Human Shape from Silhouettes using Generative HKS Descriptors and Cross-Modal Neural Networks

Endri Dibra[1], Himanshu Jain[1], Cengiz Öztireli[1], Remo Ziegler[2], Markus Gross[1]
[1]Department of Computer Science, ETH Zürich, [2]Vizrt
{edibra,cengizo,grossm}@inf.ethz.ch, jainh@student.ethz.ch, rziegler@vizrt.com

1. Supplementary

In this supplementary material, we first give some further insights on the parametric space and the components contributing to the final error. Then, we demonstrate further experiments on noise added to the silhouette, poses, comparisons to CCA applied to the network features and more applications of the $HKS-Net$ architecture (1) from the paper. To complete, we show failure cases, estimated mesh overlays over the input images for our method, and we also visually compare to the method from Bogo et al. [1].

1.1. Parametric space and final error components

We chose to use 20 PCA components to generate the shape space for fairness to methods we compare to [2, 3], that utilize the same number of components, and also because it was enough to capture 95% of the energy and avoid low-variance datasets. Starting from the original meshes and others spanning such a space, we learn a 4000 dimensional internal representation space, extracted from the HKS features and used to decode mesh vertices directly. Despite the fact that the embedding space is of higher dimensionality than the 20 parameters used in the previous works, we believe that it's higher accuracy stems from compact pose-invariant features, needed here to learn non-linear mappings to higher dimensional mesh vertex spaces. This is a better learned representation than just pure PCA applied on the triangle deformations. Furthermore, our method is also faster than the other methods for the same input and output resolution. This is one factor that contributes to the final error estimation, and also demonstrated e.g. in Tab.2 from the paper, when we compare $SF-1$ to $HS-Net-1S$. Another factor that plays an important role is the mapping from silhouette images to the embedding space. We demonstrate the decrease in error as the number of view/modes is increased during training, but remains uni-modal at testing, utilizing our novel CMNN network (e.g. Tab.2 from the paper, $SF-1$ vs $SFS-1$). If we consider the influence of the input image resolution, we believe that it does not play a role in comparison to the previous works, as we used the same input image size as in [3], which is half of the resolution used in [2]. Last but not least, the combination of the above two factors through joint training also helps decreasing the errors, as we show in Tab.1 $SFUS-HKS-1$ from the paper.

1.2. Noise

An important evaluation factor for real world systems is robustness to noise. Although for our target applications this is less of a concern, in general this is important. Hence, we generate noisy silhouettes by non-uniformly eroding or dilating the silhouette at the border, with filters of various radii (we consider 1,3,5,7 and 9 pixels). An illustration of such noise applied to the same silhouette for the various radii is depicted in Fig.1. The mean error obtained over all the body measurements when noise is applied to every input test silhouette for the $SF-1$ network, is shown in Fig.2 (top line), computed as the difference from the clean silhouette errors. As it can be observed, the increase in error for reasonable noise radius is small, and even for highest noise radius, the maximum error is below 2 cm.

Missing Limb. In addition, we perform a further experiment, where silhouette noise is understood as a missing limb part, which could represent difficulties in silhouette

