Pacific Graphics 2016
E. Grinspun, B. Bickel, and Y. Dobashi
(Guest Editors)

Volume 35 (2016), Number 7

3D Body Shapes Estimation from Dressed-Human Silhouettes

1008

Abstract

Estimation of 3D body shapes from dressed-human photos is an important but challenging problem in virtual fitting. We propose a novel automatic framework to efficiently estimate 3D body shapes under clothes. We construct a large database of 3D naked and dressed body pairs, based on which we learn how to predict 3D positions of body landmarks (which further constrain a parametric human body model) automatically according to dressed-human silhouettes. Critical vertices are selected on 3D registered human bodies as landmarks to represent body shapes, so as to avoid the time-consuming vertices correspondences finding process for parametric body reconstruction. Our method can estimate 3D body shapes from dressed-human silhouettes within 4 seconds, while the fastest method reported previously need 1 minute. In addition, our estimation error is within the size tolerance for clothing industry. We dress 3D naked bodies with one set of common clothes acquired by physically based cloth simulation technique. To the best of our knowledge, We are the first to construct a large database containing 3D naked and dressed body pairs and our database may contribute to the areas of human body shapes estimation and cloth simulation.

Categories and Subject Descriptors (according to ACM CCS): I.3.m [Computer Graphics]: Miscellaneous—Image-based modeling

1. Introduction

Virtual fitting systems provide valuable visual information and size suggestions for users to buy clothes online, offering unique immersive experience. The main difficulties of realistic virtual fitting systems are the efficiency of physically based cloth simulation, the acquisition of 3D customized human model and the determination of physical property of cloth. Virtual fitting systems need to extract relatively accurate 3D body shapes of customers, especially in specific size aspects required for fashion design (e.g., chest/waist/hip size and body length).

3D human body estimation is a hot topic in computer graphics, as it plays an important role in applications such as movies, computer games and virtual fitting. A variety of methods have been proposed to estimate body shapes and they can be classified into non-parametric methods and parametric methods [CTT*].

Non-parametric methods take 3D points as input by scanning a human body in several views. Then a body mesh is acquired through registering or merging scan views and hole filling steps. This kind of method relies on the accuracy of scanners and restricts human to keep still with minimal clothes when scanning. For ordinary online shopping customers, they absolutely cannot bear these restrictions.

Parametric methods deform a 3D template body with a series of parameters. Input information is used as constraints to work out parameters and such information can be 3D or 2D that is more accessible. At early stage, parametric methods reconstructed 3D bodies using minimally-dressed information. Nowadays, more and more researchers utilize parametric human body model to estimate 3D body shapes from dressed-human information, the input of which is more convenient for users to get.

Clothes occlude human bodies and make body shapes estimation challenging. A body space learnt from a database of 3D naked bodies can be used to alleviate clothes effects [HSR*09, WPB*14, NH14]. Due to projecting dressed shapes to the body space, the estimated body is usually fatter than ground truth. Skin areas show more confidence for shapes so that some researchers set higher weight for the exposed-skin part of input information [BB08, ZCD*15]. Consequently, this kind of method relies on skin areas and still leaves the covered body shapes ambiguous. Zhu et. al. [ZM15] allowed users to interactively estimate some body points under clothes, which depended on users' experience. Chen et. al. [CGZZ13] made the first attempt to model clothes deformation and proposed a parametric dressed body model. Compared with physically based cloth simulation, their clothes is less realistic.

Physically based cloth simulation has been researched for many years with explicit modules (e.g., cloth modeling, numerical time integration and collision handling). Given a 3D naked body and corresponding clothes, some commercial cloth simulation softwares can composite a realistic 3D dressed model and fossilization. However, it is extremely difficult and complex for the inverse operation, i.e., recovering the naked body from its dressed shape.

To analyze the relationship between a 3D naked body and its dressed shape, we construct a large database containing 1718 pairs of 3D naked and dressed bodies. Because of the labor intensity of real shapes acquisition, we synthesize 1718 male bodies with a standard standing pose using 56 real male bodies from

MPI database [HSS*09] and dress them with a long-sleeved shirt and long pants using mature cloth simulation technique. Ideally, we aim to acquire same accuracy for "undressing estimation" as "dressing simulation" with the help of our database. To our knowledge, this is the first large database that contains 3D naked and dressed human body pairs, which may benefit the areas of cloth simulation and human body shapes estimation.

Based on our database, we create training samples containing dressed-human silhouettes, initial 3D landmarks and target 3D landmarks. An effective feature descriptor is proposed to combine 3D naked body landmarks with dressed-human silhouettes, and regressors are trained for guiding landmarks movements (from initial landmarks to target landmarks) according to dressed-human silhouettes with training samples. In testing phase, given dressed-human silhouettes and a set of initial landmarks as input, we regress target 3D naked body with training results as guidance. The regressed landmarks are used to constrain SCAPE model [ASK*05] for body reconstruction.

Our work provide a tangible solution for ordinary users to access their 3D body data with a common set of clothes holding a standard standing pose. The main application of our work is virtual fitting. Users can reconstruct their own 3D bodies by our method with photos so that they could receive size recommendations and visualize their 3D views with new clothes in virtual fitting room. There are many other potential applications. Take computer games for example, it is exciting to create a virtual customized character with similar shape of the player in real world. It is also possible to combine our method with 3D printing for customized human toys and sculptures.

Our main contributions are summarized as: (a) an automatic framework for efficient body shapes estimation from dressed-human silhouettes, (b) the first large database containing 3D naked and dressed body pairs and (c) an effective feature descriptor combining 3D naked body landmarks with 2D dressed-human silhouettes.

