2016 Fourth International Conference on 3D Vision

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

Abstract

We represent human body shape estimation from binary silhouettes or shaded images as a regression problem, and describe a novel method to tackle it using CNNs. Utilizing a parametric body model, we train CNNs to learn a global mapping from the input to shape parameters used to reconstruct the shapes of people, in neutral poses, with the application of garment fitting in mind. This results in an accurate, robust and automatic system, orders of magnitude faster than methods we compare to, enabling interactive applications. In addition, we show how to combine silhouettes from two views to improve prediction over a single view. The method is extensively evaluated on thousands of synthetic shapes and real data and compared to state-of-art approaches, clearly outperforming methods based on global fitting and strongly competing with more expensive local fitting based ones.

1. Introduction

Human body shape estimation is an important problem in computer vision, but has so far not received as much attention as the closely related problems such as pose estimation. The methods so far rely on hand-crafted features and specialized algorithms with possible manual interaction. In contrast, it has been shown repeatedly that utilizing neural networks can lead to superior results for many problems such as classification [30], segmentation [34, 18], pose estimation [48] and shape classification or retrieval [47, 49, 17]. However, applying this technique to body shape estimation has not been considered so far. Estimated shapes can in turn be used for applications such as surveillance [9], biometric authentication, image retouching [56], rendering novel viewpoints [52, 7, 46] and also pose estimation, since the integration of body shape knowledge simplifies and improves pose estimation algorithms [54, 14]. A current trend is that of medical and personal measurements, garment and virtual cloth fitting [21, 50, 39, 36], which is also the focus of this paper.

A practical human body shape estimation algorithm should be accurate, robust, efficient and automatic. The existing algorithms do not satisfy these fundamental properties simultaneously. More accurate methods rely on manual input and a fitting pose [56, 28, 39], while others operate under more restrictive assumptions [5], or utilize handcrafted features [31, 44, 33]. As a further common shortcoming, most methods have prohibitive time complexity for practical applications [5, 51, 10].

In this paper, we propose an accurate, fully automatic, and very fast method that avoids handcrafted features and pose fitting by utilizing Convolutional Neural Networks (CNNs) to estimate the 3D body shape of a person, with garment fitting and personal measurements applications in mind. We analyze four possible cases as inputs to the network (a) a single frontal binary silhouette of the person scaled to a fixed size, needed in case of missing camera calibration information (b) the shaded image of the person scaled to a fixed size, with the motivation that shading withholds information complementary to the silhouette (c) a frontal silhouette which assumes known camera parameters and (d) two silhouettes simultaneously (front and side) under known distance from the camera, which in fact is a realistic assumption for the intended use-cases. In compliance with the applications, we make the assumption that people are wearing tight clothes and pose in a neutral stance that allows mild pose changes, Fig. 3 (Top-left). Our method relies on advances made in the field of Neural Networks and a human body shape model [3] obtained from thousands of 3D scans [53, 37]. Utilizing a CNN of roughly the size of AlexNet [30] our method learns a global mapping from the input to the shape parameters. In fact, we learn an end-to-end regression from an input silhouette to 20 parameters that are used to recover the underlying body shape. In addition, we show how to combine body views from two silhouettes to improve prediction over a single view. In order to comprehensively evaluate our method we validate it

978-1-5090-5407-7/16 $31.00 © 2016 IEEE
DOI 10.1109/3DV.2016.19

on thousands of body shapes, by computing error metrics on measurements used in garment fitting, showing robustness to noise and comparing it to state-of-the-art methods that work under the same restrictive assumptions as (d). We clearly outperform the state-of-the-art methods solely based on global mapping [51, 10] for all four input types, and strongly compete in accuracy with a method that additionally uses local iterative fitting [5], while being orders of magnitude faster.

Contributions In summary, this paper has the following contributions : (1) we present a fast and automatic system for human shape and body measurements estimation, from silhouettes or shaded images of people in garment fitting like poses, by learning a global mapping to shape parameters, (2) we present the first system to our knowledge, that can accurately reconstruct human shapes from images utilizing CNNs, (3) we show how to train from scratch an end-to-end fully supervised regression from CNNs with binary silhouette images as input, and demonstrate how to incorporate more evidence (e.g. a second view) in order to improve prediction, (4) we thoroughly validate the method on larger datasets, and demonstrate clear improvements in accuracy and speed over the state-of-the-art.

2. Related Work

Human body shape statistical models It is an ill-posed problem to estimate the 3D geometry of a human body from 2D imagery. Human body shape models regularize the problem by constructing a parametric model, capturing the inherently low degrees of freedom of human body shapes [3, 38, 24]. Hence, the problem of shape estimation boils down to estimating the parameters of the model. The shape models can be augmented with pose changes represented by transformations in an embedded control skeleton [3, 24, 35]. We utilize a popular human body shape model called SCAPE [3], generated on the combination of two state-of-the-art human body databases [53, 37].

Silhouette matching for body shape estimation A common approach leading to accurate 3D human body shape estimations from imagery, is matching an input silhouette to that of the projected 3D shape by correspondence [9, 22, 4, 6, 23, 56, 28]. Despite promising work [41, 42], it is difficult to establish correspondences between silhouettes, especially in the presence of occlusions and challenging poses. Thus, current methods require manual efforts to estimate pose and shape by matching silhouettes [9, 56, 28] and operate under assumptions on the view, calibration, and error metrics utilized [22, 6, 28]. The recent work of Lahner et al. [31] targets such a matching, with accurate results, however, for a retrieval task. In contrast to previous methods that directly match silhouettes, we formulate shape estimation from silhouettes as a regression problem where global and semantic information on the silhouettes are incorporated by utilizing CNNs. This leads to accurate, robust, and fast body shape estimations without manual interaction, resulting in a practical system.