Measurements	SFS-1	SF-1-CCA	SFUS-1	SFUS-1-SH
A. Head circumference	3.7±2.9	4.3±3.5	3.9±2.9	4.2±3.4
B. Neck circumference	2.3±1.8	2.8±2.1	2.2±1.7	2.2±1.9
C. Shoulder-blade/crotch length	5.3±4.2	7.2±5.5	5.4±4.1	5.8±4.5
D. Chest circumference	5.9±4.9	7.8±6.9	5.8±4.8	6.6±5.5
E. Waist circumference	7.5±5.9	9.2±7.2	7.5±5.7	8.5±6.6
F. Pelvis circumference	8.4±6.7	9.7±8.1	8.1±6.5	8.6±7.1
G. Wrist circumference	1.9±1.6	3.1±2.1	1.9±1.6	2.1±1.7
H. Bicep circumference	2.9±2.4	4.2±3.4	2.9±2.5	3.3±2.6
I. Forearm circumference	3.1±2.3	3.3±2.6	2.9±2.3	3.2±2.5
J. Arm length	3.3±2.5	4.5±3.6	3.2±2.5	3.5±2.9
K. Inside leg length	6.2±4.8	7.4±6.0	5.7±4.5	6.2±5.1
L. Thigh circumference	5.8±4.9	7.1±5.9	5.8±4.8	6.2±5.3
M. Calf circumference	3.3±2.7	4.3±3.6	3.5±2.8	3.9±3.3
N. Ankle circumference	1.9±1.5	2.1±1.6	2.1±1.5	2.3±1.7
O. Overall height	11.2±8.6	14.1±11.1	10.4±8.1	11.9±9.5
P. Shoulder breadth	2.2±1.2	2.6±2.1	2.1±1.7	2.2±1.9

Table 1. Body measurement errors comparison over various experiments considered in the supplementary and results from the paper. Errors are expressed as Mean±Std. Dev in millimeters.

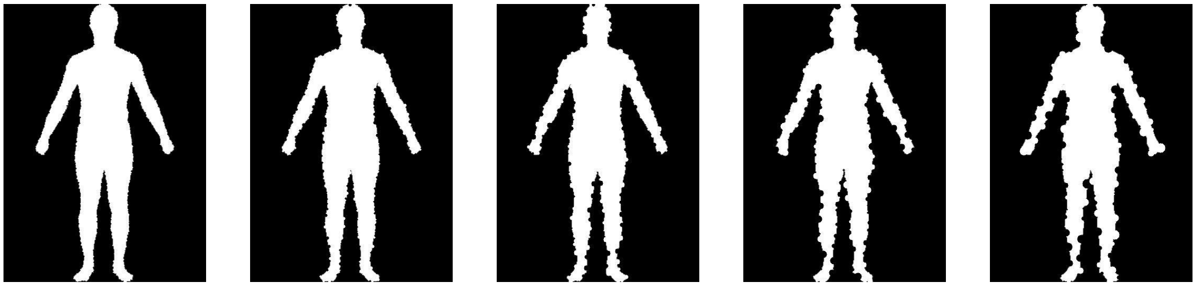

Figure 1. 5 silhouettes representing the same person with noise applied to them. Noise parameters (radii) considered 1,3,5,7 and 9 pixels.

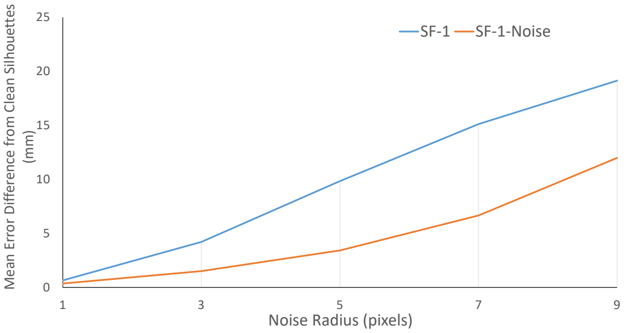

Figure 2. Error plots for the increase in the mean errors as compared to the silhouettes without noise. The top line ($SF-1$) demonstrates the errors when training is performed on clean silhouettes and testing on noisy ones. On the other hand, the bottom line ($SF-1-Noise$) demonstrates the errors when noise is inflicted into the training data. The mean errors are computed over all body measurements. The noise parameter (radii) varies from 1 to 9 pixels.

extraction over various body parts, due to motion blur, occlusion or similar foreground-background color (e.g. when a person stays in front of a wall or uniform background colour, quite often the feet project onto the floor/pavement which could be of the same color as the shoes, e.g. Fig.3 from the paper). For this, we evaluate the $SF-1$ on test data of the form depicted in Fig.3, where a limb part is missing. We observe a little increase of 3.77 mm in the overall mean error, as compared to $SF-1$ evaluated on the complete silhouettes. This could be due to the human body prior and its symmetric properties.

