Shape from Selfies: Human Body Shape Estimation Using CCA Regression Forests

Endri Dibra[1(✉)], Cengiz Öztireli[1], Remo Ziegler[2], and Markus Gross[1]

[1] Department of Computer Science, ETH Zürich, Zürich, Switzerland
{edibra,cengizo,grossm}@inf.ethz.ch
[2] Vizrt, Zürich, Switzerland
rziegler@vizrt.com

Abstract. In this work, we revise the problem of human body shape estimation from monocular imagery. Starting from a statistical human shape model that describes a body shape with shape parameters, we describe a novel approach to automatically estimate these parameters from a single input shape silhouette using semi-supervised learning. By utilizing silhouette features that encode local and global properties robust to noise, pose and view changes, and projecting them to lower dimensional spaces obtained through multi-view learning with canonical correlation analysis, we show how regression forests can be used to compute an accurate mapping from the silhouette to the shape parameter space. This results in a very fast, robust and automatic system under mild self-occlusion assumptions. We extensively evaluate our method on thousands of synthetic and real data and compare it to the state-of-art approaches that operate under more restrictive assumptions.

1 Introduction

Estimating human body shape from imagery is an important problem in computer vision with diverse applications. The estimated body shape provides an accurate proxy geometry for further tasks such as rendering free viewpoint videos [10,48,49,53], surveillance [11], tracking [16], biometric authentication, medical and personal measurements, virtual cloth fitting [17,36,40,51], and artistic image reshaping [56]. Pose estimation is also tightly coupled with shape estimation. Knowing the body shape significantly reduces the complexity and improves the robustness of pose estimation algorithms and thus expands the space of poses that can be reliably estimated [2,55].

However, in contrast to pose estimation, body shape estimation has received substantially less attention from the community. Most existing algorithms rely on either manual input [25,40,56], restrictive assumptions on the acquired images [6],

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46493-0_6) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part IV, LNCS 9908, pp. 88–104, 2016.
DOI: 10.1007/978-3-319-46493-0_6

or require information other than just 2D images (e.g. depth) [23,37,50]. Furthermore, some of the methods have prohibitive complexity for real-time applications [6,25,50]. For practical applications, it is essential to have an automatic and fast algorithm that can work with images acquired under less restrictive conditions and body poses.

In this paper, we propose a fast and automatic method for estimating the 3D body shape of a person from images, utilizing *multi-view* semi-supervised learning. Our method relies on extracting novel features from a given silhouette of a single person under minimal self-occlusion like in a selfie, and a parametric human body shape model [3]. The latter is utilized to generate meshes spanning a spectrum of human body shapes, from which silhouettes are computed over multiple views, in poses compliant with the target applications for training. We firstly estimate viewing direction with high accuracy, by solving a classification task. Utilizing the information simultaneously captured in multiple synthetic views of the same body mesh, we apply Canonical Correlation Analysis (CCA) [24] to learn informative bases where the extracted features can be projected. A random forest regressor is then adopted to learn a mapping from projected feature space to parameter space. This results in lower feature dimensionality, reducing the training and test time drastically, and improves prediction as compared to plain regression forests. We demonstrate our results on real people and in a free-view-point video (supplementary [1]), and comprehensively evaluate our method by validating it on thousands of body shapes.

Contributions. In summary, the contributions of the paper are: (1) a fast and automatic system for shape estimation from monocular silhouette/s under no fixed pose and known camera assumptions, thanks to novel features that capture robust global and local information simultaneously, (2) demonstration of how CCA multi-view learning with regression forests can be applied to the task of shape estimation, leveraging synthetic data and improving prediction over random forests with raw data, (3) extensive validation on thousands of body shapes via thorough comparisons to state-of-the-art on a new bigger dataset.

2 Related Work

General methods for shape estimation. Estimating 3D geometry of body shapes from limited imagery is an inherently ill-posed problem. Early methods used simplifying assumptions such as the visual hull [30] or simple body models with geometric primitives [15,27,34]. Although these work well for coarse pose and shape approximations, an accurate shape estimation cannot be obtained.

Human body shape statistical priors. Instead of assuming general geometry, human body shape model based methods rely on the limited degrees of freedom for the possible body shapes. These parametric models are typically constructed from collected 3D scans of people [3,22,39]. Utilizing such a prior allows us to always stay within the space of realistic body shapes, and reduces the problem to estimating the parameters of the model. Such models can also

be combined with articulation models to simultaneously represent pose as joint angles or transformations, and shape with parameters [3,22,35]. In this paper, we combine state-of-the-art 3D body shape databases [38,54] containing thousands of meshes, and utilize a deformation model based on SCAPE [3].

Fitting body shapes by silhouette matching. Once a statistical model of 3D shapes is defined, an error metric between the input silhouettes and those of the projections of the parameterized 3D body shape can be minimized [4,9,11,18,21,25,56]. Although this leads to accurate matching, despite promising results on deformable 2D shape matching [42,43], establishing correspondences between the input and output silhouettes is a very challenging problem especially when the body pose is not known or self occlusions are present. The simultaneous estimation of pose and shape is currently addressed by manual interaction to establish and refine matching or pose estimation [11,25,56], and under certain assumptions on the error metric, camera calibration, and views [9,18,25]. A recent work [29] aims at automatically finding a correspondence between 2D and 3D deformable objects by casting it as an energy minimization problem, demonstrating good results however for a shape retrieval task. Instead of fitting silhouettes directly and locally, we consider a global mapping from silhouettes to shape parameters that is invariant to various poses under mild self-occlusion assumptions. This allows us to sidestep pose estimation, avoid any manual interaction, and estimate the shape parameters for imperfect silhouettes interactively, all of which are essential components for a practical shape estimation system.

