Supplemental Material: Detailed, accurate, human shape estimation from clothed 3D scan sequences

Chao Zhang[1,2], Sergi Pujades[1], Michael Black[1], and Gerard Pons-Moll[1]

[1]MPI for Intelligent Systems, Tübingen, Germany
[2]Dept. of Computer Science, The University of York, UK

1. Implementation details

We include in this section more implementation details of the presented method.

1.1. Body model

The SMPL model [1] is gender specific. Both the female and male models consist of $N = 6890$ vertices. To model shape variations, we use $N_{\text{shape}} = 10$ shape coefficients. The pose consists of $N_{\text{pose}} = 3 \times 24 + 3 = 75$ parameters.

1.2. Optimisation parameters

We find that results are not very sensitive to optimization parameter weights. We set the weights so that the different terms in the objective are balanced; roughly in the same order of magnitude. We define the single-frame objective function as:

$$E(\mathbf{T}_{\text{Est}}, M(\boldsymbol{\beta}, 0), \boldsymbol{\theta}; \mathcal{S}) = \lambda_{\text{skin}} E_{\text{skin}} + E_{\text{cloth}} + \lambda_{\text{cpl}} E_{\text{cpl}} + \lambda_{\text{prior}} E_{\text{prior}},$$

On BUFF dataset, we use $\lambda_{\text{skin}} = 100$, $\lambda_{\text{outside}} = 100$, $\lambda_{\text{fit}} = 3$, $\lambda_{\text{cpl}} = 1$, and $\lambda_{\text{prior}} = 0.1$. At the stage of estimating per-frame shape $\mathbf{T}^k_{\text{Est}}$, we increase λ_{skin} by a factor 10 to retrieve more personalized details.

On INRIA dataset, since texture information is not available, we consider all vertices as cloth and therefore set $\lambda_{\text{skin}} = 0$. We decreased $\lambda_{\text{fit}} = 1$ to be more robust to wide clothing, and keep other weights unchanged. To make a fair comparison in this dataset, we initialize the pose using the exact same Stitched Puppet [4] landmarks computed in [3].

To create the fusion scan from the single-frame objective results, we only retain one frame every 1/3 of second. We found this sampling to be a good trade-off between accuracy and computational time.

1.3. Computation time

As all sequences can be computed in parallel, we report the computation time for one sequence. The first step is to solve for the single-frame objective with all vertices labeled as skin (Section 4.1). The computation time of this step is ∼10 seconds per frame. Then the fusion scan is created by a simple rearrangement of data. The fusion mesh computation (Section 4.2) takes ∼200 seconds. The final detail refinement computation (Section 4.3) takes ∼40 seconds per frame. All computations are executed on an 3GHz 8-Core Intel Xeon E5. The current implementation is not efficient and the computational time could be drastically reduced. In the future, we plan to parallelize the computation of the Jacobians on the GPU, which represent most of the computing time in the process.

2. Segmentation

For completeness, we describe the automatic method we used for segmenting the scan into skin and cloth vertices (Section 4 of the paper). The segmentation method is the one in [2]. Segmentation is not part of our contribution and other more sophisticated learning based methods could be used as well. Furthermore, the method presented in Section 4 is robust to inaccurate cloth-skin segmentations thanks to the smooth geodesic distance field.
By minimizing the single frame objective using all labels as skin (Section 4.2) we obtain frame-wise cloth templates $\mathbf{T}^k_{\text{cloth}}$ that align with the scans $\mathcal{S}^k$. Instead of solving the segmentation on the image domain, we solve it directly on the mesh generated by $\mathbf{T}^k_{\text{cloth}}$ to exploit useful 3D shape information. The idea is to solve a Markov Random Field (MRF) with a graph connectivity given by the template mesh.