Train with Noise. Lastly, we perform an experiment, where instead of only testing with noisy silhouettes, we also train with noisy ones. For this experiment the amount of training data grows linearly with the amount of noise radii we consider. Once again the network trained is similar to $SF-1$, and we call it $SF-1-Noise$. Evaluating on the same test silhouettes as the first noise experiment, we observe a decrease in the mean error, as shown in Fig.2 (bottom line), which shows that adding perturbations and noise to the training data makes the method more robust to it.

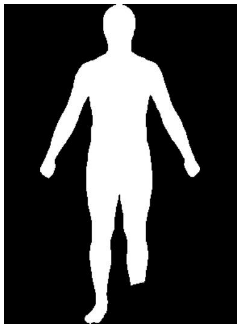

Figure 3. Visualization of the input silhouettes when a limb part is missing.

1.3. CCA

In the paper, we demonstrated that cross-correlating features during training time at later stages, by sharing weights in fully connected layers, improved predictions for our test samples. One main advantage of cross-view learning through neural networks is that the training data does not need to be stored in memory and especially, one can add as many views as desired, as compared to CCA [4] that has been practically shown for two views only. Nevertheless, for fairness also to the method from [3], we compare to a version of our network that utilizes CCA for correlation, and only considers two views (the front and side silhouette scaled and in a neutral pose). The training goes as follows : 1. We first train two networks separately, one for the front $SF-1$ and one for the side $SS-1$ to map view specific silhouettes to the embedding space. This is utilized to learn view specific features directly from the network (as opposed to [3] that extract handcrafted features); 2. Then, we extract 8064 features from the last convolutional layer over each view, for all of our training data, and since the dimensionality is quite high, similar to [3], we apply dimensionality reduction through PCA up to 500 dimensions that capture most of the energy. Starting from these 500 dimensional vectors we apply CCA, to find linear projection bases where the correlation of the projected features is maximized; 3. In the end, we train a smaller network of three fully connected layers $SF-1-CCA$, to map from the 500 CCA projected

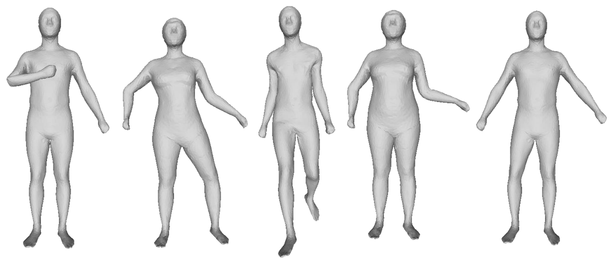

Figure 6. Illustration of people in various poses considered throughout our experiments.

features of the frontal view only (the desired one) to the embedding space. At test time, a new frontal view silhouette is first input to $SF-1$ that performs a forward pass to extract 8064 features, which are then projected onto PCA and CCA. The projection is mapped through $SF-1-CCA$ to the embedding space, which in turn reconstructs the mesh with the help of the $HKS-Net$. We demonstrate the results of this procedure for the same synthetic meshes and we compare to our cross-modal training over two views $SFS-1$ in Tab.1. It can be noticed that our method outperforms the CCA based one. The latter still performs well, however on the expense of added memory footprint and unscalability to more than two views. Furthermore, it is not trivial to train the network end-to-end without splitting it into various components. And lastly, we think that most of the learning is due to the non-linear mapping performed from $SF-1-CCA$, rather than from the linear CCA mapping.