CNNs in applications With the rebirth of neural networks, classification and recognition tasks were revised [30, 45, 25] and demonstrated more accurate results than previous works. Building on them, there have been recent works using CNNs with 3D shapes for tasks like shape classification and retrieval [49, 47, 17], pose estimation [48], image semantic segmentation [34, 18] and human re-identification [13]. Most of the methods working on shapes though, tackle retrieval or classification applications and are geared towards rigid shapes (like chairs, tables etc.). To a smaller extent, works like [48] and [29] tackle regression with CNNs, however for human or camera pose estimation. It has also been a common theme for most of the previous methods that accept a 2D input to use an RGB or grayscale image, often fine-tuning previous architectures trained on similar inputs. Unlike the above, we newly introduce a method that tries to solve a regression problem, for accurate human shape estimation, by training a CNN from scratch, on binary input images.

Mapping statistical models for body shape estimation A more recent approach for estimating 3D body shapes from silhouettes involves constructing statistical models for both the 3D bodies and 2D silhouettes [5, 12, 11, 51, 44, 10] by handcrafting features. Then, estimation is defined via a mapping between the parameters of these two models. Linear [51] or more complex non-linear models [10] can be defined. These methods rely on a global mapping between 2D and 3D and have been evaluated only on a limited set of measurements. In a concurrent work, Dibra et al. [15] define a fast mapping from specialized silhouette features, projected at correlated spaces, to shape parameters utilizing random forest regressors. A more refined version of [10], that additionally performs a local fitting has been introduced by Boisvert et al. [5], targeting applications similar to ours under more restrictive assumptions. In general, the mentioned methods are not practical for real-time applications due to their high running times. In particular, Boisvert et al. [5] demonstrate a higher accuracy over the rest but on the expense of an optimization procedure used for local fitting. We also learn a global mapping from silhouettes to a parametric 3D shape model, improving accuracy and speed significantly. Unlike the previous methods we utilize CNNs, which allow us to train an end-to-end regressor robust to mild pose changes and silhouette noise, and validate our results on thousands of 3D shapes spanning a great variety of body shapes, with an extensive set of experiments varying in the generality assumptions. We distinguish from other CNN attempts like [40, 47], in that they utilize rigid 3D shapes for matching and retrieval. We further illustrate that our architectures work for different types of inputs such as

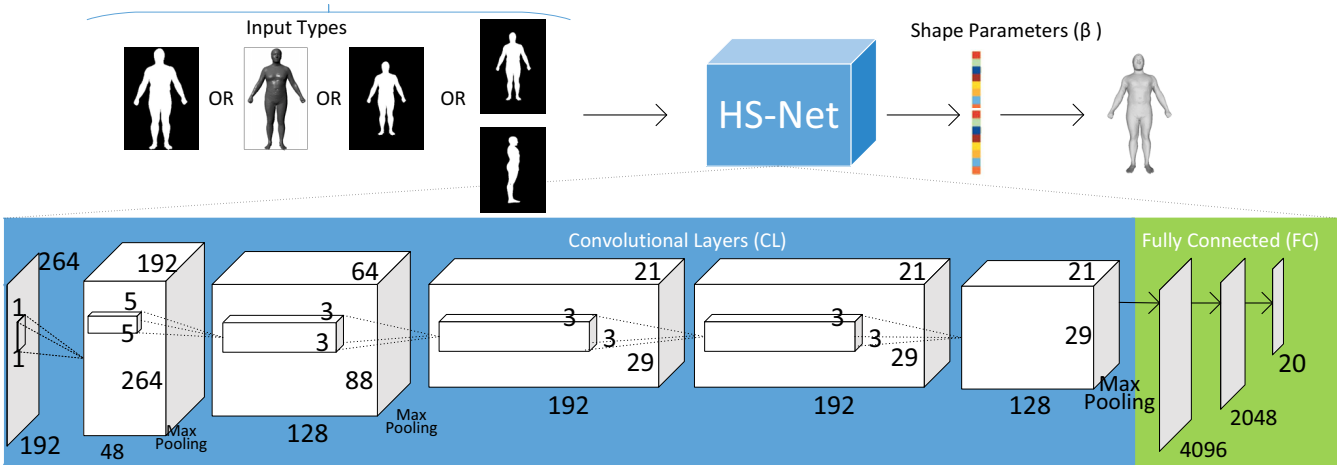

Figure 1: System Overview. **Top:** One of the four input types (scaled frontal silhouette to a fixed height, shaded image, one or two unscaled silhouettes) are fed to the Human Shape Network (HS-Net), to learn a global mapping and estimate human shape parameters (β), which can be used to reconstruct the human body shape. **Bottom:** The HS-Net architecture for the one view case.

multiple silhouettes, or images with shading information.

3. Shape Estimation Algorithm

3.1. Method Overview

Our goal is to design a fast and automatic system to accurately estimate the 3D human shape from silhouettes or images with shading information, for the garment fitting application in mind. More specifically, we would like to learn a global mapping from image evidence to parameters representing the 3D shape utilizing CNNs. With respect to the requirements (and privacy), we categorize image evidence in two groups: silhouette and shaded image. For the first, and least revealing case, extracting silhouettes in general images is not yet fully-automatic, but for our application it is realistic to assume that the person is wearing tight clothes and posing in front of a uniform color background, which simplifies the problem. For the second case on the other hand, the requirement is that the clothing is as minimalistic as possible, due to the fact that our training is based on naked body shapes (Sec. 4). In practice, a shaded image of a real person can be obtained by recovering the intrinsic image [43]. A neutral pose, allowing mild changes, is a reasonable assumption in both cases. While it is true that a 2D image withholds ambiguity per se, our goal is to generate the best approximating 3D mesh that explains the evidence.

