— Supplemental Document —
General Automatic Human Shape and Motion Capture Using Volumetric Contour Cues

Helge Rhodin[1] Nadia Robertini[1, 2] Dan Casas[1]
Christian Richardt[1, 2] Hans-Peter Seidel[1] Christian Theobalt[1]

[1]MPI Informatik, [2]Intel Visual Computing Institute

1 Sequence details

Table 1 lists details on all input sequences, including the complexity of the used models and optimization results.

Sequence	Source	# Cameras	# Frames	Motion	Environment	Ground truth	Input resolution	Stage I runtime	Stage II runtime
`Walk`	[1]	6	100	walking	outdoor, moving background, ambiguous color	—	320×180	123	24
`Cathedral`	[2]	4	20	runing, falling	outdoor	—	240×140	32	5
`Subject3`	Our	3	100	volleyball	outdoor, cluttered background	shape laser scan	180×90	65	15
`Subject2`	Our	6	100	gymnastics	studio, few colors	shape laser scan	200×160	62	15
`Subject1`	Our	6	100	gymnastics	studio, few colors	shape laser scan	200×160	68	18
`HumanEva-Walk`	[3]	3	586	walking	studio, low-quality image	markers, manual silhouettes	160×120	176	42
`HumanEva-Box`	[3]	3	382	boxing	studio, low-quality image	markers, manual silhouettes	160×120	113	65
`Marker`	[1]	2	100	walking	studio	markers, joint positions	128×128	22	8
`Skirt`	[4]	6	100	dancing	studio, green screen	—	128×128	97	28
`Monocular`	[5]	1	3×1	posing	studio, segmented	multi-view reconstruction	125×125	—	0.3
`Studio`	[6]	10	4 (300 tracking)	gymnastics, walking	studio	—	162×121	14	1.5

Table 1. List of sequences used in our paper and their characteristics. 'Input resolution' refers to the resolution used by our algorithm, not the original video resolution. Runtime is measured in minutes.

2 Implementation details of Stage I

The main document only outlines Stage I, the lifting of 2D skeletal joint detections to a consistent 3D skeleton, as it is not the main contribution of our work, and well-founded solutions already exist [7–9]. Here we give additional details on our implementation.

For detecting joints in images, we use the unary potential output from the ConvNet-based body-part detector by Tompson et al. [10]. It outputs a heat map $\mathcal{D}_{c,t,j}$ for each camera c, frame t and joint j, which contains per-pixel likelihoods to be covered by joint j. This estimation is performed separately for each frame t and camera c. There is no direct, one-to-one correspondence between detection and skeleton joint, as heat maps localize joint positions only roughly and are in general multi-modal due to detection ambiguities and the presence of multiple people. To nevertheless find a good match between 3D skeleton and 2D projection, we introduce $E_{\text{detection}}$, which measures the overlap of the heat maps $\mathcal{D}$ with the projected skeleton joints in terms of Gaussian overlap. Each joint in the model skeleton has an attached colored 3D joint-Gaussian G_j. The detection heat maps and joint-Gaussians (colored blobs) are both shown in the overview Figure 1 (main paper). Each Gaussian is projected into each camera view using the projection model of Rhodin et al. [11]. It defines the visibility $\mathcal{V}_j(u, v)$ of Gaussian G_j as seen from pixel (u, v) in the heat map. The energy term $E_{\text{detection}}$ thus measures the overlap of each Gaussian with the corresponding heat map:

$$E_{\text{detection}}(c, t, \mathbf{p}_t, \mathbf{s}) = -\sum_{(u,v)} \sum_{j} \mathcal{V}_j(u, v, \mathbf{p}_t, \mathbf{s}) \cdot \mathcal{D}_{c,t,j}. \tag{1}$$

As $E_{\text{detection}}$ is non-convex, we employ a hierarchical optimization approach.

In level I, the coarse skeleton position and orientation is determined. For this, we set joint-Gaussians to have large support (standard deviation of $\sigma = 1\,\text{m}$), and only optimize the rigid skeleton pose based on the torso joints for the first frame of the sequence. Level II refines the global pose across the whole sequence (in our experiments around 100 frames) with medium-sized Gaussians ($\sigma = 0.4\,\text{m}$). Level III adjusts bone length by optimizing the shape $\mathbf{s}$. Level IV adds elbow and knee joints with $\sigma = 0.1\,\text{m}$. Level V adds the remaining wrist and ankle joints. We observed that enabling self-occlusion for leg-Gaussians and ignoring occlusion for torso and arm joints gives best results overall.

2.1 Automatic vs. manual actor model

Our reconstructed actor model provides the Gaussian parameters and underlying skeleton dimensions. We tested our model's applicability to the volumetric Gaussian representations proposed by Stoll et al. [6] and Rhodin et al. [11] on the `Marker` and `Walking` sequences, with 3 and 10 cameras, respectively. We found that our automatically generated model matches their manually initialized and hand-crafted body dimensions, see Figure 1.

