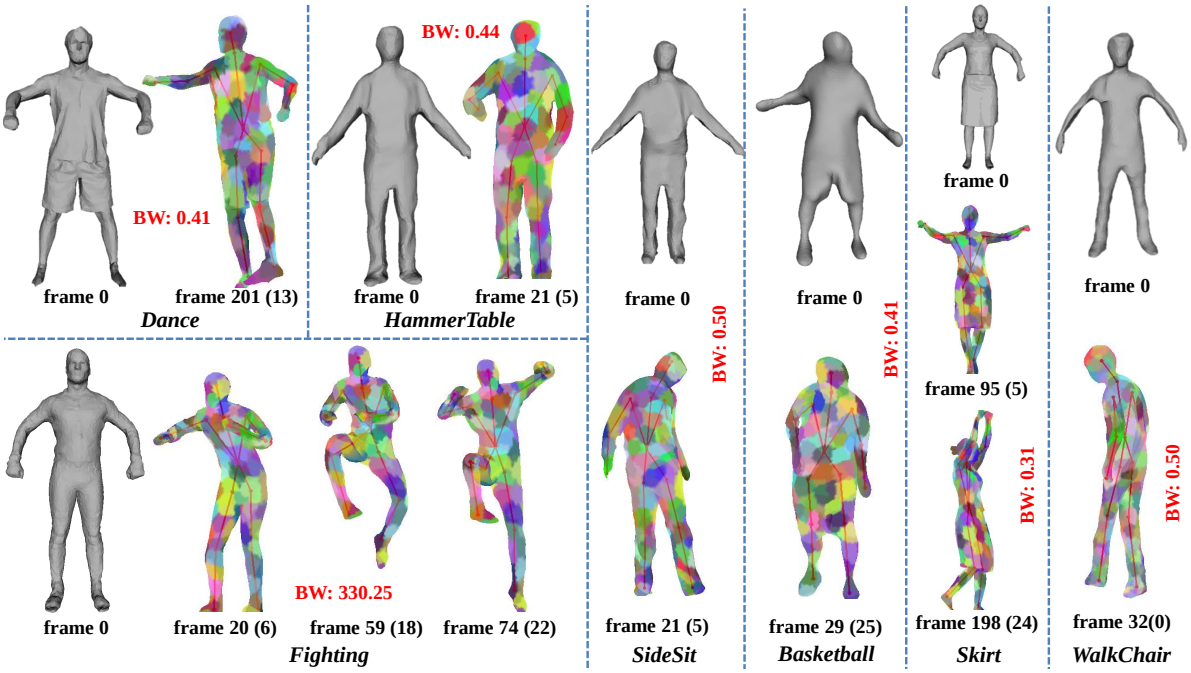


Figure 6. Generated keyframe pool of all testing sequences. Numbers in the parenthesis are the delay time.

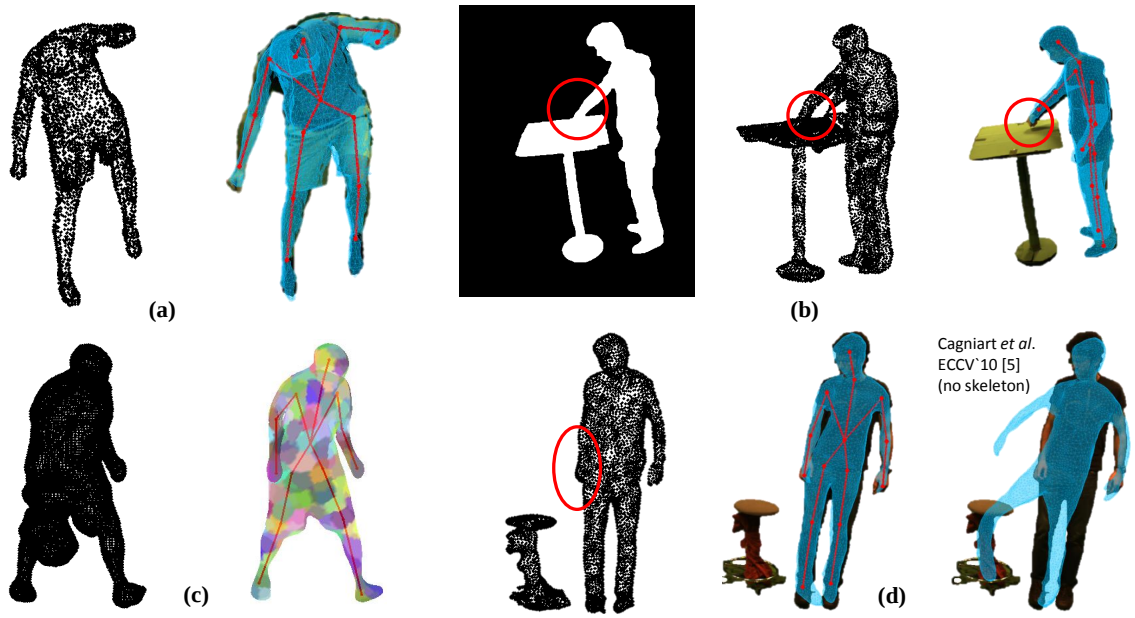


Figure 7. Results of (a) *Dance*, (b) *HammerTable*, (c) *Basketball*, and (d) *WalkChair*. Black dots are the point clouds.