A system overview is depicted in Fig. 1 (top), with the input being one of the four input types : scaled frontal silhouette to a fixed height, shaded image, one or two unscaled silhouettes, and as output the reconstructed 3D human shape. We pose shape estimation as an instance of supervised learning. Specifically, we solve a regression problem, where data is generated using a statistical human shape model (Sec. 3.2) based on SCAPE [3]. Utilising parameters spanning from the human shape space, various meshes are reconstructed, from which we obtain silhouettes or shaded images. The parameters themselves are the output, and are used to reconstruct the 3D human shapes. In order to learn a global mapping from the data to the parameters, we do not need to handcraft features as in previous works [33, 44]. We also do not apply local fitting, in contrast to previous works that focus on accuracy [5]. Instead, inspired by recent trends and outstanding results on various computer vision topics, we train CNNs (Sec. 3.3) from scratch, to find the most representative features and a mapping from the image evidence to the human shape. This results in a very fast and automatic system that clearly outperforms methods based on global mapping [51, 10] and strongly competes with expensive methods that adopt local fitting [5].

In the following, we explain the human shape model in Sec. 3.2, the CNN based learning method and architecture in Sec. 3.3, the data generation in Sec. 4 and the results on real and synthetic data in Sec. 5. In Sec. 6, we conclude with a discussion of our method including limitations.

3.2. Shape Model

For human shape estimation problems, from a few camera images [4, 44, 6, 22, 56, 28, 21], deformable shape models are typically a method of choice, in particular SCAPE [3]. That is mainly due to its simplicity. It is a low-dimensional parametric model based on triangle deformations learned from 3D range scans of different people in different poses. The deformations due to body shape and pose changes are captured simultaneously. Below we explain the

model adaptations to our needs, but we advise the reader to refer to the original work [3] for more details. Here, SCAPE is defined as a set of 12894 triangle deformations applied to a reference template 3D mesh consisting of 6449 vertices, with parameters α and β, responsible for pose and intrinsic body shape deformations respectively. Estimating the human body shape implies estimating those parameters. Let $\mathbf{e}_{i1}$ and $\mathbf{e}_{i2}$ be two edges of the i^{th} triangle of the template mesh, defined as the difference vectors between the vertices of the triangle. Given α and β, each of such edges is deformed according to the following expression

$$\mathbf{e}_{ij}^{'} = \mathbf{R}_i(\alpha)\mathbf{S}_i(\beta)\mathbf{Q}_i(\mathbf{R}_i(\alpha))\mathbf{e}_{ij}, \quad (1)$$

with $j \in \{1, 2\}$. $\mathbf{R}_i(\alpha)$, $\mathbf{Q}_i(\mathbf{R}_i(\alpha))$ and $\mathbf{S}_i(\beta)$ are matrices corresponding to joint rotations, pose induced non-rigid deformations and intrinsic shape variation, respectively. In this work, we try to estimate the β parameters. We learn the deformation space of body shapes by stacking triangle deformations of meshes in full correspondence, of different people under the same pose, and then applying PCA. The learned transformations can be written as $\mathbf{s}(\beta) = \mathbf{U}\beta + \mu$, with $\mathbf{U}$ a matrix with orthonormal columns, and μ the mean of the triangle transformations over all the meshes. Our mesh set contains roughly 5000 different meshes gathered from two available datasets put in full correspondence, from which 1500 are put aside for testing and are not used for learning the model, as explained in Sec. 4. The template mesh is computed as the mean over the remaining meshes. After computing the per-triangle deformations from the template mesh over each mesh, we apply PCA in order to extract the components capturing the largest deformation variations. We noticed that 20 components are enough to capture more than 95% of the energy, hence $\beta \in R^{20}$.

One could learn the deformation space due to joint rotations $\mathbf{R}_i(\alpha)$ applying a similar procedure. In contrast to $\mathbf{S}_i(\beta)$ though, the same human mesh but in different poses should be used by applying a similar procedure. We however, are not interested in estimating α, but want to estimate β only. This also relates to the important fact that the intrinsic shape $\mathbf{S}_i(\beta)$ should not change over different poses. Similarly, pose induced deformations $\mathbf{Q}_i(\mathbf{R}_i(\alpha))$ are also not needed, due to the fact that our application expects a neutral standing pose, allowing mild pose changes around it. Taking the common assumption that the body shape does not significantly change due to the range of poses we consider, we decouple pose changes from shape changes. In this way, solving equation 1 boils down to solving a simplified version of it hence $\mathbf{e}_{ij}^{'} = \mathbf{S}_i(\beta)\mathbf{e}_{ij}$. This is a common assumption used in previous works as well [37, 28], where fast pose changes are performed adopting an efficient method from the graphics community instead, known as Linear Blend Skinning (LBS) [32]. Reconstructing a new

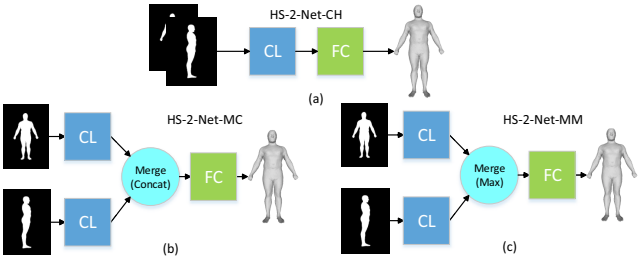

Figure 2: The three architectures considered for two input silhouettes. (a) both silhouettes are input as two channels (b) each silhouette is input into two separate convolutional layer (CL) blocks and outputs of the CL are concatenated through a *Merge* layer (c) same scenario however with a *Max* operation performed instead.

shape utilizing SCAPE involves solving a least-squares system over newly estimated shape parameters, which runs in milliseconds.

3.3. Learning A Global Mapping

We pose the global mapping as an end-to-end regression problem, from 2D input image to shape parameters. We achieve this by training from scratch a CNN similar to that of AlexNet [30] and adapting it to our inputs and regression task, as depicted in Fig. 1 (bottom). Regarding the number of input images we distinguish two cases : A frontal single view image, coming in different forms, and two images simultaneously, from front and side.