2. Related Work

2.1. 3D Human Body Reconstruction Methods

3D human body reconstruction methods can be classified to non-parametric methods and parametric methods [CTT*]. A variety of works [TCL*13] [LVG*13] [TZL*12] use non-parametric methods to obtain a 3D mesh which is close to scanned points cloud. If we use non-parametric methods to estimate body shapes, we rely on 3D scanners and restrict human to keep still with minimal clothes when scanning, which is far away from our input requirement of using dressed-human photos.

Many parametric methods have been proposed for 3D human body reconstruction. Allen et. al. [ACP03] came up with a statistical model to learn a shape space for a similar pose. Similar ideas were used for human body shapes estimation from one or more images [SYW06] [CC09] [BSWX13] and body measurements [WS13]. To allow pose variation, Anguelov et. al. [ASK*05] proposed SCAPE model which considered body deformation as the combination of pose deformation and shape deformation. SCAPE model successfully models human body variation and attracts lots of researchers.

SCAPE model was used to estimate body shapes from a single image or painting of minimally-dressed people [GWBB09]. Balan et. al. [BB08] adopted it to estimate body shapes with 4 images of normally-dressed people from different views. Such estimated mesh can be utilized to modify the input image [ZFL*10] or video [JTST10]. Weiss et. al. [WHB11] adopted it to estimate human body with noisy Kinect data. We use SCAPE model for our body reconstruction with several 3D landmarks.

2.2. 3D Body Shapes Estimation Under Clothes

Hasler et. al. [HSR*09], Wuhrer et. al. [WPB*14] and Neophytou et. al. [NH14] regarded clothes as noises. They trained their models with databases of minimally-dressed bodies to learn human body spaces which did not contain clothes. They used dressed-human information (3D dense points) as constraints to deform a template mesh and got a coarse body mesh which was affected by clothes. Then the coarse mesh was represented in the learnt body space to alleviate clothes effects. These methods work better for tight clothes and the reconstructed bodies are usually fatter than ground truth. With RGBD data acquired by Kinect, Zeng et. al. [ZCD*15] used the RGB image to detect skin areas as tight constraints for body shapes estimation. These methods should find correspondences between input 3D dense points and target mesh vertices, which is time-consuming.

Compared with 3D information obtained by scanners, images are more accessible. Balan et. al. [BB08] took dressed-human silhouettes from 4 views as input. They detected exposed-skin parts to decide weights for input information and then constrained SCAPE model with input information. The reconstruction energy was represented as pixels differences between input silhouettes and projections of target mesh in 4 views. This representation and computation are complex so that they used a gradient-free direct search simplex method to optimize the energy. Fitting takes approximately 40 minutes for a single model.

Chen et. al. [CGZZ13] made the first attempt to consider clothes and they extended SCAPE model to a dressed human shape model. They used deformation transfer [SP04] technique to construct a database of naked and dressed body pairs. For each type of clothes, only one naked and dressed body pair in database was generated by a animation software (POSER) with high quality while the other points were obtained by deformation transfer. They assumed that clothes deformation was only related to body shape deformation for a specific clothes type and learnt clothes-related coefficients for the model additionally. This also leads to the non-linear optimization problem which costs some time. They found the correspondence between 3D dressed body vertices and 2D dressed body contour points with a HMM method [KSvdP09], the computation of which is not fast.

Zhu et. al. [ZM15] predicted body shapes under clothes using orthogonal-view dressed-human photos. They allowed users to interactively estimate some body points for photos and then searched naked body silhouettes in a large database according to those points. Unlike previous work, they did not directly build the

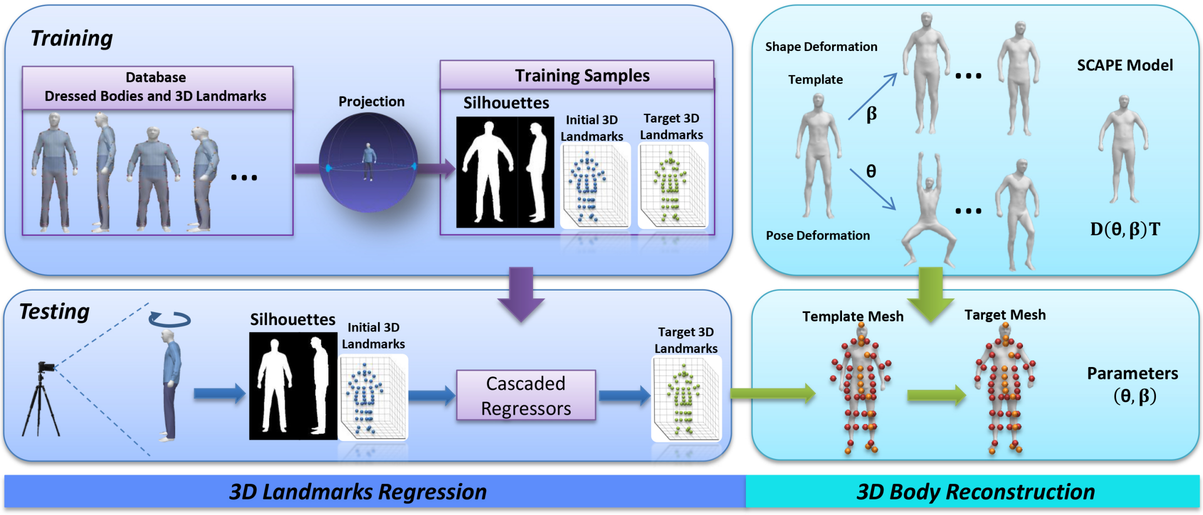

Figure 1: Overview. *Our novel automatic framework experiences two stages: 3D landmarks regression and 3D body reconstruction. We create training samples using our database that contains 3D naked and dressed body pairs. Regressors are trained to learn the predictive relationship between dressed-human silhouettes and target 3D naked body landmarks. The regressed landmarks are used as constraints to optimize SCAPE model which is trained with a database of 3D naked bodies.*

energy between 2D silhouettes and 3D bodies. To acquire efficiency, they defined 2D features for silhouettes and 3D features for bodies and learnt the relationship between them using a database of 3D naked bodies and corresponding naked body silhouettes. Finally, they optimized body mesh using 3D features.