1.4. Late Sharing

We perform a further experiment, to demonstrate the need of sharing weights at later stages in the network for the cross-modal training, as opposed to sharing at earlier stages. The motivation behind late sharing was that we first wanted to let the network separately figure out the appropriate filters to apply to the various views, and then combine higher level and more meaningful features through shared fully connected layers. To demonstrate this, we train a network considering three views, similar to $SFUS-1$, however here the weight sharing starts from the first convolutional layers, all the way to the end, which we call $SFUS-1-SH$. The evaluations of this network for the same synthetic data, with frontal scaled silhouettes as input, are depicted in Tab.1. It can be seen that the results are worse than $SFUS-1$, demonstrating the need for late sharing.

1.5. Convolutional Filters

For illustrative purposes, we also demonstrate the filter responses of one of the last convolutional layers for $SF-1-P$ when the input silhouette is of a person in three various poses (Fig.4). A more detailed version of the single view architecture is depicted in Fig.5. The network internally learns to distinguish between various body parts (e.g. limbs), as similar looking filters are applied to the same parts (e.g. hands), even though the poses vary. An illustrative figure of the various poses we consider in the paper and here is depicted in Fig.6.

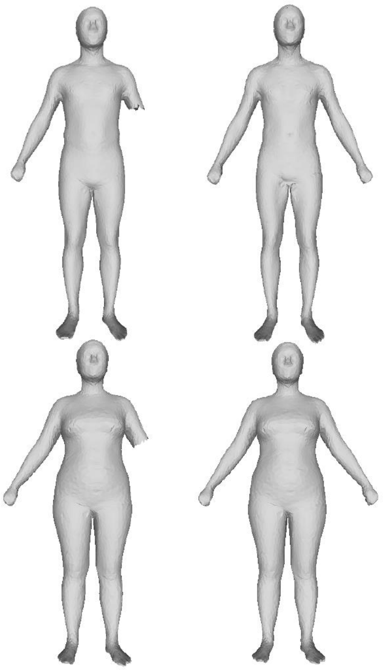

Figure 7. Mesh reconstruction (right) when a partial mesh (left) is input into the $HKS-Net$.

1.6. Experiments on HKS-Net

Despite our intention in this paper to demonstrate accurate estimation of human body shapes from silhouettes, here we present some further experiments that show some of the nice properties of the $HKS-Net$. We demonstrate visual results of the reconstructions as well as mean errors computed over all the body measurements. For each input mesh we first compute the HKS descriptor. That is then fed into the $HKS-Net$ to reconstruct the final mesh. For the quantitative results, we compute the difference of errors for each measurement, obtained for each of the experiments that we consider (which modify the original mesh), from the errors obtained when the original meshes in neutral pose are input to the $HKS-Net$.

Partial Mesh. First, we assume the mesh in a neutral pose comes with missing parts (limbs etc.). We remove the left hand over all the test meshes. The qualitative reconstructions are depicted in Fig.7. The mean overall added error is 6.72 mm. We can observe that the network has reconstructive abilities despite missing extremities.