One View The frontal view image can come in three forms. Firstly, a frontal binary silhouette of the human in a neutral pose, scaled to a fixed height is considered. This is the most general case, and assumes unknown camera calibration, hence the need for a fixed scaling. Second, if the camera parameters are known, e.g. when the person stands a known distance away from the camera, the input is a fixed size image of varying silhouette size and height. Estimating the real 3D shape from a 2D input silhouette is an ill-posed problem per se, due to the fact that a silhouette can represent various body shapes, even though we strive to reconstruct the shape that best explains it. Utilizing silhouettes only, has the advantage that no personal information is revealed, which is important for privacy protection. Allowing the problem to be a bit more relaxed, by adding further information, we lastly consider the case of using additional image cues such as shading, complementary to the scaled silhouette, similar to [22]. In order to synthetically generate training data, we render images with shading under Lambertian assumptions. In practice a similar result could be achieved by extracting the intrinsic image [43]. The input size for all the mentioned methods is set to 264×192 pixels. For each case, the single channel input images, along with the known shape parameters, are fed into our *Human*

Shape Network (HS-Net), which learns the mapping from input to the shape parameters β. *HS-Net* is a modification of Alexnet [30] customized to a regression problem, our various input types, intended application and the available hardware. The network consists of five convolutional blocks, followed by three fully connected layers as illustrated in Fig. 1 (bottom). Each layer is followed by an activation layer (ReLu). In addition, dropout layers are utilized between fully connected layers to avoid overfitting and max pooling is used after the first, second and fifth convolutional blocks. The network is trained from scratch, since the available pre-trained models are geared towards classification and RGB or grayscale images, while we tend to learn regression from binary images. We experimented with different optimization algorithms and observed that the best results were obtained using *RMSProp* [1] and *Adadelta* [55]. We decided to utilize the Adadelta optimizer due to its capacity to automatically adjust the learning rate and prevent it from becoming too small.

Two Views In compliance with the realistic scenario of estimating the body shape and the body parts measurements as accurately as possible, we additionally opted for the usage of two silhouettes simultaneously, where the person is seen from a full frontal and side view. This setting also assumes known camera parameters, same as the methods we compare to [51, 10, 5], which translates to knowing the distance from the camera. One of the challenges of this case is how to combine multiple view inputs in a way that the convolutional network can use them coherently. We explore and evaluate three different approaches to achieve this. The first approach, utilizes a model architecture very similar to the one view case, however the input images from the different views are stacked along the channel dimension to form two channel images, see Fig. 2 (a). These two channel images are then fed into the network for training. By visualizing the output filters (supplementary [1]) for different layers on various test images, we observed that the network learns some filters more pronounced towards frontal views, while others favor the side views. For the second approach, the architecture differs from the previous case, in that we add a *Merge* layer similar to the view pooling layer of [47], after two sets of convolutional layers with shared weights for each view, followed by fully connected layers. The input images from each view are fed into two separate five layer convolutional networks and merged using a concatenation operation, see Fig. 2 (b). The third approach distinguishes from the second one in that the merge layer performs a *Max* operation over each dimension, see Fig. 2 (c). The motivation behind the last two approaches, was to allow the network to separately learn features from individual images and then fuse them more discriminatively through a merging layer. The merge layer with max operation improves learning and subsequently the estimation accuracy (see Tab. 1) over the two channel network, as it combines evidence at a later stage of learning. All three methods lead to improvements over the one view case, which we demonstrate in Sec. 5, where the merging with max operation performs the best.

4. Data Generation

Our method is based on training a CNN hence it requires numerous training and validation data. Gathering a big number of human shapes is a highly non-trivial task - due to the need of specialized equipments for scanning people, the difficulty of finding large numbers of them, and more importantly, due to the necessity of scanning them under minimalistic clothing, in order to better capture the intrinsic shapes. Unfortunately, there exists no freely available dataset of real human body shapes along with measurements. A feasible solution though, would be to learn a parametric shape model from a small subset of body shapes capturing body shape variances and generate synthetic data from it. Taking advantage of the commercially available CAESAR dataset [2], containing people in an almost naked apparel, researchers have released two datasets [53, 37], consisting of meshes obtained by fitting a template mesh to subsets of the CAESAR dataset. We merge these two datasets and construct a larger one, to enable learning a more general shape model. One of them [53], consists of around 1500 registered meshes in correspondence, however of higher resolution than the other dataset [37]. The resolutions respectively are 12500 vertices 25000 triangles and 6449 vertices 12894 triangles. Mesh resolution is not so important for our application, hence we map the higher resolution meshes to the lower resolution ones. This also improves the computation time. To achieve that, we first extract a template mesh, as the mean mesh of each dataset, and then apply non-rigid ICP [2] to the two template meshes. Afterwards, closest points in both meshes are computed, using barycentric coordinates in the closest triangle. The retrieved mapping can be applied to all the remaining meshes due to the same mesh connectivity. Roughly 3000 meshes are selected from the combined dataset to learn the shape model, leaving out around 1500 meshes for experiments and validation. Applying the method from Sec. 3.2, we extract 20 principal components from the triangle deformations that capture most of the variance. For synthesizing new meshes, we sample from the 20 dimensional multivariate normal distribution spanned by the PCA space, where for a random sample $\beta = [\beta_1, \beta_2, ..., \beta_{20}]$, it holds that $\beta \sim \mathcal{N}(\mu, \Sigma)$. μ and Σ are the 20-dimensional mean vector and the 20×20 covariance matrix respectively. For training, we generate 100000 meshes from the multi-variate dis-

[1] http://www.cs.toronto.edu/ tijmen/csc321/slides/lecture_slides_lec6.pdf

[2] http://store.sae.org/caesar/

tribution over the β parameters. To avoid potentially unrealistic humans, we restrict the parameters to be in the $\pm 3 \times Std.Dev$ range in each PCA dimension. We generate silhouettes by projecting the meshes into frontal and side views, and shaded images by rendering the meshes with shading under lambertian assumptions, similar to [47]. These are fed to the respective CNNs, by first scaling them to 264×192 pixel resolution. Our evaluation and testing set comprises of the meshes left out from the training dataset, as well as of real images of people standing in front of a wall.