More formally, let us introduce one random variable v_i, for every node in the template mesh $i \in \mathcal{T}$. These random variables can take discrete values $\{0 \ldots N_{\text{gar}}\}$, where 0 corresponds to the skin and N_{gar} are the number of garments the person is wearing. For the purposes of this paper, we are only interested in separating skin from cloth. Then we solve for the collection of random variables $\mathbf{v} = \{v_i \mid i \in \mathcal{T}\}$ that minimize the following cost function

$$E(\mathbf{v}) = \sum_{i \in \mathcal{T}} \varphi_i(v_i) + \sum_{(i,j) \in \mathcal{T}} \psi_{ij}(v_i, v_j) \tag{1}$$

where φ_i is a node-dependent unary term and ψ_{ij} is a binary term.

Unary term: The unary term encodes the negative log likelihood of node i taking label v_i

$$\varphi_i(v_i) = \sum_{j \in \mathcal{B}_i(\mathcal{S})} -\log(p_j(v_i)) + \epsilon_i(v_i) \tag{2}$$

where the first term is the *data likelihood term* and the second term ϵ is a *semantic body prior* over body parts.
Data likelihood term: We fit a Gaussian mixture model (GMM) to the appearance of each of the garments. In order to be more robust to illumination changes we fit the GMM in HSV space instead of RGB. For every scan point $\mathbf{x}_j$ in the neighborhood $\mathcal{B}_i(\mathcal{S})$ of node i, we evaluate the likelihood under the fitted GMM:

$$p_j(s) = \sum_{m=0}^{N} \pi_s^m \mathcal{N}(I(\mathbf{x}_j) | \mu_s^m, \Sigma_s^m) \tag{3}$$

where $I(\mathbf{x}_j)$ is the HSV appearance of scan point $\mathbf{x}_j$, and μ_s^m, Σ_s^m are the mean and covariance of mixture mode m of segmentation class s.

To train the GMM, we select one frame, and use the information-theoretic criteria (BIC) to determine the number of modes that represents the data well. Then, the skin model is automatically selected, and the rest of the modes fused into a non-skin mode. The selected frame for training is the first of each sequence. We tested randomly selecting other frames obtaining almost identical results. The training of the GMM for each subject is fully automatic.

Although the appearance model is powerful, it is sensitive to noise, shadows and illumination changes. Consequently, we add a semantic prior term that encodes prior knowledge of plausible garment segmentations.
Semantic body prior: This encodes intuitive information such as: the torso nodes are more likely to be T-shirt, hands and head have to be skin. To that end, we leverage a segmentation into parts of our body model. Formally, we define two kinds of priors,

1. a node i is more likely to be label $s = l$ in which case $\epsilon_i(s) = 1 - \delta(s - l)$
2. a node i can not take a certain label in which case $\epsilon_i(s) = \delta(s - l)$.

In particular, for Case 1, the nodes of the head, hands should be labeled as “skin,” the nodes of the spine or torso should be labeled “shirt,” and the nodes of the thighs should be labeled “ trousers.” For Case 2, we penalize when the upper-body nodes are labeled as “trousers” or when the calf and feet are labeled as “T-shirt.”

This sort of intuitive prior knowledge has proven to be very effective enabling segmentation where it would have been impossible otherwise. The prior term helps to correctly segment hands and feet, for which the scan often contains noise and missing data.

Pairwise term: This is a smoothness term, encouraging neighboring pixels to have the same label. This term is as simple as it can be; given the adjacency matrix $\mathbf{W}$ of size $|\mathcal{T}| \times |\mathcal{T}|$ of our template mesh, the term evaluates to

$$\psi_{ij}(v_i, v_j) = \mathbf{W}_{ij} \left(1 - \delta(v_i - v_j)\right) \tag{4}$$

taking cost 1 if nodes i and j are neighbors and take different labels, and cost 0 otherwise. The resulting segmentation is a per node label s_i indicating to which layer a node belongs.

The MRF is solved using the alpha expansion method.