Posed Mesh. Secondly, we test over meshes coming in

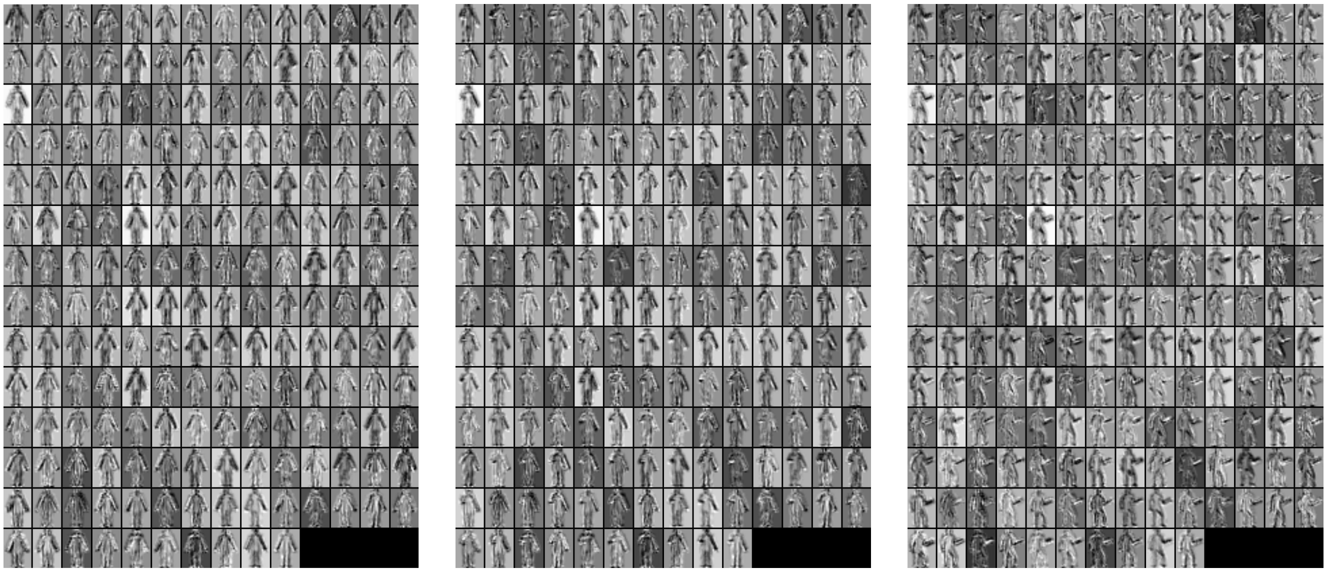

Figure 4. Visualization of filter responses on the last convolutional layers of $SF-1-P$. The same person in three various poses is shown.

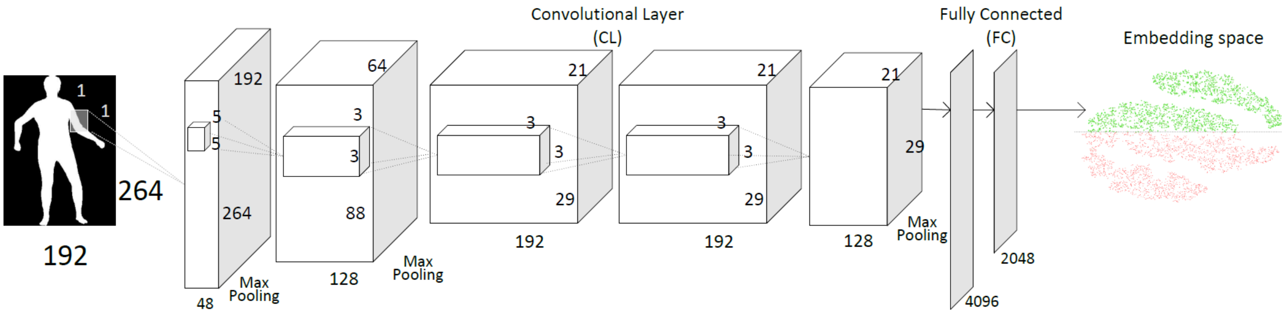

Figure 5. Network architecture for a single view case trained with $SF-1-P$ architecture from the paper. For other types of inputs, such as side view etc., the architecture is the same.

poses obtained from Linear Blend Skinning (LBS) and different from the neutral one. This experiment is important for applications where the computation of a neutral pose of a given posed-mesh is needed. This would allow for mesh alignment, matching as well as consistent measurement computations. Some qualitative reconstructions are depicted in Fig.8. The mean overall added error is 3.72 mm. This almost implies invariance to isometric deformations, however due to LBS artefacts the errors increase a bit as opposed to the neutral pose reconstruction.

Noisy Mesh. Lastly, we evaluate robustness to mesh noise for the $HKS-Net$. For this we apply random vertex displacements to the original ground truth meshes. The qualitative reconstructions are depicted in Fig.9. The mean overall added error is almost negligible, 0.2 mm, which implies robustness to mesh noise.