5. Validation and Results

Our method targets the application of human body shape and body parts estimation. In order to assess its reliability, one can not rely only on the visual reconstruction of the mesh. Rigorous quantitative experiments are necessary, especially of measurements over various important body parts. If the latter can be estimated accurately, fitting clothes virtually or even buying clothes online becomes more intuitive and appealing. Measuring different body parts consistently in real datasets is difficult, as even the most trained individuals are reported to deviate up to 10 mm [20]. Hence, we evaluate on synthetic meshes, obtained by fitting a parametric model to real people scans from the CAESAR dataset, similar to the methods we compare to [5]. Performing the evaluations on this dataset, in addition to the shapes being very close to the real ones, has the advantage that they are in full correspondence. Thus, it becomes easy to automatically measure various body parts. Additionally, the poses adopted from the real human scans, deviate from the neutral pose specified by experimenters while they are being scanned, Fig. 3 (top-left). These meshes are quite realistic and in compliance with the variation of the poses that people adopt for our target applications. Different openings of the arms, legs and even shoulders can be noticed, while we show results of more pronounced poses in the supplementary material [1]. In our experiments, we apply the same measurements as in [5], Fig. 3 (top-right). For our evaluation we use 1500 meshes and 4 real people on 16 body measurements, which to the best of our knowledge is the most complete one so far, as compared to related work. Boisvert et al. [5] evaluate on 220 meshes and 4 real people, Xi et al. [51] on 24 meshes and two real people, Sigal et al. [44] for two measurements only on two subjects and Balan et al. [4] for silhouette errors and height measurement on a few individuals.

General Training and Set-Up details For each of the 100000 generated meshes, Sec. 4, we generate silhouettes from frontal and side views, as well as shaded images under lambertian assumptions with Maya[3]. As preprocessing, the images are centered, normalized to the [0,1] interval and fixed to the 264×192 pixels resolution for all the cases. The resolution was chosen such that it neither impedes learning of shape variations, nor is too big, due to the hardware and time constraints we had. We use 95000 images for training and 5000 for the network validation. As explained above, the testing is performed on 1500 unseen samples and real human ones. The network architecture is detailed in Fig. 1 (bottom) for the one view input case. For the various experiments that we perform, we change the networks as explained in Sec. 3.3 and adopt the following nomenclature : *HS-1-Net-S* and *HS-1-Net* for the scaled and unscaled input silhouette and *HS-1-Net-Im* for the scaled shaded image input. Training usually converges between 15-25 epochs depending on the experiment. The batch size was set to 32, to not be a proper divisor of the number of training samples per epoch, which is equal to half of total training samples. This provides an easy way to simulate shuffling without hitting memory constraints for such big datasets. We also experimented with batch normalization [27] right after the convolutional layers, resulting in slight error increase. Applying batch normalization after the fully connected layers though, caused the network to converge to constant functions.

We experimented with the RMSprop, Adagrad [16] and Adadelta [55] optimizers, in order to minimize the manual learning rate adjustments. We observed that RMSprop (with an initial learning rate of 0.001) and Adadelta (with decay rate of 0.95) converged faster than Adagrad, also with a smaller test error. Thus, all the reported experiment results are for the models trained using Adadelta. We experimented with the squared loss, with and without multiplying the last fully-connected layer by custom weights. The weights are set to be the eigenvalues of the covariance matrix obtained from PCA, during the data generation step Sec. 3.2 and normalized to 1, such that we emphasize large scale changes in 3D body shapes. As expected, using squared loss with custom weights performed better. For all the networks, we utilized Glorot uniform weight initialization [19].

For the two view case, we used the best performing network configurations from the one view case, however the architectures were modified to fit the input extension, as shown in Fig. 2. The two selected views were the frontal and the side one. We also distinguish between three cases here : *HS-2-Net-CH* for the input silhouettes passed as two channels of a single image, *HS-2-Net-MM* for separately training the two inputs as different single channel images and applying a merge layer, that performs a max operation over each dimension right after the output of the last convolutional layers (CL), and *HS-2-Net-MC* for the same architecture that concatenates the output of CL, instead of max operation. All the CL have shared weights.

Quantitative Experiments We perform 16 3D measurements on the test meshes which consist of males and fe-

[3] http://www.autodesk.com/products/maya/

males in roughly equal numbers, similar to [5]. The measurements are illustrated in Fig. 3 (top-right) and are widely used in garment fitting. We compute the Euclidean distance between two extreme vertices for the straight line measurements, while for the ellipsoidal ones, the perimeter is computed on the body surface. For each measurement we calculate the difference between the value estimated and the ground truth, and report the mean error and standard deviation computed over the error values for all the test meshes in Table 1. Additionally, we show how the mean error over all measurements varies, for each different input type that we consider, in Fig. 3. *HS-2-Net-MM* has the lowest error of 4.02 mm, as compared to 11 mm of [5], which utilizes a more expensive local fitting algorithm. For completeness, we compare to the work of Helten et al. [26], that utilizes an RGB-D camera for capturing the body shapes, and a full RMSE map per vertex to measure the differences. They report an error of 10.1 mm, evaluating on 6 individuals from two depth maps, while we report an error of 7.4 mm on 1500 meshes.