3. Approach Overview

There are two obstacles for "undressing estimation" and we use 3D naked body landmarks to overcome them. One obstacle that hinders the effectiveness of previous methods is the way to remove clothes. We construct a database of naked and dressed body pairs and propose a data-driven method to predict 3D naked body landmarks from dressed-human silhouettes to solve this obstacle. Previous work should find correspondences between the vertices of target mesh and points from input information for body reconstruction, which is another obstacle blocking the efficiency. In our work, the landmarks indices of target mesh are pre-defined and we use predicted naked body landmarks to constrain target mesh, avoiding time-consuming correspondences mapping process.

As figure 1 shows, our automatic framework experiences two stages: 3D landmarks regression and 3D body reconstruction. To analyze the relationship between naked body and its dressed shape, we construct a large database of naked and dressed body pairs. With the help of our database, we create training samples consisting of dressed-human silhouettes, initial 3D landmarks and target 3D landmarks. Database construction and training samples preparation are introduced in section 5. We propose an effective feature descriptor to combine 3D naked body landmarks with dressed-human silhouettes, and train regressors for guiding landmarks movements (from initial landmarks to target landmarks) according to dressed-human silhouettes with training samples. Our regression framework is explained in section 6.1 and feature descriptor is illustrated in section 6.2. In the testing phase, with training results as guidance and a set of initial landmarks, we regress target 3D naked body landmarks. The regressed landmarks are used to constrain SCAPE model [ASK*05] for our body reconstruction, which is introduced in section 4. SCAPE model is trained with a database of 3D naked bodies.

4. Parametric Body Reconstruction

We adopt SCAPE model [ASK*05] for our parametric body reconstruction. SCAPE model decouples human body deformation into pose deformation and shape deformation which are separately controlled by pose parameter $\boldsymbol{\theta}$ and shape parameter $\boldsymbol{\beta}$. Given parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, the vertices positions $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_V\}$ of target mesh are solved by minimizing the least square error:

$$\arg\min_{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_V} \sum_{k=1}^{K} \sum_{d=2}^{3} \| \boldsymbol{R}_{p[k]}(\boldsymbol{\theta}) \boldsymbol{S}_k(\boldsymbol{\beta}) \boldsymbol{Q}_k(\boldsymbol{\theta}) \hat{\boldsymbol{v}}_{d,k} - (\boldsymbol{y}_{d,k} - \boldsymbol{y}_{1,k}) \|^2 \quad (1)$$

where K denotes the total number of triangles and V is the total number of vertices. $\boldsymbol{y}_{1,k}$, $\boldsymbol{y}_{2,k}$ and $\boldsymbol{y}_{3,k}$ are three vertices of a triangle k. $\hat{\boldsymbol{v}}_{d,k}$ is an edge of template mesh, and $\boldsymbol{y}_{d,k} - \boldsymbol{y}_{1,k}$ represents the corresponding edge of target mesh. The human body is divided into 17 partitions and $p[k]$ means the partition p that triangle k locates at. $\boldsymbol{R}_{p[k]}(\boldsymbol{\theta})$ is a 3×3 matrix with a 3-dimensional parameter $\boldsymbol{\theta}$, which represents the rigid rotation for partition p. $\boldsymbol{Q}_k(\boldsymbol{\theta})$ is a 3×3 matrix that shows the non-rigid deformation (e.g., muscle bulging) induced by pose variation. $\boldsymbol{S}_k(\boldsymbol{\beta})$ is a 3×3 matrix explaining the shape variation between different individuals.

The formulation of $\boldsymbol{R}_{p[k]}$, $\boldsymbol{S}_k$ and $\boldsymbol{Q}_k$ can be found in [ASK*05]. We train SCAPE model using MPI database [HSS*09] which consists of pose database and shape database. Pose database contains

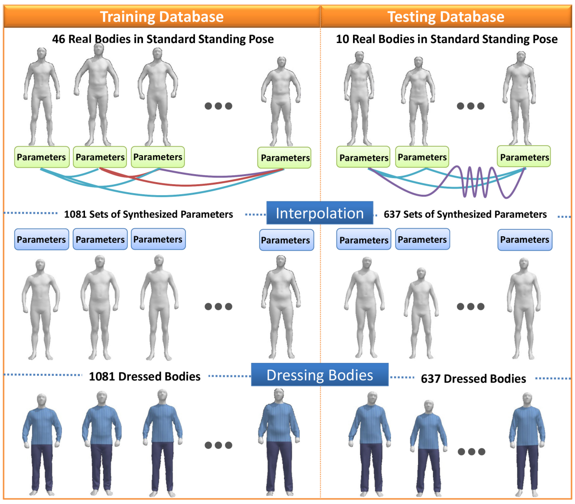
Training Database
Testing Database
46 Real Bodies in Standard Standing Pose
10 Real Bodies in Standard Standing Pose
Parameters
1081 Sets of Synthesized Parameters
Interpolation
637 Sets of Synthesized Parameters
1081 Dressed Bodies
Dressing Bodies
637 Dressed Bodies

Figure 2: Database construction. *Our database consists of 1718 synthesized naked bodies and corresponding dressed bodies with suitable clothes. 56 male bodies with standard standing pose in MPI database are represented as parameters of SCAPE model. These parameters are used for interpolation to generate more sets of parameters which determine our synthesized bodies. We dress the synthesized bodies using a physically based software with one common clothes type.*

one individual with 35 different poses and it is used to train the relationship between $\boldsymbol{Q}_k$ and $\boldsymbol{\theta}$. Shape database contains 56 individuals with one standard standing pose and we utilize it to train the relationship between $\boldsymbol{S}_k$ and $\boldsymbol{\beta}$.