1.7. Failure Cases

One typical example of a failure case is that of a single view ambiguity, e.g. Fig.10 (bottom), where we show a synthetic mesh of a man with pot belly that is not captured from the frontal silhouette, hence the reconstruction (on the right) tries to best explain it. Other examples are bodies that do not reside in the shape space from which we generate the data, e.g. the muscular male in Fig.10 (top).

1.8. Mesh Overlaps

We show the estimated meshes (third column in gray), utilizing our method $SF-1-P$ from the paper, for three input photos from Dibra et al. [3] (same individuals as in Fig.3 from the paper) along with the estimated meshes (last column in pink) from the method of Bogo et al. [1], in Fig.11. We also show the meshes estimated with our method and that from [1] overlayed on the input images, in Fig.12 and Fig.13 respectively, in order to emphasize the differences in estimations from both methods. It can be noticed that our

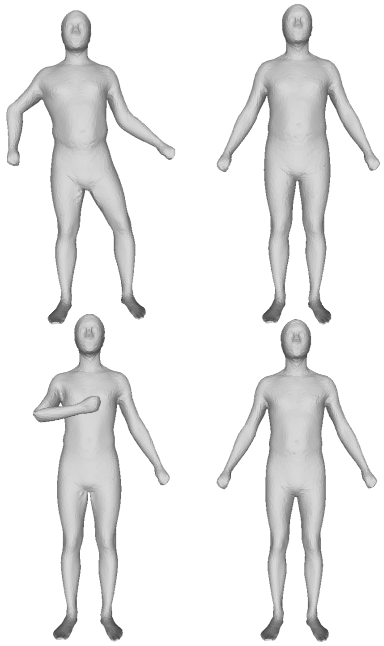

Figure 8. Mesh reconstruction (right) when a posed mesh (left) is input into the $HKS - Net$.

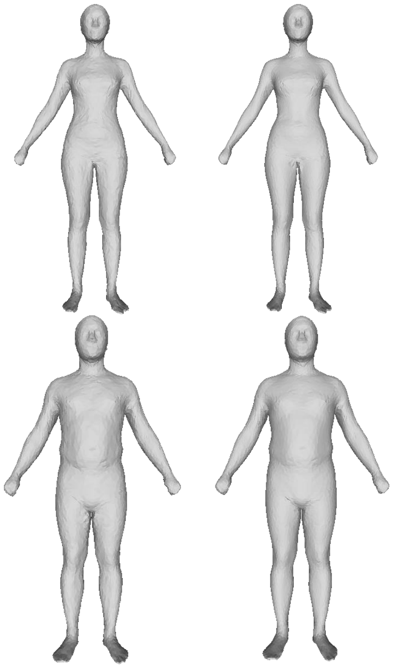

Figure 9. Mesh reconstruction (right) when a noisy mesh (left) is input into the $HKS - Net$.

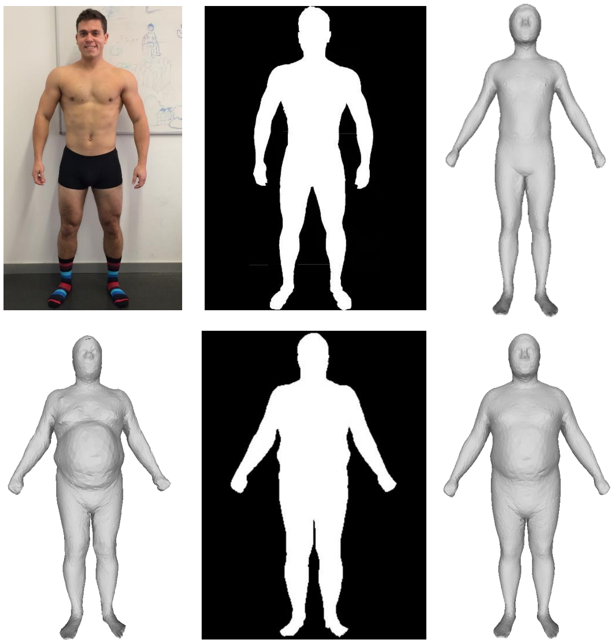

Figure 10. Examples of two failure cases.