Fitting body shapes by statistical models. A recent body of works rely on global mappings between silhouettes and 3D body shapes [6,12–14,46,52]. These methods rely on a statistical model for body shapes as well as silhouettes. In this case, the problem reduces to estimating this mapping by various linear [52], or more complex techniques such as the shared Gaussian process latent variable model [12]. In order to generate robust and accurate shapes, these techniques typically require pre-defined and accurate poses [6,12,52], and have been validated with limited measurements except for the recent work of Boisvert et al. [6]. The running times can also be prohibitive for real-time applications [6,13]. We also define a mapping from the silhouette to a statistical shape space. However, we aim at robustness to pose changes and silhouette noise via computing specialized features, projecting them at correlated spaces and training a regressor with random forests. This allows close to real-time performance, unlocking further applications. We further present an extensive evaluation of our shape estimation with thousands of test cases and tens of body measurements.

Multi-view learning. Canonical Correlation Analysis (CCA) [24] and Kernel-CCA [20] are statistical learning techniques that find maximally correlated linear and non-linear projections of two random vectors. The projected spaces learn representations of two data views such that each view‘s predictive ability is mutually maximized. Hence, information present in either view that is uncorrelated with the other view is automatically removed in the projected space. That is a

helpful property in predictive tasks. The aforementioned methods have been used for unsupervised data analysis with multiple views [19], fusing learned features for better prediction [41], reducing sample complexity using unlabeled data [26], or when multiple views are hallucinated from one single view [33]. A generalized version of CCA [45] has also been proposed but for a classification and retrieval task. Despite its power, CCA in combination with regression has found little usage since its proposal [26]. It has only been empirically evaluated for linear regression [33], and utilized for an action recognition classification task [28]. To the best of our knowledge, we are the first to apply CCA in a non-linear regression task for shape estimation, illustrating its power for such non-linear problems.

3 Shape Estimation Algorithm

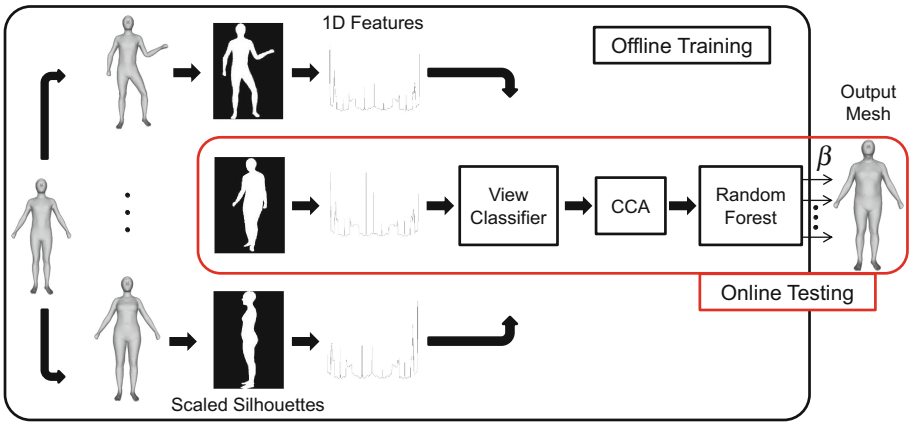

Fig. 1. Overview of our system. **Training:** Silhouettes from 36 views are extracted from meshes generated in various shapes and poses (Sect. 3.2). A View Classifier is learned (Sect. 3.4) from extracted silhouette features (Sect. 3.3). View specific Regression Forests are then trained to estimate shape parameters by first projecting features in CCA correlated spaces (Sect. 3.5). **Testing:** The extracted features from an input silhouette are used to first infer the camera view, and then the shape parameters by projecting them into CCA spaces and feeding them into the corresponding Regression Forest.

3.1 Method Overview

The goal of our system is to infer the 3D body shape of a person from a single or multiple monocular images fast and automatically. Specifically, we would like to estimate the parameters of a 3D body shape model (Sect. 3.2) such that the corresponding body shape best approximates the 3D body of the subject depicted in the input images. Despite the ambiguity that the 2D silhouette withholds, the projection of the transformed mesh in the image should at least best explain it.

An overview of our system is depicted in Fig. 1. The input to the shape estimation algorithm is a 2D silhouette of the desired individual under minimal self-occlusion (e.g. a selfie), which can be computed accurately for our target scenarios, by learning a background model through Gaussian mixture models and using Graphcuts [7]. The word "selfie" here is used interchangingly to describe the activity of taking a selfie in front of a mirror, and also as a label for poses representing mild self-occlusion (Fig. 2). We then compute features extracted from the silhouettes (Sect. 3.3). These are first used to train a classifier on the camera viewing direction (Sect. 3.4). The features from silhouettes of a particular view are then projected into bases obtained by CCA, such that the view itself and the most orthogonal one to it (e.g. front and side) are used to capture complementary information into the CCA correlated space, and fed to a Random Forest Regressor (Sect. 3.5) trained for each camera view. At test time, the extracted features from an input silhouette are used to first infer the camera view, and then the shape parameters by projecting them into CCA spaces and feeding them into the corresponding Regression Forest. The parameters are used to generate a mesh by solving a least-squares system on the vertex positions (Sect. 3.2). The generated mesh can then be utilized for various post-processing tasks such as human semantic parameter estimation, free view-point video with projective texturing, further shape refinement [6,56], or pose refinement [25].

3.2 Shape as a Geometric Model

We utilize the SCAPE model [3], which is a low-dimensional parametric model learned from 3D range scans of different people in different poses that captures correlated deformations due to shape and pose changes simultaneously. Specifically, SCAPE is defined as a set of triangle deformations applied to a reference template 3D mesh. Estimating a new shape requires estimating parameters α and β, which determine the deformations due to pose and intrinsic body shape, respectively. Given these parameters, each of the two edges $\mathbf{e}_{i1}$ and $\mathbf{e}_{i2}$ of the i^{th} triangle of the template mesh (defined as the difference vectors between the vertices of the triangle), is deformed according to the following expression

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

with $j \in \{1, 2\}$. The matrices $\mathbf{R}_i(\alpha)$ correspond to joint rotations, and $\mathbf{Q}_i(\mathbf{R}_i(\alpha))$ to the pose induced non-rigid deformations, e.g. muscle bulging. $\mathbf{S}_i(\beta)$ are matrices modeling shape variation as a function of the shape parameters β. The body shape deformation space is learned by applying PCA to a set of meshes of different people in full correspondence and same pose, with transformations written as $\mathbf{s}(\beta) = \mathbf{U}\beta + \mu$, where $\mathbf{s}(\beta)$ is obtained by stacking all transformations $\mathbf{S}_i(\beta)$ for all triangles, $\mathbf{U}$ is a matrix with orthonormal columns, and μ is the mean of the triangle transformations over all meshes (please refer to Anguelov et al. [3] for further details). We therefore obtain the model by computing per-triangle deformations for each mesh of the dataset from a template mesh, which is the mean of all the meshes in the dataset (Fig. 2 left), and then applying PCA in

Fig. 2. 6 meshes from our database. The leftmost one is the mean mesh in the rest pose. The others are from different people in various poses.

order to extract the components capturing largest deformation variations. We chose to use 20 components ($\beta \in R^{20}$).

