HS-Nets : Estimating Human Body Shape from Silhouettes with Convolutional 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

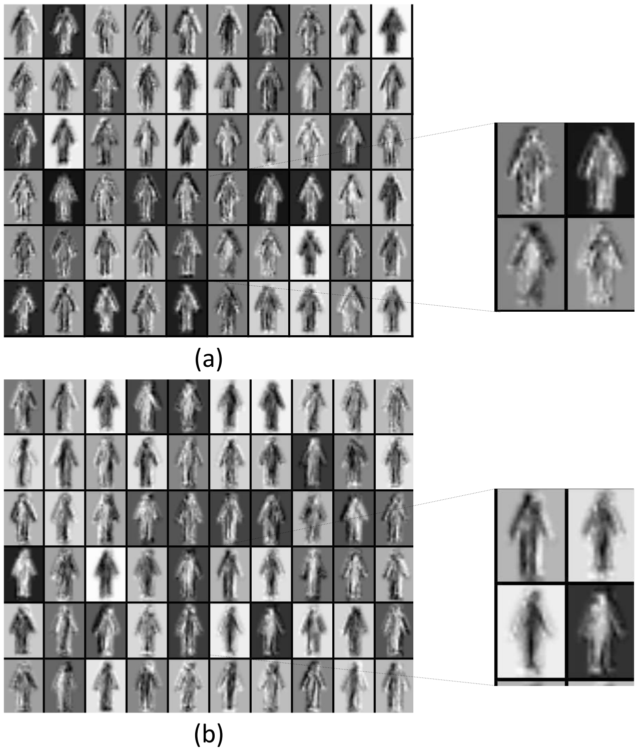

Figure 1: Visualization of randomly chosen 60 convolutional filters on a test input for 3rd layer (*left*) and zoomed in view of four selected filters (*right insets*) for (a) one view case (HS-1-Net) and (b) two views case (HS-2-Net-CH).

1. Supplementary Material

In this supplementary material, we present new quantitative results for three additional set-ups in Table 1, illustration of the convolutional filters for a test image over the one and two view HS-Net (Fig.1), and estimated meshes for four synthetic test examples (Fig.4).

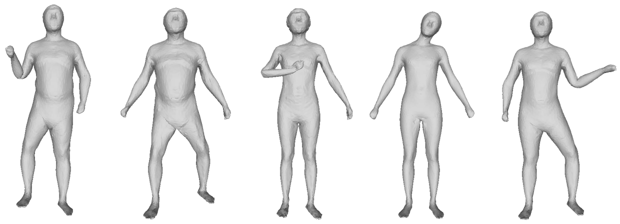

Figure 2: Meshes in various poses

1.1. Quantitative Results

In the paper, we presented quantitative results of the mean error and standard deviation for 16 body measurements performed on the mesh. There, we compared to state-of-the-art methods for full body silhouettes in a neutral pose under varying assumptions. Here, we present some more experiments that show the extensibility of our approach. These include: (1) an experiment with more pronounced poses (2) partially visible silhouettes and (3) images rendered under specularity assumption. All the following experiments were performed assuming unknown camera calibration, hence scaled silhouettes similar to *HS-1-Net-S* from the paper.

Poses. We generated 95000 meshes in 10 different poses, Fig.2 and compare to the results of *HS-1-Net-S* Tab.1, Col. 1. Except for the Arm Length (J) measurement (which has an added error of 40 mm), we observe very similar results. The added error for J can be due to the fact that we include poses similar to the one from Fig.2 (middle). This pose introduces self occlusions, handling of which is a limitation of our current system.

Half Body. We think that an interesting stress case is that of partially visible people, e.g. an upper body selfie, or without loss of generality, a female wearing a skirt that would impede the reliable estimation of the lower body part. Directly applying *HS-1-Net-S* to such inputs resulted in increased errors, implying that with the current training set, it