3. Additional results

3.1. Pose estimation with manual markers

We report here for the sake of completeness the alternative method we used to report pose estimation error at the time of submission. In [3], landmarks correspondence to MoCap were manually defined. To obtain a correspondence we fit the SMPL model to S-SCAPE model to transfer the landmark locations. In Fig.1 we report the error. The results in Fig.1 do not reflect the quality of pose estimation due to two reasons: 1) the manual definition of landmarks produces a systematic error and 2) fitting SMPL to S-SCAPE produces another source of systematic error. Therefore, the curves are dominated by the error in computing the correspondence between body vertices and marker locations. Consequently, in the paper in Fig.10 we automatically computed the correspondences in 10 different frames for each sequence and averaged the vertex to marker error for all frames and correspondence sets. This heavily reduces the effect of wrong correspondence estimation for both methods and allows to evaluate pose estimation.

3.2. Robustness to segmentation

In order to evaluate the robustness of the method when skin/cloth segmentation is not available we evaluate our method labeling the scans of BUFF as *all cloth*. The obtained errors are presented in Tab. 1. While the obtained shapes are less detailed (specially in the face), they are still accurate.

cloth style	t-shirt, long pants						soccer outfit					Avrg.
tilt twist left	00005	00096	00032	00057	03223	00114	00005	00032	00057	03223	00114	Avrg.
all cloth	3.03	3.05	2.69	2.86	3.21	2.74	2.80	2.66	2.90	3.28	2.64	2.90
detailed	**2.52**	**2.83**	**2.36**	**2.44**	**2.27**	**2.31**	**2.44**	**2.59**	**2.28**	**2.17**	**2.23**	**2.40**
hips	00005	00096	00032	00057	03223	00114	00005	00032	00057	03223	00114	Avrg.
all cloth	3.20	3.10	2.91	2.96	3.41	2.99	2.94	2.75	2.98	3.42	2.85	3.05
detailed	**2.75**	**2.64**	**2.63**	**2.55**	**2.40**	**2.56**	**2.58**	**2.59**	**2.50**	**2.38**	**2.51**	**2.55**
shoulders mill	00005	00096	00032	00057	03223	00114	00005	00032	00057	03223	00114	Avrg.
all cloth	2.76	3.22	3.08	3.25	3.41	2.86	2.78	2.92	2.91	3.26	2.72	3.01
detailed	**2.49**	**2.85**	**2.72**	**2.37**	**2.26**	**2.59**	**2.83**	**2.82**	**2.28**	**2.33**	**2.51**	**2.55**

Table 1. Comparison of the numerical results obtained disregarding the skin/cloth information and using it. When no skin/cloth information is available, our method still obtains numerically accurate results.