According to formula (1), Cheng at. al. [CTT*] derived the linear representation of $\boldsymbol{Y}$ with reference to $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ respectively. When fixing shape parameter $\boldsymbol{\beta}$ and rotation deformation $\boldsymbol{R}_{p[k]}$, $\boldsymbol{Y}$ is represented as equation (2). $\boldsymbol{c}$ and $\boldsymbol{d}$ are determined by shape parameter $\boldsymbol{\beta}$ and rigid rotation $\boldsymbol{R}_{p[k]}$. Similarly, $\boldsymbol{Y}$ is represented as equation (3) when fixing pose parameter $\boldsymbol{\theta}$. $\boldsymbol{f}$ and $\boldsymbol{g}$ are decided by pose parameter $\boldsymbol{\theta}$. Both the detailed explanation of $\boldsymbol{c}$, $\boldsymbol{d}$, $\boldsymbol{f}$ and $\boldsymbol{g}$ and the derivation of equation (2) and (3) can be found in [CTT*].

$$\boldsymbol{Y} = \boldsymbol{c} \cdot \boldsymbol{\theta} + \boldsymbol{d} \tag{2}$$

$$\boldsymbol{Y} = \boldsymbol{f} \cdot \boldsymbol{\beta} + \boldsymbol{g} \tag{3}$$

As we address in section 3, we use body landmarks to constrain SCAPE model for our body reconstruction to leave out the time-consuming correspondences mapping process. We compute parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ by alternately optimizing two energies:

$$E(\boldsymbol{\theta}) = \sum_{l=1}^{L} \| \boldsymbol{y}_l(\boldsymbol{\theta}) - \boldsymbol{P}_l \|^2 + w_\theta \sum_{p_1, p_2} \| \boldsymbol{\theta}_{p_1} - \boldsymbol{\theta}_{p_2} \|^2 \tag{4}$$

$$E(\boldsymbol{\beta}) = \sum_{l=1}^{L} \| \boldsymbol{y}_l(\boldsymbol{\beta}) - \boldsymbol{P}_l \|^2 + w_\beta (\frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{\Lambda} \boldsymbol{\beta}) \tag{5}$$

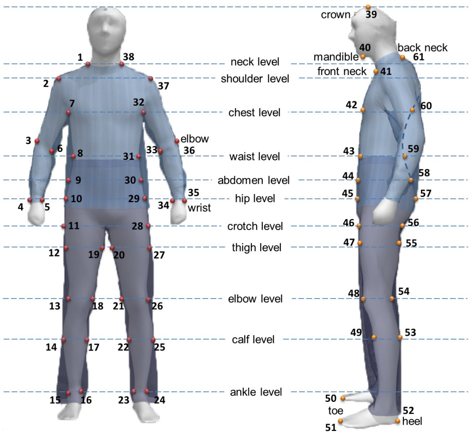

Figure 3: Landmarks. *Landmarks on naked body under clothes. Red vertices are used for front view while yellow ones are used for side view.*

Suppose there are L body landmarks. $\boldsymbol{P}_l$ denotes the position of body landmark l and $\boldsymbol{y}_l$ is the corresponding vertex of target mesh. p_1 and p_2 are two adjacent partitions of body. The second term of equation (4) is a quadratic smoothness term to keep the adjacent body parts change continuously. In equation (5), $\boldsymbol{\Lambda} = diag(1/\sigma_1^2, 1/\sigma_2^2, \cdots, 1/\sigma_B^2)$. σ_i^2 is an eigenvalue from SCAPE shape parameters and B is the dimension of $\boldsymbol{\beta}$. The second term in equation (5) is to regularize $\boldsymbol{\beta}$. w_θ and w_β are weight coefficients and we set $w_\theta = 0.143$ and $w_\beta = 0.002$ for our implementation. The value of w_θ, w_β and the number of iterations are validated in Appendix II.

5. Database Construction and Training Samples Preparation

Because of the difficulty in transformation from dressed body model to naked body model, we use a data-driven approach to indicate naked body information from dressed-human silhouettes. To the best of our knowledge, this is the first large database containing 3D naked and dressed body pairs. Due to the massive manual work of dressing simulation, we currently only use male bodies with standard standing pose in MPI database [HSS*09] for our database construction. We use our database to validate our automatic body shapes estimation framework. The framework can be applied to female situation directly when provided with a database of female naked and dressed body pairs.

Figure 2 illustrates our database construction. We can find that small differences of pose appearance exist between different individuals, even though they are required to hold the same standing pose. Parameters ($\boldsymbol{\theta}$, $\boldsymbol{\beta}$) for each body are acquired by representing the mesh with SCAPE model, and one 3D body mesh is determined by one set of parameters. 1081 bodies are synthesized using 46 real bodies to construct training database. We interpolate one set of parameters from two known sets of body parameters with a random weight between 0 and 1. Similarly, 637 synthesized bodies are obtained using 10 real bodies for testing database. Finally, we design a long-sleeved shirt and long pants using physically based cloth simulation software (Marvelous Designer: www.marvelousdesigner.com) and dress 1718 bodies in suitable sizes.