We would like to estimate the shape parameters β regardless of the pose. We take the common assumption that the body shape does not significantly change due to the range of poses we consider. Hence, we ignore pose dependent shape changes given by $\mathbf{Q}_i(\mathbf{R}(\alpha))$. Decoupling pose and shape changes allows us to adopt a fast and efficient method from the graphics community known as Linear Blend Skinning (LBS) [31] for pose changes, similar to previous works [25,38]. Starting from a rest pose shape with vertices $\mathbf{v}_1, ..., \mathbf{v}_n \in \mathbf{R}^4$ in homogenous coordinates, LBS computes the new position of each vertex by a weighted combination of the bone transformation matrices $\mathbf{T}_1, ..., \mathbf{T}_m$ in a skeleton controlling the mesh, and skinning weights $w_{i,1}, ..., w_{i,m} \in \mathbf{R}$ for each vertex $\mathbf{v}_i$, as given by the following formula:

$$\mathbf{v}_i^{'} = \sum_{j=1}^{m} w_{i,j}\mathbf{T}_j\mathbf{v}_i = \left(\sum_{j=1}^{m} w_{i,j}\mathbf{T}_j\right)\mathbf{v}_i \tag{2}$$

In our model, the skinning weights are computed for a skeleton of 17 body parts (1 for the head, 2 for the torso, 2 for the hips and 3 for each of the lower and upper limbs) for the mean shape mesh using the heat diffusion method [5]. It has to be noted that $w_{i,j} \geq 0$ and $w_{i,1} + \cdots + w_{i,m} = 1$.

3.3 Feature Extraction

We extract novel features from the scaled silhouettes as the input to our learning method. These features are designed to capture local and global information on the silhouette shape, and be robust to pose and slight view changes. For each point in the silhouette, two feature values are calculated, namely the *(weighted) normal depth* and the *curvature*. In order to extract these, we first compute the 2D point normal for every point in the silhouette, and then smooth all normals with a circle filter of radius of 7 pixels. As different people have different silhouette lengths, we sample 1704 equidistant points from each silhouette starting from the topmost pixel of the silhouette. The sample size is set according to the smallest silhouette length over all our training data. Our feature vector per silhouette then consists of 3408 real valued numbers.

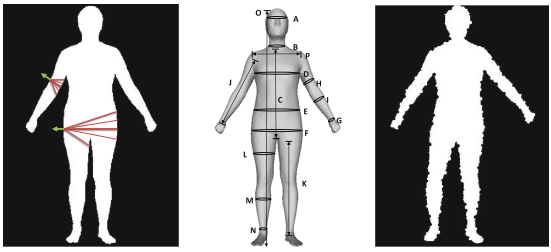

Fig. 3. (left) Normal depth computation in 2 different points. The arrows are the silhouette normals. The normal depth is computed as the weighted mean of the lengths of the red lines. (middle) 3D measurements on the meshes used for validation. (right) Noisy silhouette. (Colour figure online)

The normal depth is computed as follows. For any point from the sampled set, we send several rays starting from the point itself and oriented along the opposite direction of its normal, until they intersect the inner silhouette boundary. The lengths of the ray segments are defined as the normal depths as illustrated in Fig. 3 (left). The normals are represented in green and the ray segments in red for two different points in the silhouette. We allow an angle deviation of 50° from the silhouette normal axis. The feature for a point is defined as the weighted average of all normal depths falling within one std. dev. from the median of all the depths, with weights defined as the inverse of the angle between the rays and the normal axis.

The *normal depth* is a feature inspired by 3D geodesic shape descriptors [44,47], and different from the *Inner-Distance* 2D descriptor [32] used for classification of different object types while being noise sensitive, and the spectral features utilized in [29] for a shape retrieval task. The main ideas behind our feature are (a) for the same individual in different poses, under mild self-occlusions, the features look very similar with small local shifts, (b) each point feature serves as a robust body measurement, correlated with the breadth of the person in various parts of the body, which is analogous to estimating body circumference at each vertex of the real body mesh, and (c) the feature is robust to silhouette noise due to the median and averaging steps. The measure might differ though in some parts of the silhouette (e.g. elbow) for the same person in different poses. In order to alleviate this limitation, we apply smoothing on small neighborhoods of the silhouette. The *curvature* on the other hand is estimated as the local variance of the normals. Despite being a local feature, it provides a measure of roundness, especially around the hips, waist, belly and chest, which helps in discriminating between various shapes.

We illustrate that the combination of normal depth that captures global information on the silhouette and curvature encoding local details leads to estimators robust to limited self-occlusions, and discriminative enough to describe the silhouette and reconstruct the corresponding shape in Sect. 4.

3.4 View Direction Classification

To increase robustness with respect to view changes, we decided to train view-specific Regression Forests for 36 viewing directions around the body. In order to discriminate between the views, we train a Random Forest Classifier utilizing the 3408 features extracted (Sect. 3.3) from 100,000 silhouettes of people in multiple poses, shapes and views, having as labels the views numbered 1 to 36. We achieve a high accuracy of 99 % if we train and test on neutral and selfie-like poses. The accuracy decreases to 85.7 % if more involved poses (e.g. walking, running etc.) are added. However, by investigating class prediction probabilities, we observed that false positives are assigned only to the views that are contiguous to the view with the correct label. As it will be shown in Sect. 4, Table 2, a 10° view difference has a low reconstruction error when the features are projected into CCA bases.