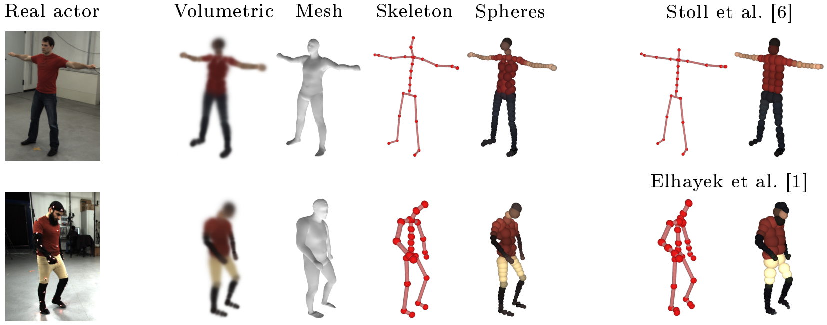

Fig. 1. Comparison of the estimated actor models for the `Walk` sequence (top) and `Marker` sequence (bottom) to the manually created skeletons of Stoll et al. [6] and Elhayek et al. [1]. For comparison between methods, we represent Gaussians as spheres with radius equal to one standard deviation.

We additionally tested the model quality by tracking the same sequence once with the automatically estimated body model and once with the original, manually created models. The overall tracking performance of our model was equivalent to Rhodin et al.'s model, and improved on Stoll et al.'s model. Tracking results are shown in the supplemental video.

2.2 Body shape space generalization

We qualitatively assess the generalization capability of our body shape model by comparing against representing meshes directly as a vector of vertex positions [12], and using per-triangle rotation and shear with respect to a rest shape, similar to SCAPE [13]. For each database mesh instance, we build a combined feature vector by stacking $(\boldsymbol{\gamma}_i, \mathbf{b}_i)$, vertex positions $\mathbf{v_i}$, and per-triangle shear $\mathbf{a_i}$ into a single vector. We perform PCA on the combined features, which generates principal vectors that jointly express the variation in all three representations. To test generalization capability, we explore different PCA coefficients and analyze the mesh predicted by each representation.

Our volumetric skinning is computationally more efficient than per-triangle encodings, SCAPE-like model [13], as per-triangle deformations 'explode' the mesh, and fusing requires solving a linear system, while still yielding comparable shape generalization, see Figure 2. Direct encoding in terms of vertex positions is as efficient as our volumetric skinning – in both cases, vertex positions depend linearly on the PCA coefficients –, however, it exhibits stronger artifacts, see Figure 2. We believe this is due to the coupling of each vertex to neighboring Gaussians in our model, which introduces an implicit spatial smoothing regularization. Each vertex is influenced by multiple Gaussians, and each Gaussian was registered based on all neighboring vertices, which compensates for inaccuracies in the mesh registration.

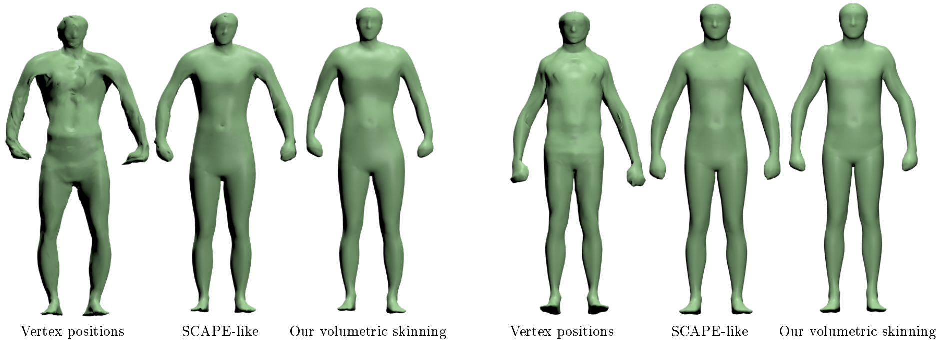

Fig. 2. Comparison of PCA body shape spaces: vertex positions, SCAPE-like per-triangle transformations, and our volumetric skinning. Our volumetric skinning is computationally as efficient as vertex encodings, and yields comparable shape generalization to the SCAPE-like method.

2.3 Model components

We quantitatively assess the contribution of each of our model components by comparing estimated poses to the marker-based ground truth of the `Marker` sequence of Elhayek et al. [1]. We execute all algorithm variants on 3 camera views and 100 frames, the mean Euclidean joint error is plotted in Figure 3. Stage II consistently improves the initial estimates of Stage I. Without the smoothness term E_{smooth}, temporal jitter emerged. Disregarding contour direction and image gradient direction in E_{contour} results in doubling the reconstruction error, which indicates that the integration of contour direction is crucial for the success of the proposed algorithm.