Based on the observation of landmarks selection in previous work [SI03, ZMK13], we set several critical landmarks (Figure 3) to better represent body shapes, especially for virtual fitting. Some landmarks are used to describe a body shape in front-view silhouette while the others are for side-view silhouette. Since the bodies in our database have the same number of vertices (each of which owns an index) and same topology, we just need to select a series of indices, avoiding manual landmarks annotation for every body.

The constructed database is used to prepare training samples for landmarks regression. As shown in figure 1, an training sample contains initial 3D landmarks, target 3D landmarks and dressed-human silhouettes. We project dressed bodies in front and side views to prepare dressed-human silhouettes. For each body, we have its target 3D landmarks. We obtain initial 3D landmarks by randomly using other samples' target landmarks. To enhance training samples, we assign E sets of initial positions to each of 1081 target positions in our database. Thus, the number of training samples N equals to $1081 \times E$. Effects caused by different enhance extents of training samples are shown in Appendix I.A.

6. 3D Landmarks Regression

3D landmarks regression process plays an important role in our efficient automatic framework. For one hand, 3D landmarks are used to bridge naked body shapes and dressed-human silhouettes, which is a novel idea for "undressing estimation". For another hand, 3D landmarks leave out time-consuming correspondences mapping for body reconstruction, promising the efficiency for our work.

Our regression idea is inspired by previous work while we face new challenges. Dollar et. al. [DWP10] proposed to learn a fixed linear sequence of weak random ferns regressors to predict facial landmarks from a facial image. However, Cao X. et. al. [CWWS14] pointed out that such regressors were too weak and they adopted a two-level boosted regression instead. Cao C. et. al. [CWLZ13] adopted the same regression idea as [CWWS14], but extended it to 3D facial landmarks. Their results showed that this method could be applied to indicate 3D facial landmarks. Cheng et. al. [CTT*] regressed body landmarks with a depth image (of people with tight clothes) to acquire the efficiency of body reconstruction. Unlike facial landmarks, body landmarks lack distinction in image intensity appearance. The input requirement for our work is more challenging — dressed-human silhouettes. We could only make full use of 2D dressed-human silhouettes, where the naked body is occluded by clothes.

6.1. Regression Framework

We try to learn 3D landmarks movements from initial landmarks to target landmarks according to dressed-human silhouettes using training samples. As we illustrated in section 5, we prepare N training samples with our database and each sample consists of initial positions of 3D landmarks, target positions of 3D landmarks and dressed-human silhouettes. Imagine that in the testing phase, we have a testing sample (a pair of dressed-human silhouettes and a set of initial landmarks) and we want to use the landmarks movements of training samples whose relationships between initial landmarks and silhouettes are in similar situation as the testing sample to guide our landmarks movements. Based on this observation, we firstly need an effective feature (denoted as f) to combine landmarks with appearances of silhouettes. Then the feature are used to classify training samples and landmarks movements are computed for each classification.

$$\boldsymbol{P_i} = \boldsymbol{P}_{i-1} + R_i(\boldsymbol{I}, \boldsymbol{f}_i, \boldsymbol{P}_{i-1}) \quad i = 1, 2, \cdots, m \tag{6}$$

As equation (6) shows, 3D body landmarks are regressed in an additive manner. $\boldsymbol{P}_i$ represents positions of landmarks in i^{th} stage. Our training target R_i is a function of silhouettes $\boldsymbol{I}$, feature$\boldsymbol{f}_i$ and current positions of landmarks $\boldsymbol{P}_{i-1}$, whose goal is to reduce the differences between current landmarks and target landmarks of training samples. Suppose training sample j is classified into class Ω_c in i^{th} stage, then its R_i is computed as:

$$R_i = \frac{\sum_{j \in \Omega_c} \delta \boldsymbol{P}_i^j}{|\Omega_c|} \tag{7}$$

where $\delta \boldsymbol{P}_i^j = \boldsymbol{P}_T^j - \boldsymbol{P}_{i-1}^j$. $\boldsymbol{P}_T^j$ denotes the target positions of landmarks and $\boldsymbol{P}_{i-1}^j$ is current positions of landmarks for training sample j. $|\Omega_c|$ represents the total number of training samples in class Ω_c.

Notice that increment R_i can be approximated using equation (7) when the training samples in same classification own similar target landmarks movements. So we propose an effective feature for classification to promise this point, which is introduced in next section.

6.2. Feature Descriptor and Classification

Since both dressed-human silhouettes and target landmarks have intrinsic relationship with ground truth body shapes, we assume that similar target landmarks movements (from current landmarks to target landmarks) mean similar relationship between current landmarks and dressed-human silhouettes. We propose a novel feature descriptor to describe the relationship between 3D landmarks and 2D dressed-human silhouettes. Previous work [DWP10] [CWLZ13] [CWWS14] [RCWS14] [CTT*] used image intensity to define feature descriptor while image intensity is so weak for our black-and-white silhouettes.