3.5 Learning Shape Parameters

We pose shape parameter estimation as a regression task. Given the silhouette features, using supervised learning, we would like to estimate the shape parameters such that the reconstructed shape best explains the silhouette. To make the features more discriminative, we propose to correlate features extracted from silhouettes viewed from different directions. More specifically, we apply Canonical Correlation Analysis (CCA) [24] over features extracted from a pair of silhouettes from two camera views.

At training time, the views are selected such that they capture complementary information. While the first one is the desired view from which we want to estimate the shape (one of 36 views), the second one is chosen to be as orthogonal as possible to the first, e.g. (front and side view). Because the human body is symmetric, a complementary view to a desired one is always searched in the zero to 90° angle range to that view. In practice, we round the complementary view to the closest extreme (i.e. front or side view) to ease the offline computations.

We first apply PCA to reduce the dimensionality of the extracted features from 3408 to 300 in each view. Then, we stack the PCA projected features for all mesh silhouettes from the first and second views into the columns of the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. Then, CCA attempts to find basis vector pairs $\mathbf{b}_1$ and $\mathbf{b}_2$, such that the correlations between the projections of the variables onto these vectors are mutually maximized by solving:

$$\underset{\mathbf{b}_1, \mathbf{b}_2 \in R^N}{\arg\max}\, corr(\mathbf{b}_1^T \mathbf{X}_1, \mathbf{b}_2^T \mathbf{X}_2), \tag{3}$$

where $N = 300$. This results in a coordinate free mutual basis unaffected by rotation, translation or global scaling of the features. The features projected onto this basis thus capture mutual information coming from both views. The subsequent basis vector pairs are computed similarly, with the assumption that the new projected features are orthogonal to the existing projected ones. We use 200 basis pairs with CCA projections covering 99 % of the energy.

The final training is done on the 200 projected features extracted from one view, which is one of the 36 views we consider. These projected features are input to a Random Forest Regressor [8] of 4 trees and a maximum depth of 20. The labels for this regressor are the 20-dimensional shape parameter vectors β. Each component of β is weighted with weights set to the eigenvalues of the covariance matrix defined in Sect. 3.2 in the computation of the shape deformation space, and normalized to 1, to emphasize the large scale changes in 3D body shapes. At test time, the raw features extracted from a single given silhouette are first classified into a view. These are then projected with the obtained PCA and CCA matrices for that view to obtain a 200 dimensional vector. The projected features are finally fed into the corresponding Random Forest Regressor, in order to obtain the desired shape parameters β.

4 Validation and Results

Previous shape-from-silhouette methods lack extensive evaluation. Xi et al. [52] demonstrate results on two real images of people and 24 subjects in synthetic settings, Sigal et al. [46] validate on two measurements and two subjects in monocular settings, and Balan et al. [4] report silhouette errors for a few individuals in a sequence and height measurement for a single individual. To the best of our knowledge, only Boisvert et al. [6] perform a more extensive validation, for 220 synthetic humans consisting of scans from the CAESAR database [39], and four real individuals' front and side images. We present the largest validation experiment with 1500 synthetic body meshes as well as real individuals.

Data Generation. In order to learn a general model, we merge two large datasets [38,54] consisting of 3D models extracted from the commercially available CAESAR dataset[1]. We select 2900 meshes from the combined dataset for learning the shape model, leaving out around 1500 meshes for testing and experiments. In order to synthesize more training meshes, we sample from the 20 dimensional multivariate normal distribution spanned by the PCA space (Sect. 3.2), such that for a random sample $\beta = [\beta_1, \beta_2, ..., \beta_{20}]$, it holds that $\beta \sim \mathcal{N}(\mu, \Sigma)$ with μ being the 20-dimensional mean vector and Σ the 20×20 covariance matrix of the parameters. To synthesize meshes in different poses, we gather a set of animations comprising of various poses (e.g. selfie, walking, running, etc.). After transferring a generated pose to the template mesh using LBS, we compute the resulting per-triangle deformations $\mathbf{R}_i$. For a given mesh with parameters β, the final pose is then given by $\mathbf{e}'_{ij} = \mathbf{R}_i \mathbf{S}_i(\beta)\mathbf{e}_{ij}$, where $\mathbf{e}_{ij}$ are the edges of the template mesh (Sect. 3.2).

As the training set, we randomly generate 100000 samples from the multivariate distribution over the β parameters, and restrict them to fall into the $\pm 3 \times Std.Dev$ range for each dimension of the PCA projected parameters to avoid getting unrealistic human shapes. We project the generated meshes in each of the 36 camera viewpoints around the mesh (Sect. 3.4). The silhouette is

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

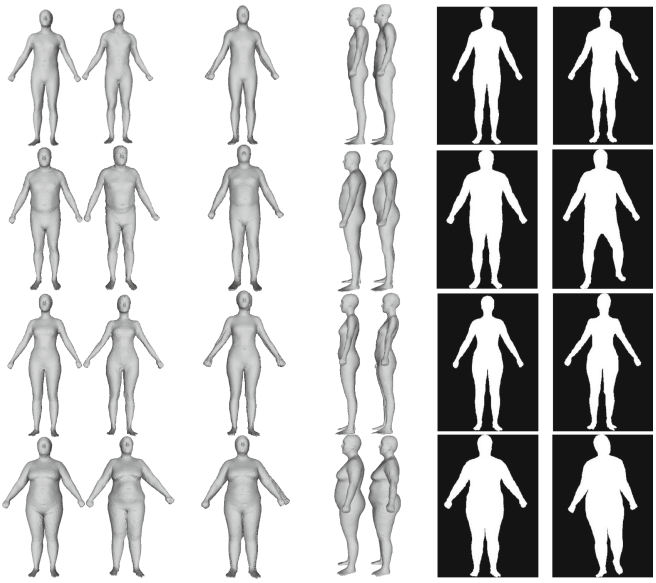

Fig. 4. Visual results for predictions on 4 test meshes. From left to right: predicted mesh, ground truth mesh, the two meshes frontally overlapping, the two meshes from the side view, silhouette from the predicted mesh, input silhouette.