We observed that using weights with squared loss function increases the accuracy of the model. The model trained on silhouettes with known camera parameters performs significantly better than the one with unknown camera calibration. The shaded images network *HS-1-Net-Im*, performs also slightly better than the corresponding silhouette one *HS-1-Net-S*, implying that shading information possibly improves the shape estimation accuracy, but could also be related to added information due to grayscale input as opposed to a binary one. Lastly, *HS-2-Net-CH* demonstrates more accuracy for the ellipsoidal errors while *HS-2-Net-MC* for the euclidean ones, despite their overall similar performance. In comparison to the other methods, our network clearly outperforms the global methods [51, 10] for all the input types, as well as the more involved method of [5], except for the Overal Height measurement (O) and Inside Leg length (K). Adding a second view gives better results than a single view with noticeable improvements in the height and waist estimation.

Noise Due to the imperfection of silhouette extraction algorithms, we evaluated the robustness of our model under the influence of noise. We apply noise to the silhouette by randomly eroding or dilating the silhouette at the border, with filters of various radii, evaluating it for 1,3,5,7 and 9 pixels. We plot the errors of each body measurements and show examples of noisy silhouettes for a radius of 1, 5 and 9 pixels in Fig. 4. The method achieves performance similar to the noiseless case within a reasonable noise radius, where even for the highest noise parameter the maximum error (in the height) is below 5 mm, implying robustness to noise.

Qualitative Results We demonstrate results of frontal body shapes, obtained by applying *HS-1-Net-S* over scaled silhouettes, extracted from images of real people in Fig. 5.

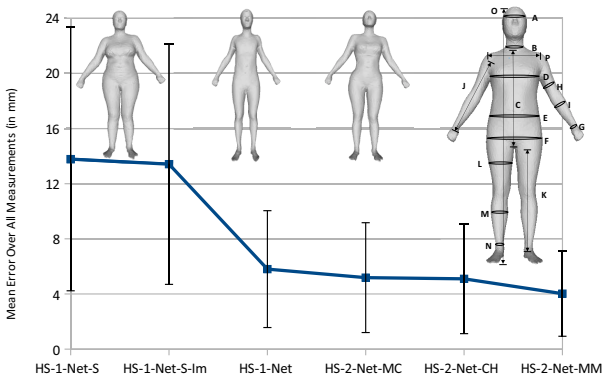

Figure 3: Mean error over all measurements for different input types. (top-left) 3 test meshes in slightly changing poses. (top-right) Illustration of the body measurements (A - P) on the template mesh.

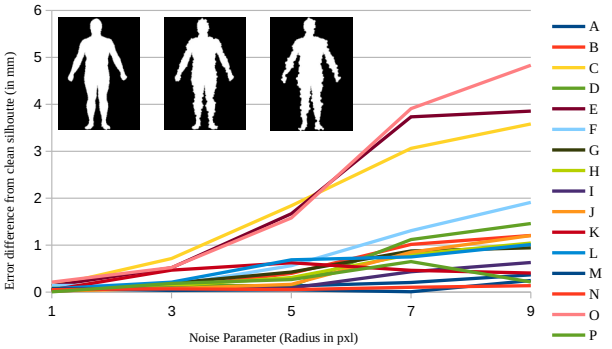

Figure 4: Error plots for each of the body measurements (A - P) when noise is applied, as compared to clean silhouettes. (top-left) 3 silhouettes with noise parameters 1, 5 and 9. (Figure best seen in colors).

Figure 5: Mesh reconstruction for 4 real subjects in mildly varying poses. (left) Input image (middle) Extracted Silhouette (right) Reconstruction of the estimated shape.

The individuals adopt a neutral pose, however please note the variations in the arms and legs openings. Our method manages to reconstruct accurate shapes, also backing up our

Measurement	HS-1-Net-S	HS-1-Net-S-Im	HS-1-Net	HS-2-Net-MC	HS-2-Net-CH	HS-2-Net-MM	[5]	[12]	[51]
A. Head circumference	4±4	4±4	2±4	2±3	2±3	**2±3**	10±12	23±27	50±60
B. Neck circumference	8±5	6±4	3±1	2±1	3±1	**2±1**	11±13	27±34	59±72
C. Shoulder-blade/crotch length	20±15	20±14	7±7	5±6	4±5	**3±5**	4±5	52±65	119±150
D. Chest circumference	13±7	13±6	4±1	2±1	4±2	**2±1**	10±12	18±22	36±45
E. Waist circumference	19±13	19±13	8±7	6±7	8±7	**7±5**	22±23	37±39	55±62
F. Pelvis circumference	19±14	19±12	6±5	5±4	6±5	**4±4**	11±12	15±19	23±28
G. Wrist circumference	5±3	5±3	3±2	2±1	3±2	**2±2**	9±12	24±30	56±70
H. Bicep circumference	8±4	8±3	2±1	2±1	2±1	**2±1**	17±22	59±76	146±177
I. Forearm circumference	7±4	6±3	2±1	2±1	2±1	**1±1**	16±20	76±100	182±230
J. Arm length	12±8	12±8	6±4	5±4	5±4	**3±2**	15±21	53±73	109±141
K. Inside leg length	20±14	19±13	12±8	13±9	11±7	9±6	**6±7**	9±12	19±24
L. Thigh circumference	13±8	12±7	8±5	7±4	7±4	**6±4**	9±12	19±25	35±44
M. Calf circumference	12±7	11±6	5±2	5±2	4±2	**3±1**	6±7	16±21	33±42
N. Ankle circumference	6±3	5±2	3±1	2±1	3±1	**2±1**	14±16	28±35	61±78
O. Overall height	50±39	49±37	20±15	19±15	16±13	12±10	**9±12**	21±27	49±62
P. Shoulder breadth	4±4	3±4	3±4	2±4	2±4	**2±4**	6±7	12±15	24±31

Table 1: Error comparisons on body measurements for the various inputs and presented training modalities, as well as state-of-the-art methods (last three columns). The measurements are illustrated in Fig. 3 (top-right). Errors are represented as Mean±Std. Dev and are expressed in millimeters. Our best achieving method *HS-2-Net-MM* is highlighted.

claim that mild pose changes do not affect our robustness. We provide further qualitative and quantitative results on synthetic meshes in the supplementary material [1].