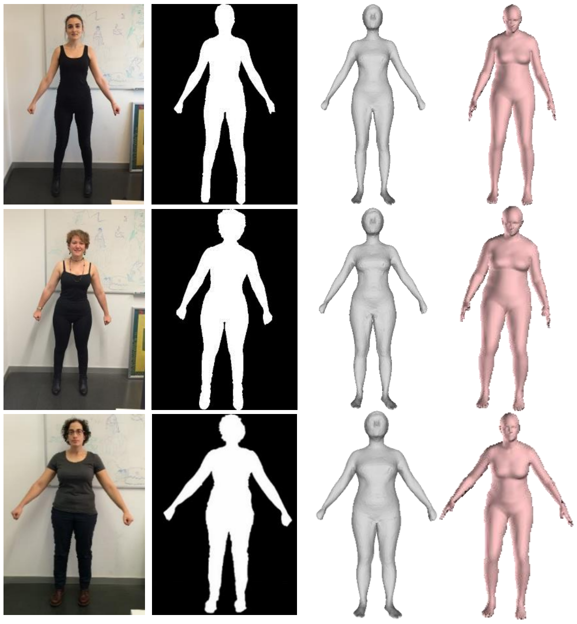

Figure 11. Results for predictions on the test images from Dibra et al. [3]. From left to right: the input image in a rest pose, the corresponding silhouette, the estimated mesh by our method $SF-1-P$, and by the method of Bogo et al. [1].

method gives more accurate estimations for these individuals, with a tendency of the method from [1] to overestimate, also visible by the difference in silhouette projection, especially on the torso and around the waste in Fig.13. Additionally, in Fig.12 we show the overlay on the scanned mesh of another individual from the testing dataset. Please note that we did not apply linear blend skinning to change the neutral pose to fit perfectly the input silhouette, in order to