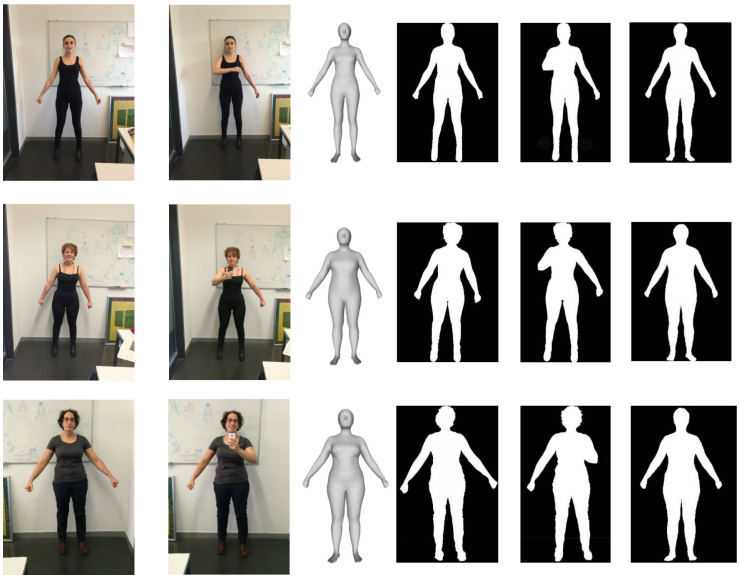

Fig. 5. Visual results for predictions on 3 females. From left to right: the two input images in a rest and selfie pose, the estimated mesh - same estimation is obtained for both poses, the two silhouettes from which features are extracted for each pose, the silhouette of the estimated mesh.

Table 1. Comparisons to state-of-the-art methods, variations of our method (*RF*, *CCA-RF-1*, *CCA-RF-2*) and ground truth, via various measurements. The measurements are illustrated in Fig. 3 (middle). Errors are represented as Mean ± Std. Dev and are expressed in millimeters. Note that we operate under a significantly more general setting than the state-of-the-art methods, please refer to the text.

Measurement	[6]	[13]	[52]	RF	CCA-RF-1	CCA-RF-2	GT
A. Head circumference	10 ± 12	23 ± 27	50 ± 60	16 ± 13	13 ± 10	$\mathbf{8 \pm 8}$	13 ± 9
B. Neck circumference	11 ± 13	27 ± 34	59 ± 72	13 ± 10	10 ± 8	$\mathbf{7 \pm 7}$	6 ± 6
C. Shoulder-blade/crotch length	4 ± 5	52 ± 65	119 ± 150	22 ± 18	18 ± 9	$\mathbf{18 \pm 17}$	14 ± 11
D. Chest circumference	10 ± 12	18 ± 22	36 ± 45	38 ± 31	30 ± 24	$\mathbf{25 \pm 24}$	24 ± 24
E. Waist circumference	22 ± 23	37 ± 39	55 ± 62	35 ± 28	29 ± 25	$\mathbf{24 \pm 24}$	16 ± 14
F. Pelvis circumference	11 ± 12	15 ± 19	23 ± 28	33 ± 26	30 ± 25	$\mathbf{26 \pm 25}$	14 ± 12
G. Wrist circumference	9 ± 12	24 ± 30	56 ± 70	10 ± 8	6 ± 5	$\mathbf{5 \pm 5}$	5 ± 5
H. Bicep circumference	17 ± 22	59 ± 76	146 ± 177	16 ± 13	13 ± 11	$\mathbf{11 \pm 11}$	9 ± 10
I. Forearm circumference	16 ± 20	76 ± 100	182 ± 230	14 ± 11	11 ± 9	$\mathbf{9 \pm 8}$	8 ± 8
J. Arm length	15 ± 21	53 ± 73	109 ± 141	19 ± 14	15 ± 12	$\mathbf{13 \pm 12}$	8 ± 8
K. Inside leg length	6 ± 7	9 ± 12	19 ± 24	26 ± 19	23 ± 18	$\mathbf{20 \pm 19}$	9 ± 9
L. Thigh circumference	9 ± 12	19 ± 25	35 ± 44	22 ± 18	19 ± 16	$\mathbf{18 \pm 17}$	11 ± 11
M. Calf circumference	6 ± 7	16 ± 21	33 ± 42	18 ± 13	14 ± 12	$\mathbf{12 \pm 12}$	7 ± 8
N. Ankle circumference	14 ± 16	28 ± 35	61 ± 78	10 ± 7	18 ± 6	$\mathbf{6 \pm 6}$	5 ± 5
O. Overall height	9 ± 12	21 ± 27	49 ± 62	60 ± 45	50 ± 42	$\mathbf{43 \pm 41}$	14 ± 11
P. Shoulder breadth	6 ± 7	12 ± 15	24 ± 31	15 ± 14	13 ± 6	$\mathbf{6 \pm 6}$	12 ± 11

computed by projecting all the mesh edges for which two coinciding triangles have normals pointing in opposite directions. The silhouettes are then uniformly scaled such that the height of the bounding box is equal to 528 pixels, and the width to 384 pixels. For testing, we evaluate our method with the meshes left out from the training dataset, as well as on real images.

Quantitative Experiments. We distinguish two test datasets, D1 and D2. D1 consists of 1500 meshes neither used to learn the parametric shape model nor to train the regression forests (RF) and D2 of 1000 meshes used to learn the parametric model but not to train the RF. These meshes consist of 50 % males and 50 % females, and are in roughly the same rest pose. In order to properly quantify our method, similar to Boisvert et al. [6], we perform 16 three-dimensional measurements on the meshes, which are commonly used in garment fitting as illustrated in Fig. 3 (middle). For the measurements represented with straight lines, we compute the Euclidean distance between the two extreme vertices. The ellipses represent circumferences and are measured on the body surface. For each of the 16 measurements, we compute the difference between the one from the ground truth mesh and the estimated mesh. We report the mean error and the standard deviation for each of the measurements in Table 1. We name our main method *CCA-RF*, with CCA applied to the features before passing them to the random forest, specifically *CCA-RF-1* and *CCA-RF-2* respectively tested on D1 and D2. Similarly, *RF*, for the method trained on raw features and tested on D1. The last table column provides the ground truth (GT) mean errors for D1, computed between the original test meshes and their reconstructions obtained

by projecting them into the learned PCA space. This provides a lower limit for the obtainable errors with our 20 parameters shape model.