Figure 4 shows our feature descriptor. Firstly, we project 3D landmarks to get 2D image points. For simple illustration, we call those points 2D landmarks in the following content. The camera configuration for projection is the same as that used in projecting 3D dressed bodies to get silhouettes. When we use real photos for testing, we estimate the camera configuration, which will be illustrated in section 7.2. Secondly, suppose G sampling points are acquired by sampling around 2D landmarks with a Gaussian distribution whose mean value is 20 pixels and standard deviation value is 2 pixels for a 800×600 image. Finally, we define our feature descriptor as the displacement from a sampling point to its nearest dressed body contour point. We have tried several feature descriptors and choose a more effective and stable one. The comparison

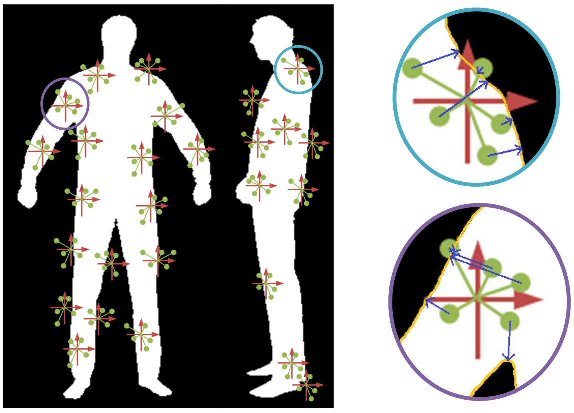

Figure 4: Sampling points and feature descriptor. *2D landmarks are marked as the centers of red crosses and green points represent sampling points. For explicit display, this figure does not show all points. The feature descriptor is defined as displacement from sampling point to its nearest contour point (blue arrow).*

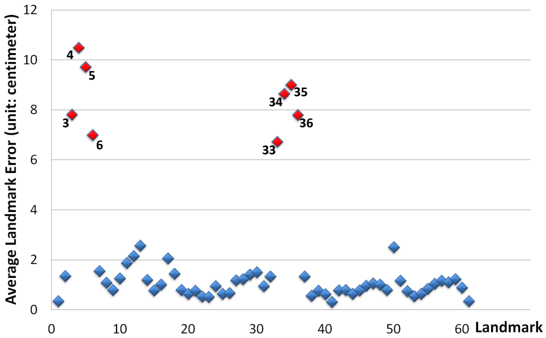

Figure 5: Average error for 61 landmarks. *The location of each landmark is illustrated in figure 3. Each point shows the mean value of landmark error in Euclidean distance for* 637 *samples.*

between effects of different feature descriptors is shown in section 8.2.

The displacement contains values in both x and y direction, so each training sample has $2G$ feature descriptors. A straightforward idea for classification is using a $2G$-dimensional vector that containing all feature descriptors as feature, which is awkward for our regression task. Instead, we adopt a random fern algorithm [OCLF10] to acquire feature reduction and select F out of $2G$ feature descriptors constitute a F-bit label as feature. The values of G and F are explained in Appendix I.C and I.B.

Take the j^{th} training sample in i^{th} stage for example, the values of its selected F feature descriptors are compared with corresponding preset thresholds. If the value of feature descriptor is less than its threshold, the corresponding fern is set to 0. Otherwise, it is set to 1. Consequently, each training sample gets a F-bit binary label and all samples are classified into 2^F classifications in each stage.

We propose a variance-ranked method to select F feature descriptors. Feature descriptor i for N training samples forms a set D^i $(i = 1, 2, \cdots, 2G)$ which contains N elements. Then we compute the variance of each set and choose the top F maximal ones. We reduce the variance after using it to make full use of all feature descriptors. Comparison with other feature selection method can be found in our Appendix I.D.

6.3. Testing Phase

Both the configuration for regression (e.g., sampling vector, selected feature descriptors in each regressor and corresponding preset thresholds) and training results (i.e., landmarks movements for each classification in each regressor) are recorded to guide 3D landmarks regression for testing. During testing phase, we use landmarks of the mean body shape as initial landmarks for the testing sample. In each stage of cascaded regression, feature is computed and the testing sample is classified into one classification. Then landmarks positions are updated using the landmarks movements for that classification.

7. Experiment Results

Our program is run on a 64-bit desktop machine with 3.5GHz Intel(R) Xeon(R) CPU and 32GB RAM. With single-core and single-thread programming, 3D landmarks regression only takes average of 0.028 second using two silhouettes with resolution 800 × 600. 3D body reconstruction costs 3.583 seconds on average.

To demonstrate the accuracy of our method, we show the statistical landmarks regression error and body reconstruction error when high-quality silhouettes are provided (section 7.1). To demonstrate the practicability of our method, we use real human photos as input to estimate 3D body shapes, the silhouettes of which may contain noises (section 7.2). Section 7.3 highlights our advantages when compared with previous work.

7.1. Testing Database Trials

Our testing database consists of 637 naked and dressed body pairs, which is not overlapped with training database. We project those dressed bodies in front and side views to get silhouettes that are free of noises. The positions of the mean body's landmarks are used as initial positions for testing samples.

We evaluate the regression method through computing 3D landmarks error. For each landmark, we compute the average error for 637 testing samples and show the error for 61 landmarks in figure 5. Those 8 landmarks with large error locate at arms (index 3-6, 33-36), where pose ambiguity exists when we only have front-view and side-view silhouettes. However, we aim to estimate 3D body shapes more than accurate poses so that we tolerate the large error for those points. Excluding those 8 landmarks, the average error for other 53 landmarks in x/y/z direction and Euclidean distance are shown in table 1. The coordinate for human body is shown in Figure 7, and z direction points from human body to camera.

Table 1: *Average 3D landmarks error*

X Direction	Y Direction	Z Direction	Euclidean Distance
0.37 cm	0.44 cm	0.65 cm	1.04 cm

For 637 testing samples, let the registered estimated bodies and ground truth bodies own the same pose parameter and we compute vertex error to evaluate shape divagation. Figure 6 shows the average error for each vertex. We further evaluate shape estimation in body measurements (height, chest size, waist size and hip size) and the cumulative error distribution is displayed in figure 7. About 50% of testing samples' body measurements error is less than $1cm$ and 90% is less than $3cm$, which satisfy the size tolerance for ordinary clothes [ZM15].