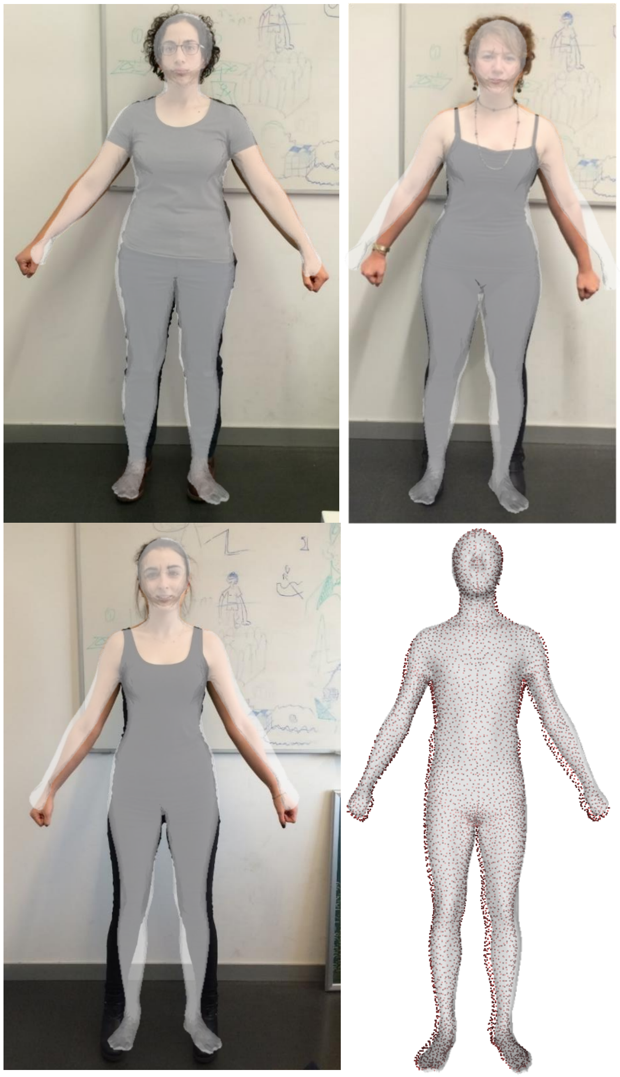

Figure 12. Estimated overlayed meshes utilizing our method overlayed on the input images or scans (bottom-right).

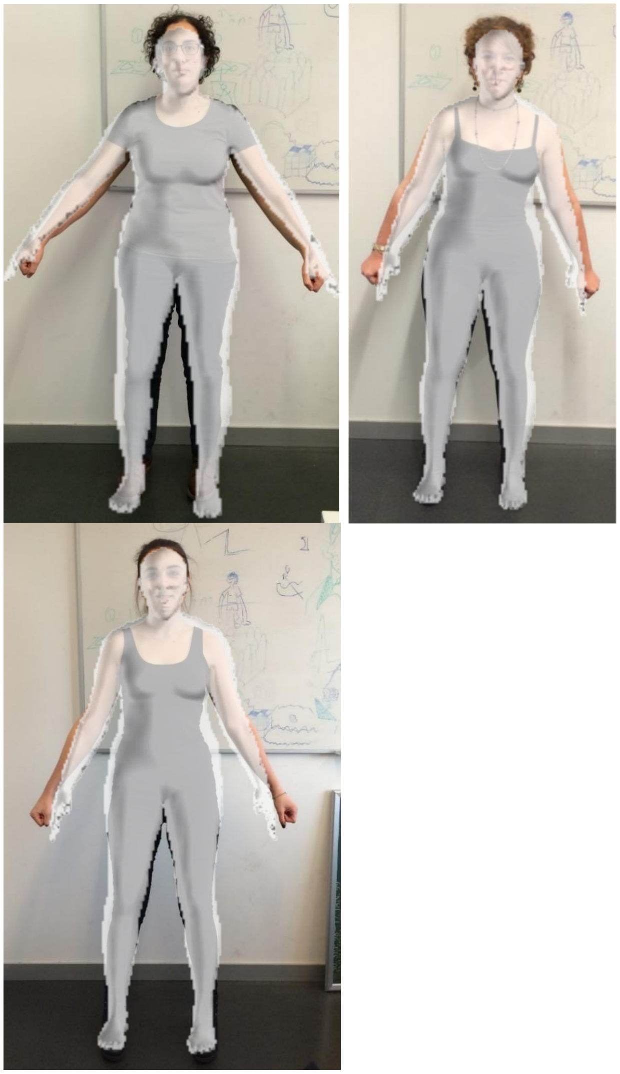

Figure 13. Estimated overlayed meshes utilizing the method from Bogo et al. [1] overlayed on the input images.

enhance the fact that for the application of automatic body measurement a fixed pose is not needed. The method from Bogo et al.[1] on the other hand attempts to more accurately estimate the 3D body pose, which is also the main purpose of their work.

References

[1] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016. 1, 4, 5, 6

[2] E. Dibra, H. Jain, C. Oztireli, R. Ziegler, and M. Gross. Hs-nets : Estimating human body shape from silhouettes with convolutional neural networks. In *Int. Conf. on 3D Vision*, October 2016. 1

[3] E. Dibra, C. Öztireli, R. Ziegler, and M. Gross. Shape from selfies: Human body shape estimation using cca regression forests. In *European Conference on Computer Vision*, pages 88–104. Springer, 2016. 1, 2, 4, 5

[4] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, Dec. 1936. 2