The influence of using only two or three cameras is analyzed on the same sequence, see Figure 4. Automatic reconstruction with three cameras is as accurate as tracking with the handcrafted model and method of Rhodin et al. [11]. Pose reconstruction from only two cameras is still accurate for large parts of the sequence, and sometimes more accurate than tracking with two cameras and a manual actor model. Dramatic errors occur only occasionally in the second half of the sequence. Shape estimation nevertheless succeeds due to the robustness provided by the underlaying parametric model. Please note that our skeleton model has a slightly different structure than the ground-truth skeleton, which likely explains some of the error.

In our experiments E_{smooth} is weighted by 0.01 and E_{flat} by 0.05.

References

1. Elhayek, A., de Aguiar, E., Jain, A., Tompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., Theobalt, C.: Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In: CVPR. (2015) 3810–3818

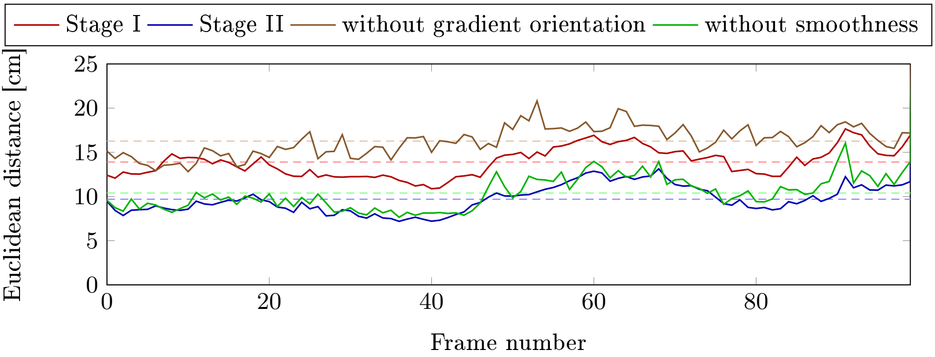

Fig. 3. Model component influence evaluation. All model components are important: stage II consistently improves on stage I; smoothness term E_{smooth} removes temporal jitter; and without contour direction in E_{contour} the reconstruction error doubled.

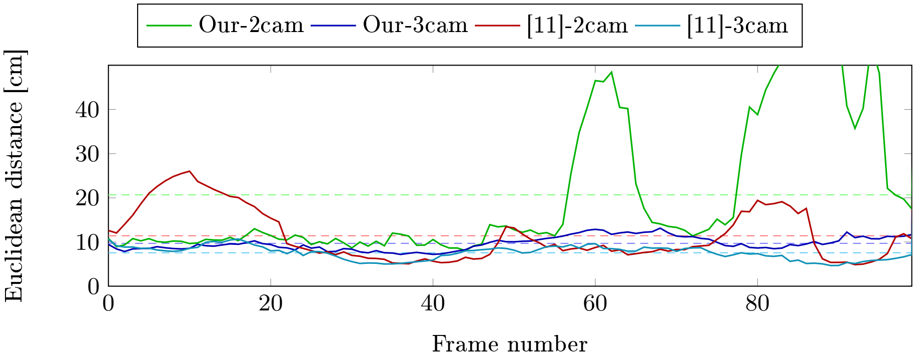

Fig. 4. Influence of the number of camera views. Using three cameras, our reconstruction is as accurate as tracking with the manual model. It is still accurate for two cameras for large parts of the sequence.

2. Kim, H., Hilton, A.: Influence of colour and feature geometry on multi-modal 3D point clouds data registration. In: 3DV. (2014) 202–209
3. Sigal, L., Bălan, A.O., Black, M.J.: HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision (87) (2010) 4–27
4. Gall, J., Stoll, C., de Aguiar, E., Theobalt, C., Rosenhahn, B., Seidel, H.P.: Motion capture using joint skeleton tracking and surface estimation. In: CVPR. (2009) 1746–1753
5. Guan, P., Weiss, A., Bălan, A.O., Black, M.J.: Estimating human shape and pose from a single image. In: ICCV. (2009) 1381–1388
6. Stoll, C., Hasler, N., Gall, J., Seidel, H.P., Theobalt, C.: Fast articulated motion tracking using a sums of Gaussians body model. In: ICCV. (2011) 951–958
7. Sigal, L., Isard, M., Haussecker, H., Black, M.J.: Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. International Journal of Computer Vision **98**(1) (2012) 15–48

8. Amin, S., Andriluka, M., Rohrbach, M., Schiele, B.: Multi-view pictorial structures for 3D human pose estimation. In: BMVC. (2013)
9. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3D pictorial structures for multiple human pose estimation. In: CVPR. (2014) 1669–1676
10. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: NIPS. (2014) 1799–1807
11. Rhodin, H., Robertini, N., Richardt, C., Seidel, H.P., Theobalt, C.: A versatile scene model with differentiable visibility applied to generative pose estimation. In: ICCV. (2015)
12. Allen, B., Curless, B., Popović, Z.: The space of human body shapes: reconstruction and parameterization from range scans. ACM Transactions on Graphics **22**(3) (2003) 587–594
13. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: shape completion and animation of people. ACM Transactions on Graphics **24**(3) (2005) 408–416