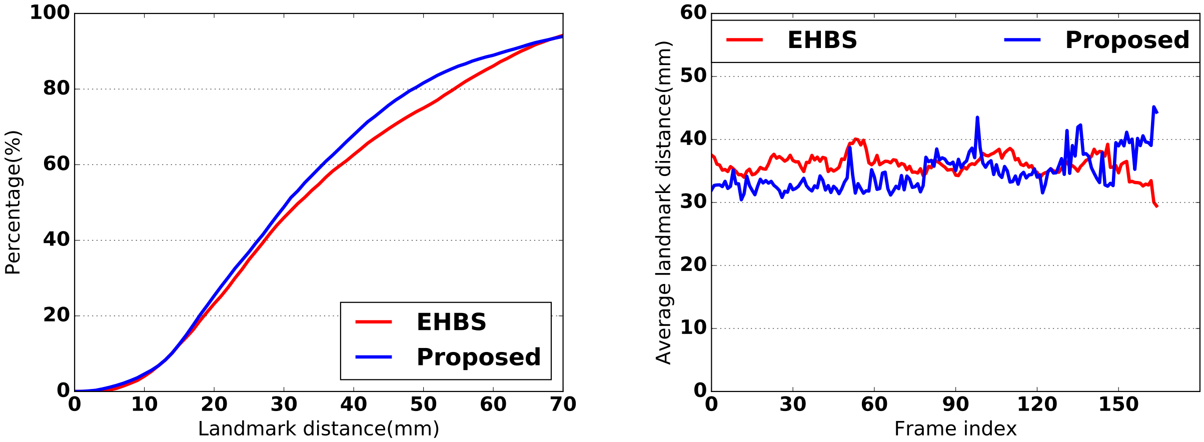

Figure 1. Comparison of pose estimation accuracy on INRIA dataset. Left: Percentage of landmarks with error less than a given distance (horizontal axis) in mm. Right: per frame average landmark error. EHBS is [3]. This figure is shown for completeness and **does not reflect pose estimation accuracy** but rather a systematic correspondence error between markers and body. We address this issue to produce Fig. 10 in the paper, see text.

3.3. Comparison to fitting SMPL to visible skin parts

We show here that our approach is more accurate than simply fitting SMPL to the visible skin parts. The results can be seen in Fig. 2. Simply fitting to visible skin ignores the cloth constraints and therefore the shape has large errors at occluded areas.

3.4. More results

We show a comparison of our method to Yang *et al.* [3] on different challenging wide and layered clothing sequences on INRIA dataset in Fig.3. Our estimated naked body shapes are more realistic. By overlaying the results and the original scans, we observe that our estimations fit better inside clothing. One can also observe that our method is robust to layered and wide clothing movement. Furthermore, to illustrate tracking robustness, in Fig.4 we compare the results of our method with [3] on several frames of one layered clothing sequence. Given a few equally sampled frames of the sequence, we are able to demonstrate the improvement of our method when dealing with 3D clothed sequences. Notice that our method can accurately capture head orientations which are not recovered by [3]. For more results and comparisons please refer to the video in http://buff.is.tue.mpg.de/.

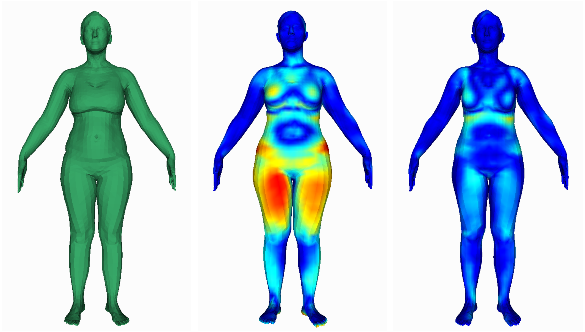

Figure 2. Result for a sequence with t-shirt and long pants. (a) Ground truth minimally clothed shape, (b) heatmap of SMPL fit to visible skin, (c) heatmap of our approach.