Before analyzing the results, it is crucial to highlight the differences between the settings and goals of the methods we compare to. Boisvert et al. [6] employ a setting where the pose is fixed to a rest pose and the distance from the camera is also fixed. The shape estimation method is based on utilizing silhouettes from two different views (front and side), with the application of garment fitting in mind. The same setting is considered for the other two methods mentioned above [13,52]. In contrast, we train and test for a more general setting, where we have a single silhouette as the input at test time, the pose can change, and no assumptions on the distance from the camera are made. Furthermore, our tests involve a significantly larger dataset with high variations.

Even though our method operates under a significantly more general setting than the previous works [6,13,52], with a single silhouette input and no distance information, it outperforms the non-linear and linear mapping methods. The mean absolute error for all the models is 19.4 mm for *CCA-RF-1* and 16.18 mm for *CCA-RF-2*. The errors are very close to those of GT, illustrating the accuracy of our technique. Note that some errors for *CCA-RF-2* are smaller than those of the GT, due to the different training as explained above. The higher error for D1 is due to the body shapes that cannot be represented with the parametric model learned from the rest of the shapes. The error is higher for the overall height, due to the fixed scale in the training and testing silhouettes that we use. It is important to note the differences in errors between the *RF* and *CCA-RF-1*. There is an overall decrease of error when CCA is utilized, which shows that the projection with the CCA bases significantly improves prediction. Additionally, we evaluate the performance of our method when the input comes from a less favorable view, the side view, achieving an error of 22.45 mm which is very close to the one from only the frontal view. For completeness, we compare also to Helten et al. [23], who utilize an RGB-D camera for capturing the body shapes, and a full RMSE map per vertex to measure the differences. Using two depth maps, fitting to the pose and testing only on 6 individuals they report a mean error of 10.1 mm while we have a mean error of 19.19 mm on 1500 meshes.

Poses, Views and Noise. We investigated accuracy in the presence of silhouette noise, various poses, and different or multiple views. We run the experiments with the data setup D1, explained above. For each experiment, we show the mean and standard deviation either of the accuracy gain or of the errors over all the body measurements in Table 2.

The first three columns show the *accuracy gain* of applying *CCA-RF* to the front view (F), side view (S) or when concatenating both views together (FS), as compared to *RF*. A larger gain is obtained in the side view as compared to the front view, due to additional information that is injected from the frontal view (the most representative one) in the projected space. An even bigger gain is obtained if both views are utilized for training and testing. This is very important, as it shows that having potentially more views improves the predictor. In fact, we have observed that utilizing the same amount (100000) of training

Table 2. Columns 1–3 show accuracy gain of applying CCA for the Frontal, Side and Frontal Side view altogether, over raw features. (VE) shows the error due to 10° view change and (VG), the gain of applying CCA. (N) is the error due to silhouette noise. (P12) shows the error of testing on 12 poses different from the training one, and the rest (Columns 8–11) demonstrate the errors while gradually adding more difficult poses from the training ones. Mean and Std. Deviation is computed over all the body measurements.

Measurement	(F)	(S)	(FS)	(VE)	(VG)	(N)	(P12)	(P1)	(W)	(R)	(PWR)
Mean (mm)	4.9	5.2	6.6	2.2	1.8	2.3	9.3	1.7	1.6	3.9	8.5
Std. Deviation (mm)	2.4	2.6	4.0	1.9	1.5	1.8	5.6	1.0	1.0	2.3	5.2

data, and training and testing on two views with the raw features, degrades the result as compared to just one view. This is alleviated with the CCA projection, improving the results as singular view noise in the data is removed.

The fourth column (VE), displays the *errors* obtained by testing on features extracted from a view 10° rotated from the frontal view, for a *CCA-RF* trained on the frontal view. The column for (VG) displays the *gain* of *CCA-RF* over *RF* for the same scenario. The CCA-RF is again more accurate, however the error for both is generally low, implying that a classification error of the camera view of 10° can be allowed in our system. (N) demonstrates the *error* due to random noise added to the silhouettes, as in Fig. 3 (right), showing robustness to noise to a certain extent. (P12) shows the *error* induced by training only on a rest pose, and testing on 12 different poses as in Fig. 2, as compared to testing on the same meshes in a rest pose, and (P1) describes the same measurement, however by training on 12 poses and testing on a different unseen one, demonstrating robustness to pose changes under minimal self occlusions. The last three columns demonstrate similar measurements, however, by increasing the articulations in the poses, with (W) consisting of poses from a walking sequence, (R) from a running sequence (supplementary [1]), and (PWR) combining all poses we have. The error increases in the latter case especially due to the introduction of poses with more self occlusions. However, when trained on individual sequences, the errors are lower, implying that for an application where a certain activity is known, one could adapt specialized regressors, especially due to the very fast training in the low dimensional spaces.

Algorithm Speed. The method is significantly faster than previous works, allowing for interactive applications. The method of Boisvert et al. [6] needs 6 s for body shape regression, 30 s for the MAP estimation, and 3 min for the silhouette based similarity optimization, with 6 s for their implementation of sGPLVM [13] (on an Intel Core i7 CPU 3 GHz and single-threaded implementation). We, on the other hand, reach 0.3 s using a single threaded implementation on an Intel Core i7 CPU 3.4 GHz (0.045 s for feature computation, 0.25 s for mesh computation, and 0.005 s for random forest regression), with even more speed-up opportunities as the feature computation and mesh vertices computation can be highly parallelized.

Qualitative Results. In Fig. 4, we show example samples from our tests. In each row, first the predicted mesh is shown along with the ground truth test mesh. Then, their overlap is illustrated. This is followed by the side views, and the silhouette of the estimated mesh and the input silhouette. Note that the input silhouettes are in different poses, but we show the estimated meshes in rest poses for easy comparisons. Our results are visually very close to the ground truth shapes even under such pose changes.