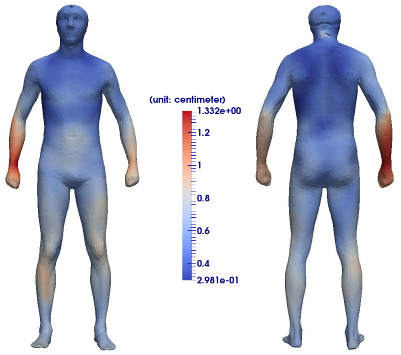

Figure 6: Average vertex error. *The colors of points show the average error of vertices for* 637 *samples.*

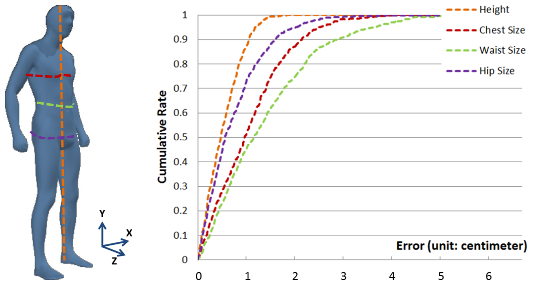

Figure 7: Cumulative error distribution in body measurements. *Cumulative error distribution in height, chest size, waist size and hip size aspects.*

7.2. Real Human Photos Testing

We take photos of 9 persons wearing casual clothes (similar to the ones in our database) holding the standard standing pose with only 1 phone camera at a fixed distance (3.5 meters), and figure 8 shows one example. Dressed-human silhouettes are obtained with the help of Photoshop. After being captured the front view, the user turns to his right side holding the same pose. The pose clues provided by orthogonal-view silhouettes do not contradict with each other with small pose differences. We use camera calibration toolbox for Matlab (`http://www.vision.caltech.edu/bouguetj/calib_doc/`) to estimate camera intrinsic parameters and opencv solvePnP function to work out the extrinsic parameters of phone camera. Table 2 shows the mean error between estimated body and ground truth in body measurements for 9 persons. Some of the virtual results are shown in figure 9.

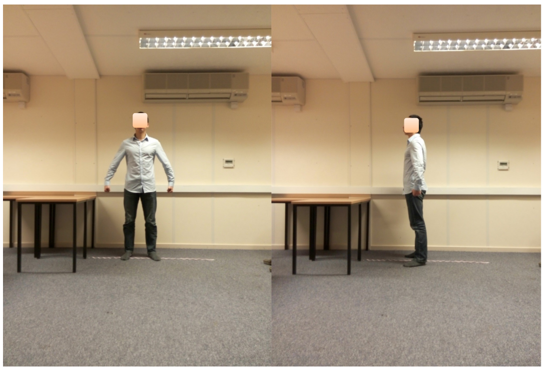

Figure 8: Testing photos. *An example of testing photos for 2 views.*

Table 2: *Average body measurements error*

Height	Chest Size	Waist Size	Hip Size
1.76cm	1.78cm	1.67cm	1.84cm

7.3. Comparison with Previous Work

We use dressed-human silhouettes as input, which is more convenient for users. Some works [HSR*09, WPB*14, NH14, ZCD*15] taking 3D scanning points as input required 3D scanners. Zhu et. al. [ZM15] made users to interactively estimate some body points under clothes.

The details of 3D body reconstruction of previous work are explained in section 2.2. Balan et. al. [BB08] used a gradient-free direct search simplex method for optimization which costs 40 minutes. Hasler et. al. [HSR*09] spent 11.5 minutes on optimization and Chen et. al. [CGZZ13] spent 1 minute. We adopt an efficient regression method to acquire 3D landmarks positions within 0.03 second. We avoid time-consuming vertices correspondence finding process because those landmarks are preset on registered bodies. Our body shape optimization target is minimizing the Euclidean distance between regressed 3D landmarks and corresponding vertices of a template mesh and the process takes less than 4 seconds.

Balan et. al. [BB08] tested 6 persons wearing 6-10 kinds of ordinary clothes with 11 poses. Wuhrer et. al. [WPB*14] had 18 scans of 5 persons in casual office clothes with up to 5 poses. For average height/chest size/waist sizes error, Balan et. al. achieved about