Measurement	HS-1-Net-S	HS-1-Net-SH	HS-1-Net-SHS-Half	HS-1-Net-SHS-Full	HS-1-Net-Im	HS-1-Net-IP
A. Head circumference	4±4	5±5	5±5	5±5	4±4	4±4
B. Neck circumference	8±5	8±5	8±5	8±5	6±4	6±4
C. Shoulder-blade/crotch length	20±15	20±15	21±16	21±16	20±14	**17±12**
D. Chest circumference	13±7	14±7	15±7	14±6	13±6	13±8
E. Waist circumference	19±13	19±14	20±15	20±14	19±13	19±14
F. Pelvis circumference	19±14	20±14	21±15	20±14	19±12	19±14
G. Wrist circumference	5±3	6±3	6±4	6±4	5±3	5±3
H. Bicep circumference	8±4	8±4	9±4	9±4	8±3	8±4
I. Forearm circumference	7±4	7±4	7±4	7±4	6±3	6±4
J. Arm length	12±8	12±8	13±8	12±7	12±8	12±8
K. Inside leg length	20±14	19±13	19±13	19±13	19±13	19±14
L. Thigh circumference	13±8	13±8	13±8	12±8	12±7	12±8
M. Calf circumference	12±7	12±6	12±6	12±6	11±6	11±6
N. Ankle circumference	6±3	6±3	5±3	5±3	5±2	5±3
O. Overall height	50±39	51±38	52±39	52±39	49±37	**47±37**
P. Shoulder breadth	4±4	4±4	4±4	4±4	3±4	4±4

Table 1: Results of the additional experiments with errors represented as Mean±Std. Dev (in mm). All experiments are done with the input scaled to a fixed height. Experiments (from *left* to *right*): full body silhouettes (from the paper); only half body silhouettes; trained on both half and full body silhouettes but tested only on half; same as previous but tested on full body silhouettes; grayscale images with shading under Lambertian assumptions (from the paper); grayscale images with Phong shading (we highlight the most significant decrease in error due to phong shading as compared to the Lambertian case).

is not possible to accurately estimate full bodies from partial silhouettes.

To tackle this, we train a network similar to *HS-1-Net-S*, however with silhouettes from the upper half of the body, as input, (Fig.3, Left). As illustrated in Tab. 1, Col. 1 and 2, the results for this network (*HS-1-Net-SH*) are similar to that of the full body case (*HS-1-Net-S*). We think that the ability of this network to accurately estimate full body shapes from partial (half body) images is due to the high correlation between different body measurements, as represented by the shape parameters β.

Finally, we train a network that can have either full or half body silhouette as possible input (*HS-1-Net-SHS*), and test separately on each input type. We demonstrate the results in Tab.1, Col. 4 and 5, and notice that the performance for both full and half body silhouettes is similar to the network trained separately on each input type (*HS-1-Net-S* and *HS-1-Net-SH*), with a maximum added error of 2 millimeters for individual measurements. This shows that our network can simultaneously learn to accurately estimate 3D body shapes, from both full and partial images, unlocking further applications.

Images under Phong Shading. We perform a final experiment to illustrate the effects of specularity on the grayscale shaded images, e.g. selfies with flash on or people wearing clothes of specular materials. For this, we train a network (*HS-1-Net-IP*) with images extracted using Phong shading (Fig.3, right) and observe a slight improvement over the experiment with shading under Lambertian assumptions (*HS-1-Net-Im*), Tab.1 Col. 5 and 6. We think that this is because of the extra information from specularity in Phong shading. This again shows that the network is able to utilize the added information from shaded images.

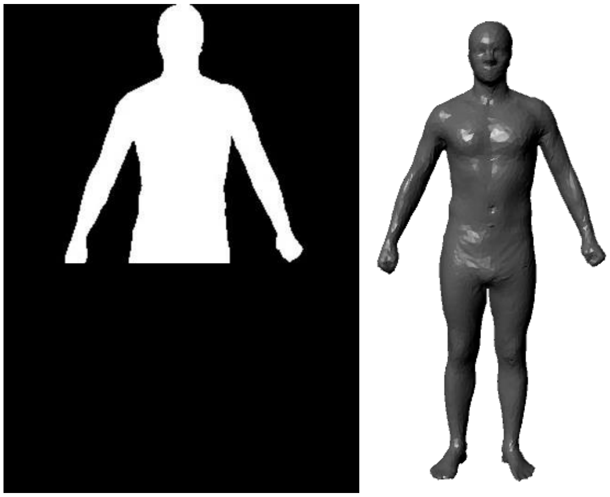

Figure 3: Examples of a half body silhouette (*left*) and a grayscale image with Phong shading (*right*)

1.2. Qualitative Results

Filters. We illustrate some of the convolution filters from the third convolutional layer, visualized on a sample test input (Fig.1). For the zoomed in version of two views treated as two channels of an image (HS-2-Net-CH) (Fig.1(b), *right inset*), we can observe that the side view is