Method Speed We conducted our experiments on an Intel(R) Core(TM) i7 CPU 950 3.0 GHz with NVIDIA GTX 980TI (6GB) GPU. The training code was implemented in Python using Keras framework [4] with Tensorflow as backend. The usual training time is around 30 minutes per epoch and the testing time was about 0.2 seconds per image. Generating a mesh from the estimated parameters takes around 0.25 seconds (significant further speed-up is possible via parallelization of this step). Our full algorithm runs in 0.45 seconds and is significantly faster than the methods we compare to with 3 minutes and 36 seconds for the full optimization of [5] and 6 seconds for the global mapping of [12].

6. Discussion and Conclusions

We presented a novel technique that can estimate 3D human body shape from silhouettes or shaded images quite accurately utilizing CNNs. We posed the problem as regression, where we try to find a global mapping from the various inputs that we presented to shape parameters. We extensively evaluated our technique on thousands of human bodies and real people.

In compliance with our main target applications, e.g. garment fitting, we mainly focused on shape estimation of people in neutral poses allowing mild pose changes, from one scaled or unscaled, two binary silhouettes as well as shaded images as input. We showed that we outperform methods based on global mapping and achieve similar results to more expensive methods that employ local fitting.

In the scope of the networks that we experimented with, we showed how to simultaneously combine two binary silhouettes in order to improve prediction over a single one, and evaluated three different methods. We believe that this sets a ground for future works in human shape estimation from multiple views.

We also demonstrated in a synthetic experiment, that if shading information is present, better results are achievable. Due to lack of real data though, it is difficult to assess its performance on real humans and believe that as intrinsic image extraction algorithms improve, it will lead to future works in this domain .

Even though silhouette extraction is not a bottleneck for our target scenarios, due to assumptions on uniform backgrounds, we evaluated the performance under the influence of noise of different levels, and showed that our method is robust to silhouette noise under reasonable assumptions. We further assumed humans in tight clothes. Applying our method to a scenario where clothed people are present deteriorates the results, however in contrast to previous works the reconstructions remain in the space of plausible human bodies.

A limitation to our method is that with the current training, it can not handle poses that differ significantly from the neutral pose and contain self-occlusions. We could handle that by generating a larger training set including more pronounced poses, which goes beyond the scope of the paper.

Lastly we showed that our system is orders of magnitude faster than the methods we compare to. Based on recent works [8] that try to compress Neural Networks as well as the possible speed-up of our mesh computation, our algorithm which already runs at interactive rates, could be integrated into smartphones in the foreseeable future.

Acknowledgment. This work was funded by the KTI-grant 15599.1.

[4] http://keras.io/

References

[1] https://cgl.ethz.ch/publications/papers/paperDib16b.php. 5, 6, 8

[2] B. Amberg, S. Romdhani, and T. Vetter. Optimal step nonrigid ICP algorithms for surface registration. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*, 2007. 5

[3] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: Shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05, pages 408–416, New York, NY, USA, 2005. ACM. 1, 2, 3, 4

[4] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*, 2007. 2, 3, 6

[5] J. Boisvert, C. Shu, S. Wuhrer, and P. Xi. Three-dimensional human shape inference from silhouettes: reconstruction and validation. *Mach. Vis. Appl.*, 24(1):145–157, 2013. 1, 2, 3, 5, 6, 7, 8

[6] A. O. Bălan and M. J. Black. The naked truth: Estimating body shape under clothing. In *Proceedings of the 10th European Conference on Computer Vision: Part II*, ECCV '08, pages 15–29, Berlin, Heidelberg, 2008. Springer-Verlag. 2, 3

[7] D. Casas, M. Volino, J. Collomosse, and A. Hilton. 4d video textures for interactive character appearance. *Computer Graphics Forum (Proc. Eurographics 2014)*, 33(2), 2014. 1

[8] W. Chen, J. T. Wilson, S. Tyree, K. Q. Weinberger, and Y. Chen. Compressing convolutional neural networks. *CoRR*, abs/1506.04449, 2015. 8

[9] X. Chen, Y. Guo, B. Zhou, and Q. Zhao. Deformable model for estimating clothed and naked human shapes from a single image. *The Visual Computer*, 29(11):1187–1196, 2013. 1, 2

[10] Y. Chen and R. Cipolla. Learning shape priors for single view reconstruction. In *ICCV Workshops*. IEEE, 2009. 1, 2, 3, 5, 7

[11] Y. Chen, T. Kim, and R. Cipolla. Silhouette-based object phenotype recognition using 3d shape priors. In *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pages 25–32, 2011. 2

[12] Y. Chen, T.-K. Kim, and R. Cipolla. Inferring 3d shapes and deformations from single views. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *ECCV (3)*, volume 6313 of *Lecture Notes in Computer Science*, pages 300–313. Springer, 2010. 2, 8

[13] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016. 2

[14] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. In *ACM SIGGRAPH 2008 Papers*, SIGGRAPH '08, pages 98:1–98:10, New York, NY, USA, 2008. ACM. 1

[15] E. Dibra, C. Oztireli, R. Ziegler, and M. Gross. Shape from selfies : Human body shape estimation using cca regression forests. In *Proceedings of the 14th European Conference on Computer Vision: Part IV*, ECCV '16, 2016. 2

[16] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011. 6

[17] Y. Fang, J. Xie, G. Dai, M. Wang, F. Zhu, T. Xu, and E. Wong. 3d deep shape descriptor. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1, 2

[19] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS10). Society for Artificial Intelligence and Statistics*, 2010. 6