Finally, we show an experiment where real pictures of three females are taken in a rest and a selfie-like pose along with the estimated meshes in Fig. 5. It is important to note that despite the pose change, the retrieved mesh for each person is the same. Another important observation is that even though the input is scaled to the same size, the estimated parameters yield statistically plausible heights, which turned out to be sufficient in obtaining an ordering based on relative height between the estimated meshes. We believe that this is due to the statistical shape model, where semantic parameters like height and weight are correlated in the PCA parameter space. To the best of our knowledge, no previous work can resolve this task. For example, in the work by Sigal et al. [46], the mesh needs to be scaled if no camera calibration is provided.

5 Discussion and Conclusions

In this paper, we presented a novel technique that estimates 3D human body shape from a single silhouette. It allows different views, poses with mild occlusions, and various body shapes to be estimated. We extensively evaluated our technique on thousands of human bodies, by utilizing one of the biggest databases available to the community.

In the scope of this paper, we focused on shape extraction from a single silhouette because of its various applications such as selfies or utilizing limited video footage. However, this is an inherently ill-posed problem. Further views can be incorporated to obtain more accurate reconstructions, similar to methods we compare to. This would lead to a better estimation especially in the areas around the belly and chest, hence decrease the elliptical body measurement errors.

The accuracy of our method is tied to silhouette extraction. For the difficult cases of dynamic backgrounds or very loose clothes, the large scale silhouette deformations would skew our results. This could be tackled by fusing results over multiple frames. Unlike [13] though, our results always remain in the space of plausible human bodies. For small scale deformations (Fig. 3 right), we show in Table 2 (N) that our results stay robust.

We assume that the silhouettes come in poses with limited partial occlusion. Under this assumption, we showed robustness, the same mesh estimation is achieved from different poses (e.g. Fig. 5). However, under more pronounced occlusions, our results start degrading (Table 2 (PWR)), which could be alleviated by increasing the number of training poses and utilizing deeper learning.

Although we aimed at precise measurements for the evaluation, errors due to discretization are inevitable, hence a standardized procedure on a standard

mesh dataset is needed as a benchmark. We believe that this work along with that of Boisvert et al. [6] has set an important step towards this direction.

Since our system is designed for a general setting, we apply a fixed scale to the silhouette, losing height information. We showed a fairly good performance on estimating the relative height and demonstrate better absolute height estimation, if camera calibration is incorporated (supplementary [1]).

Our fast system, running in minutes for training and milliseconds for execution in single core CPU's, while being memory lightweight due to the low feature dimensionality, could be integrated into smart phones, allowing body shapes to be reconstructed with one click of a button. Simultaneously, it can be used for 3D sport analysis, where estimation of a 3D shape of a player seen from a sparse set of cameras can improve projections of novel-views.

Finally, we showed how CCA, which captures relations in an unsupervised linear way, can be used to correlate different views in the data to improve the prediction power and speed of the algorithm. We believe that capturing non-linear relations with Kernel CCA's or deep architectures should lead to even better results. Our method illustrates the utility of CCA for other vision applications where two or more views describing the same object or event exists, such as multi-view pose estimation, video-to-text matching, or shape from various sources of information.

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

References

1. https://cgl.ethz.ch/publications/papers/paperDib16a.php
2. de Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. In: SIGGRAPH (2008)
3. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: Shape completion and animation of people. In: SIGGRAPH (2005)
4. Balan, A.O., Sigal, L., Black, M.J., Davis, J.E., Haussecker, H.W.: Detailed human shape and pose from images. In: CVPR (2007)
5. Baran, I., Popovic, J.: Automatic rigging and animation of 3d characters. ACM Trans. Graph. **26**, 1–8 (2007)
6. Boisvert, J., Shu, C., Wuhrer, S., Xi, P.: Three-dimensional human shape inference from silhouettes: reconstruction and validation. Mach. Vis. Appl. **24**, 145–157 (2013)
7. Boykov, Y., Jolly, M.: Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In: ICCV (2001)
8. Breiman, L.: Random forests. Mach. Learn. **26**, 123–140 (2001)
9. Bălan, A.O., Black, M.J.: The naked truth: estimating body shape under clothing. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5303, pp. 15–29. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88688-4_2
10. Casas, D., Volino, M., Collomosse, J., Hilton, A.: 4d video textures for interactive character appearance. Comp. Graph. Forum(Proc. Eurographics) **33**, 371–380 (2014)
11. Chen, X., Guo, Y., Zhou, B., Zhao, Q.: Deformable model for estimating clothed and naked human shapes from a single image. Vis. Comput. **29**, 1187–1196 (2013)