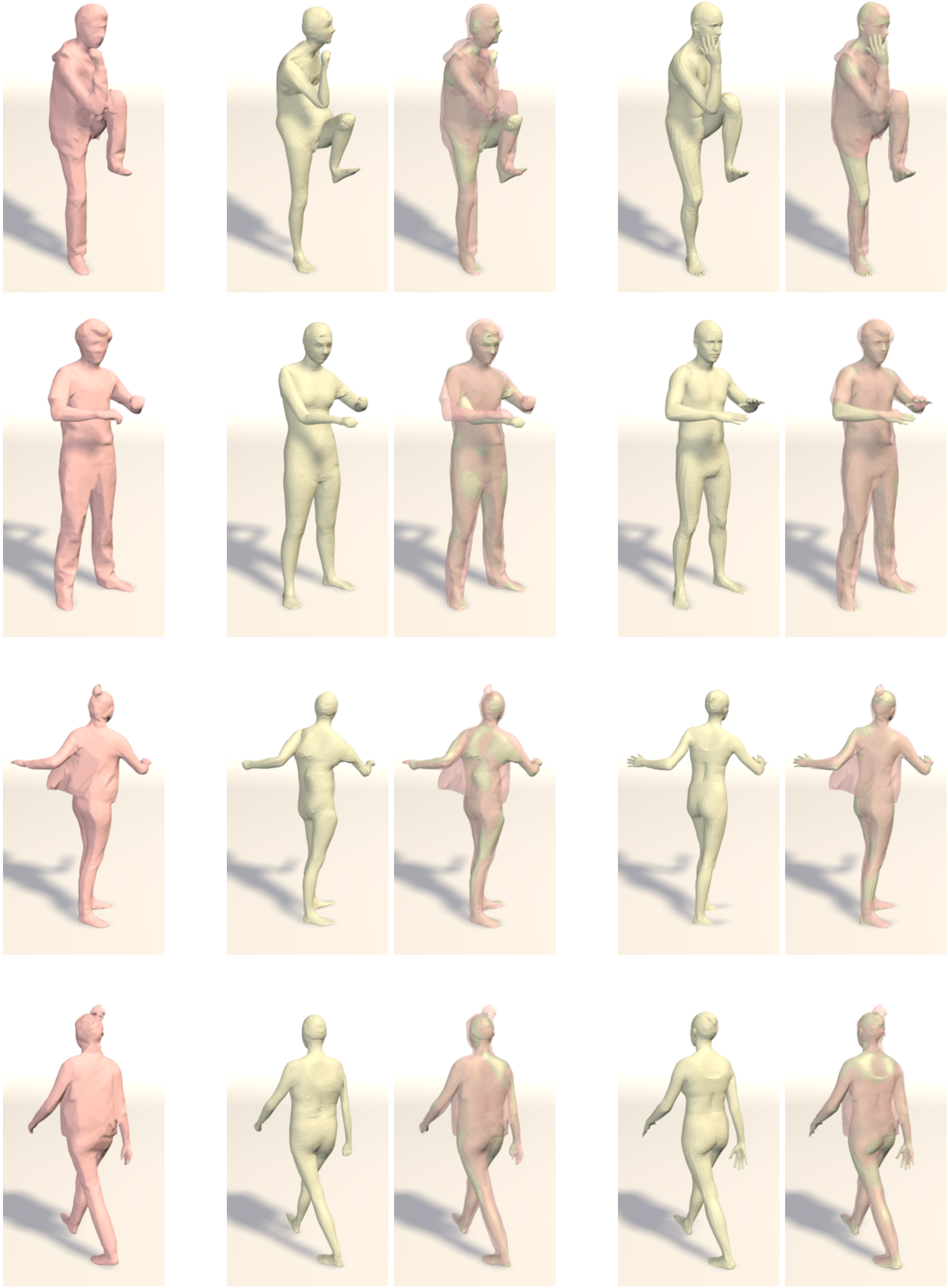

Figure 3. Comparison of our results and Yang *et al*. [3] on representative Inria wide clothing sequences. From left to right: scan, result of [3], our result. Results are shown in pairs: on the left the estimated shape, on the right the estimated shape with the scan overlayed on top. From top to bottom: sequences are s1_layered_knee, s3_wide_spin, s6_layered_spin, and s6_layered_walk.