[20] C. C. Gordon, T. Churchill, C. E. Clauser, B. Bradtmiller, and J. T. McConville. Anthropometric survey of US Army personnel: Summary statistics, interim report for 1988. Technical report, DTIC Document, 1989. 6

[21] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black. Drape: Dressing any person. *ACM Trans. Graph.*, 31(4):35:1–35:10, July 2012. 1, 3

[22] P. Guan, A. Weiss, A. O. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 1381–1388, 2009. 2, 3, 4

[23] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormählen, and H. Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 1823–1830, 2010. 2

[24] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H. Seidel. A statistical model of human pose and body shape. *Comput. Graph. Forum*, 28(2):337–346, 2009. 2

[25] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 2

[26] T. Helten, A. Baak, G. Bharaj, M. Müller, H. Seidel, and C. Theobalt. Personalization and evaluation of a real-time depth-based full body tracker. In *2013 International Conference on 3D Vision, 3DV 2013, Seattle, Washington, USA, June 29 - July 1, 2013*, pages 279–286, 2013. 7

[27] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 448–456, 2015. 6

[28] A. Jain, T. Thormählen, H.-P. Seidel, and C. Theobalt. Moviereshape: Tracking and reshaping of humans in videos.

ACM Trans. Graph. (Proc. SIGGRAPH Asia 2010), 29(5), 2010. 1, 2, 3, 4

[29] A. Kendall, M. Grimes, and R. Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2938–2946, 2015. 2

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1, 2, 4, 5

[31] Z. Lahner, E. Rodola, F. R. Schmidt, M. M. Bronstein, and D. Cremers. Efficient globally optimal 2d-to-3d deformable shape matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2

[32] J. P. Lewis, M. Cordner, and N. Fong. Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pages 165–172, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co. 4

[33] H. Ling and D. W. Jacobs. Shape classification using the inner-distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(2):286–299, Feb. 2007. 1, 3

[34] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 1, 2

[35] A. Neophytou and A. Hilton. Shape and pose space deformation for subject specific animation. In *Proceedings of the 2013 International Conference on 3D Vision*, 3DV '13, pages 334–341, Washington, DC, USA, 2013. IEEE Computer Society. 2

[36] A. Neophytou and A. Hilton. A layered model of human body and garment deformation. In *2nd International Conference on 3D Vision, 3DV 2014, Tokyo, Japan, December 8-11, 2014*, pages 171–178, 2014. 1

[37] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele. Building statistical shape spaces for 3d human modeling. *CoRR*, abs/1503.05860, 2015. 1, 2, 4, 5

[38] K. M. Robinette and H. A. M. Daanen. The caesar project: A 3-d surface anthropometry survey. In *2nd International Conference on 3D Digital Imaging and Modeling (3DIM '99), 4-8 October 1999, Ottawa, Canada*, pages 380–387, 1999. 2

[39] L. Rogge, F. Klose, M. Stengel, M. Eisemann, and M. Magnor. Garment replacement in monocular video sequences. *ACM Trans. Graph.*, 34(1):6:1–6:10, Dec. 2014. 1

[40] M. Savva, F. Yu, H. Su, M. Aono, B. Chen, D. Cohen-Or, W. Deng, H. Su, S. Bai, X. Bai, et al. Shrec16 track large-scale 3d shape retrieval from shapenet core55. 2

[41] F. R. Schmidt, D. Farin, and D. Cremers. Fast matching of planar shapes in sub-cubic runtime. In *ICCV*, pages 1–6. IEEE Computer Society, 2007. 2

[42] F. R. Schmidt, E. Töppe, and D. Cremers. Efficient planar graph cuts with applications in computer vision. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, Florida, Jun 2009. 2

[43] J. Shen, X. Yang, Y. Jia, and X. Li. Intrinsic images using optimization. In *CVPR*, pages 3481–3487. IEEE Computer Society, 2011. 3, 4

[44] L. Sigal, A. O. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *NIPS*. Curran Associates, Inc., 2007. 1, 2, 3, 6

[45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 2

[46] J. Starck, G. Miller, and A. Hilton. Video-based character animation. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '05, pages 49–58, New York, NY, USA, 2005. ACM. 1

[47] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 945–953, 2015. 1, 2, 5, 6

[48] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014. 1, 2

[49] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2

[50] S. Wuhrer, L. Pishchulin, A. Brunton, C. Shu, and J. Lang. Estimation of human body shape and posture under clothing. *Computer Vision and Image Understanding (CVIU)*, June 2014. 1

[51] P. Xi, W. Lee, and C. Shu. A data-driven approach to human-body cloning using a segmented body database. In *Proceedings of the Pacific Conference on Computer Graphics and Applications, Pacific Graphics 2007, Maui, Hawaii, USA, October 29 - November 2, 2007*, pages 139–147, 2007. 1, 2, 3, 5, 6, 7, 8

[52] F. Xu, Y. Liu, C. Stoll, J. Tompkin, G. Bharaj, Q. Dai, H.-P. Seidel, J. Kautz, and C. Theobalt. Video-based characters: Creating new human performances from a multi-view video database. In *ACM SIGGRAPH 2011 Papers*, SIGGRAPH '11, pages 32:1–32:10, New York, NY, USA, 2011. ACM. 1

[53] Y. Yang, Y. Yu, Y. Zhou, S. Du, J. Davis, and R. Yang. Semantic parametric reshaping of human body models. In *2nd International Conference on 3D Vision, 3DV 2014, Tokyo, Japan, December 8-11, 2014, Volume 2*, pages 41–48, 2014. 1, 2, 5

[54] M. Ye and R. Yang. Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 2353–2360. IEEE, 2014. 1

[55] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012. 5, 6

[56] S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han. Parametric reshaping of human bodies in images. *ACM Trans. Graph.*, 29(4), 2010. 1, 2, 3