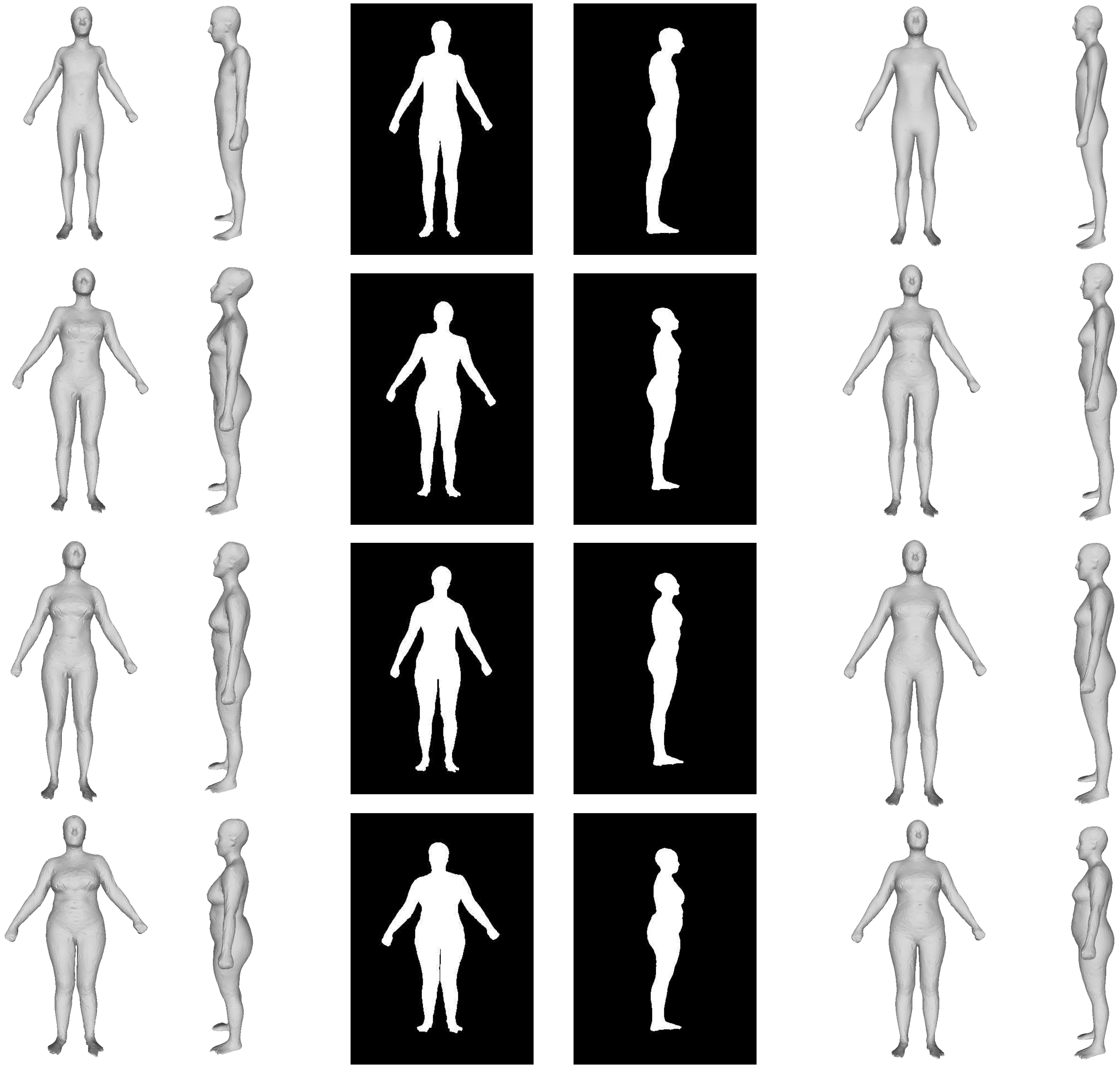

Figure 4: Reconstructed meshes for various test inputs for two views network *HS-2-Net-MM* from the paper. (*left* to *right*) Original meshes (front and side view); Silhouettes (front and side view); Reconstructed Meshes (front and side view)

more pronounced in the *bottom-left* (Fig.1(b), *right inset*) filter while the front view is more pronounced in the *top-left* (Fig.1(b), *right inset*) one, the other two have combined features from both views.

Synthetic Meshes. We present estimated meshes for four synthetic test examples with slight pose changes in Fig.4.

1.3. Summary

We have shown various real scenarios that can be handled with our method, along with its limitations. We think that it can be further extended by combining datasets of various input types, for example, poses and partial images.