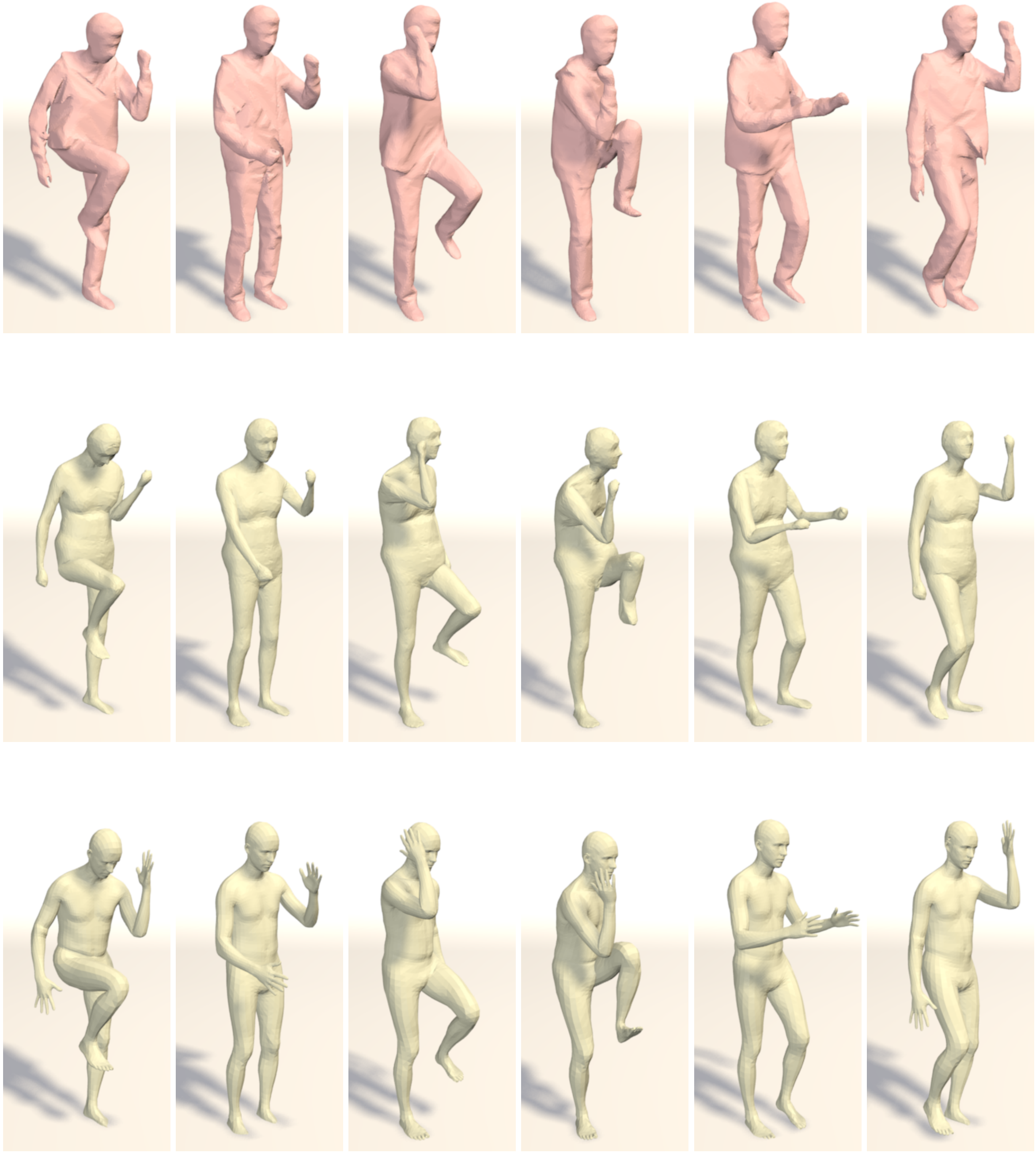

Figure 4. Comparison of our results and Yang *et al*. [3] on s1_layered_knee sequence of Inria dataset. From left to right: frame index = [0, 10, 20, 30, 40, 50]. From top to bottom: scans, [3] , our results.

References

[1] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1

[2] G. Pons-Moll, S. Pujades, S. Hu, and M. Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH) [to appear]*, 2017. 1

[3] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Estimation of Human Body Shape in Motion with Wide Clothing. In *European Conference on Computer Vision 2016*, Amster-

dam, Netherlands, Oct. 2016. 1, 2, 3, 4, 5

[4] S. Zuffi and M. J. Black. The stitched puppet: A graphical model of 3d human shape and pose. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3537–3546. IEEE, 2015. 1