12. Chen, Y., Cipolla, R.: Learning shape priors for single view reconstruction. In: ICCV Workshops (2009)
13. Chen, Y., Kim, T.-K., Cipolla, R.: Inferring 3D shapes and deformations from single views. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6313, pp. 300–313. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15558-1_22
14. Chen, Y., Kim, T., Cipolla, R.: Silhouette-based object phenotype recognition using 3d shape priors. In: ICCV (2011)
15. Delamarre, Q., Faugeras, O.: 3d articulated models and multi-view tracking with silhouettes. In: ICCV (1999)
16. Guan, L., Franco, J., Pollefeys, M.: Multi-object shape estimation and tracking from silhouette cues. In: CVPR (2008)
17. Guan, P., Reiss, L., Hirshberg, D.A., Weiss, A., Black, M.J.: Drape: dressing any person. ACM Trans. Graph **31**, 1–10 (2012)
18. Guan, P., Weiss, A., Balan, A.O., Black, M.J.: Estimating human shape and pose from a single image. In: ICCV (2009)
19. Hardoon, D.R., Mourão Miranda, J., Brammer, M., Shawe-Taylor, J.: Unsupervised analysis of fmri data using kernel canonical correlation. NeuroImage **37**, 1250–1259 (2007)
20. Hardoon, D.R., Szedmak, S.R., Shawe-taylor, J.R.: Canonical correlation analysis: an overview with application to learning methods. Neural Comput. **16**, 2639–2664 (2004)
21. Hasler, N., Ackermann, H., Rosenhahn, B., Thormählen, T., Seidel, H.: Multilinear pose and body shape estimation of dressed subjects from image sets. In: CVPR (2010)
22. Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., Seidel, H.: A statistical model of human pose and body shape. Comput. Graph. Forum **28**, 337–246 (2009)
23. Helten, T., Baak, A., Bharaj, G., Müller, M., Seidel, H., Theobalt, C.: Personalization and evaluation of a real-time depth-based full body tracker. In: 3DV (2013)
24. Hotelling, H.: Relations between two sets of variates. Biometrika **28**, 321–377 (1936)
25. Jain, A., Thormählen, T., Seidel, H.-P., Theobalt, C.: MovieReshape: tracking and reshaping of humans in videos. ACM Trans. Graph. **29**(6), 148:1–148:10 (2010). doi:10.1145/1882261.1866174
26. Kakade, S.M., Foster, D.P.: Multi-view regression via canonical correlation analysis. In: Bshouty, N.H., Gentile, C. (eds.) COLT 2007. Lecture Notes in Artificial Intelligence (LNAI), vol. 4539, pp. 82–96. Springer, Heidelberg (2007). doi:10.1007/978-3-540-72927-3_8
27. Kakadiaris, I.A., Metaxas, D.: Three-dimensional human body model acquisition from multiple views. IJCV **30**, 191–218 (1998)
28. Kim, T.K., Wong, S.F., Cipolla, R.: Tensor canonical correlation analysis for action classification. In: CVPR (2007)
29. Lahner, Z., Rodola, E., Schmidt, F.R., Bronstein, M.M., Cremers, D.: Efficient globally optimal 2d-to-3d deformable shape matching. In: CVPR (2016)
30. Laurentini, A.: The visual hull concept for silhouette-based image understanding. PAMI **16**, 150–162 (1994)
31. Lewis, J.P., Cordner, M., Fong, N.: Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In: SIGGRAPH (2000)
32. Ling, H., Jacobs, D.W.: Shape classification using the inner-distance. PAMI **29**, 286–299 (2007)

33. McWilliams, B., Balduzzi, D., Buhmann, J.M.: Correlated random features for fast semi-supervised learning. In: NIPS (2013)
34. Mikic, I., Trivedi, M., Hunter, E., Cosman, P.: Human body model acquisition and tracking using voxel data. IJCV **53**, 199–223 (2003)
35. Neophytou, A., Hilton, A.: Shape and pose space deformation for subject specific animation. In: 3DV (2013)
36. Neophytou, A., Hilton, A.: A layered model of human body and garment deformation. In: 3DV (2014)
37. Perbet, F., Johnson, S., Pham, M.T., Stenger, B.: Human body shape estimation using a multi-resolution manifold forest. In: CVPR (2014)
38. Pishchulin, L., Wuhrer, S., Helten, T., Theobalt, C., Schiele, B.: Building statistical shape spaces for 3d human modeling. CoRR (2015)
39. Robinette, K.M., Daanen, H.A.M.: The caesar project: a 3-d surface anthropometry survey. In: 3DIM (1999)
40. Rogge, L., Klose, F., Stengel, M., Eisemann, M., Magnor, M.: Garment replacement in monocular video sequences. ACM Trans. Graph. **34**, 1–10 (2014)
41. Sargin, M.E., Yemez, Y., Erzin, E., Tekalp, A.M.: Audiovisual synchronization and fusion using canonical correlation analysis. Trans. Multimedia **9**, 1396–1403 (2007)
42. Schmidt, F.R., Farin, D., Cremers, D.: Fast matching of planar shapes in sub-cubic runtime. In: ICCV (2007)
43. Schmidt, F.R., Töppe, E., Cremers, D.: Efficient planar graph cuts with applications in computer vision. In: CVPR (2009)
44. Shapira, L., Shamir, A., Cohen-Or, D.: Consistent mesh partitioning and skeletonisation using the shape diameter function. Visual Comput. **24**, 249–259 (2008)
45. Sharma, A., Kumar, A., Daume III, H., Jacobs, D.W.: Generalized multiview analysis: a discriminative latent space. In: CVPR (2012)
46. Sigal, L., Balan, A.O., Black, M.J.: Combined discriminative and generative articulated pose and non-rigid shape estimation. In: NIPS (2007)
47. Slama, R., Wannous, H., Daoudi, M.: Extremal human curves: a new human body shape and pose descriptor. In: FG (2013)
48. Starck, J., Miller, G., Hilton, A.: Video-based character animation. In: ACM SIGGRAPH Eurographics SCA (2005)
49. Stoll, C., Gall, J., de Aguiar, E., Thrun, S., Theobalt, C.: Video-based reconstruction of animatable human characters. In: SIGGRAPH Asia (2010)
50. Weiss, A., Hirshberg, D.A., Black, M.J.: Home 3d body scans from noisy image and range data. In: ICCV (2011)
51. Wuhrer, S., Pishchulin, L., Brunton, A., Shu, C., Lang, J.: Estimation of human body shape and posture under clothing. CVIU **127**, 31–42 (2014)
52. Xi, P., Lee, W., Shu, C.: A data-driven approach to human-body cloning using a segmented body database. In: Pacific Graphics (2007)
53. Xu, F., Liu, Y., Stoll, C., Tompkin, J., Bharaj, G., Dai, Q., Seidel, H.P., Kautz, J., Theobalt, C.: Video-based characters: Creating new human performances from a multi-view video database. In: SIGGRAPH (2011)
54. Yang, Y., Yu, Y., Zhou, Y., Du, S., Davis, J., Yang, R.: Semantic parametric reshaping of human body models. In: 3DV (2014)
55. Ye, M., Yang, R.: Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In: CVPR (2014)
56. Zhou, S., Fu, H., Liu, L., Cohen-Or, D., Han, X.: Parametric reshaping of human bodies in images. ACM Trans. Graph. **29**(4), 126:1–126:10 (2010). doi:10.1145/1778765.1778863