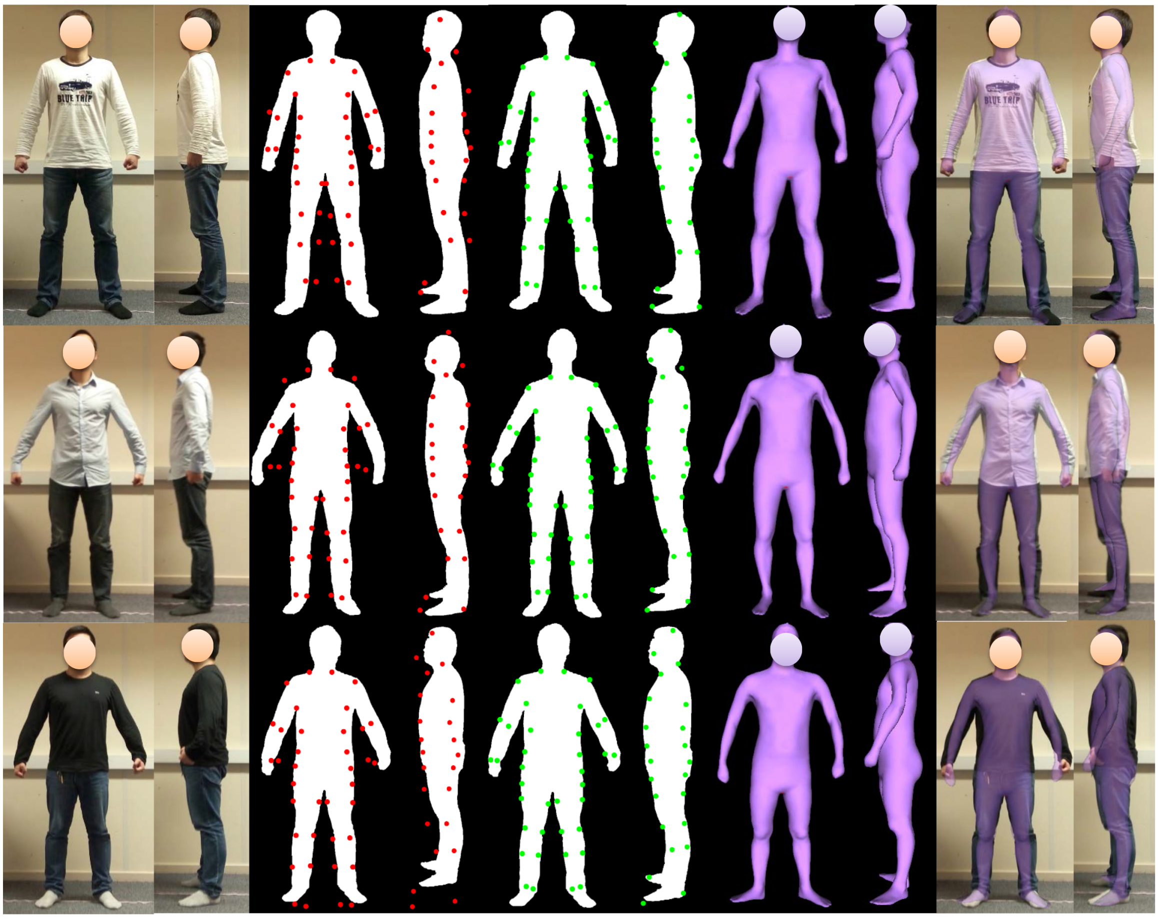

Figure 9: Visualization of body shapes estimation under clothes. *From left to right pairs: photo in front/side views, location of projection points for initial 3D landmarks in front/side views, location of projection points for estimated 3D landmarks in front/side views, estimated 3D body shapes in front/side views, visualization of estimated body shapes under clothes in front/side views.*

1.03/4.65/4.73 centimeters while Wuhrer et. al. got 2.52/14/15.8 centimeters. Our results are shown in table 2.

8. Discussion

We have tried using one front-view silhouette as our input and some other feature descriptors, and their performances are discussed in the following content.

8.1. One or Two Silhouettes

If we only use the front-view silhouette as input, the landmarks used are red points in figure 3. Compared with Table 1, the landmarks regression error in x direction decreases while the others increase (Table 3). It is understandable because the side-view silhouette provides more information for y and z directions. Figure 10 validates that adding side-view silhouette improves body reconstruction a lot.

Table 3: *Average 3D landmarks error for 1 silhouette*

X Direction	Y Direction	Z Direction	Euclidean Distance
0.25 cm	0.71 cm	3.50 cm	3.66 cm

8.2. Feature Descriptors

We have tried different feature descriptors (illustrated in figure 11) and compare their performances to decide the one which is more suitable for our method. We test different feature descriptors using the same testing database containing 637 samples. The performance for landmarks regression error is illustrated in table 4 and body shape estimation error is compared in figure 12.

Figure 12 shows that feature descriptor (b) and (e) performs better than the others in body shapes estimation. Since the average landmarks regression errors of feature descriptor (b) and (e) are nearly the same, we further compare the standard deviation value of them (table 5). Finally, we choose (e) as our feature descriptor.

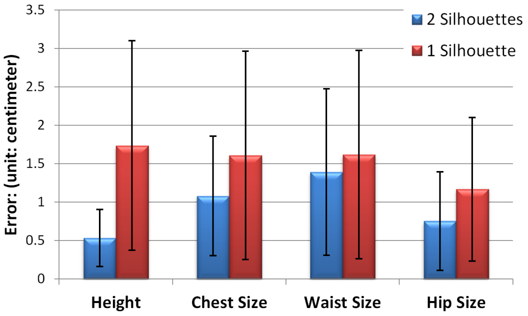

Figure 10: Comparison in body measurements error for using 1 and 2 silhouettes as input. *The height of bar shows the mean value of error while the length of black line equals to double standard deviation of error.*

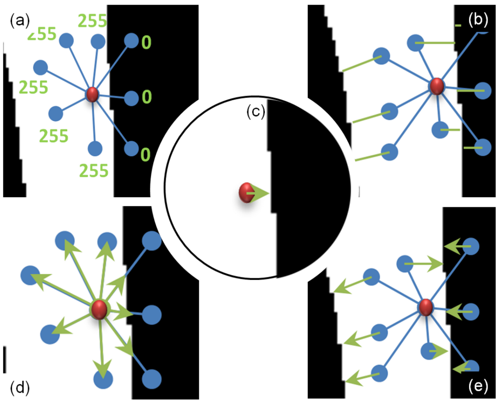

Figure 11: Different feature descriptors. *2D landmarks are marked as red points. Blue points show sampling points around 2D landmarks. The feature descriptor is respectively defined as: (a)pixel value of sampling point; (b)distance between sampling point and its nearest contour point (the value is positive if sampling point locates outside of human contour, otherwise negative); (c)displacement from 2D landmark to its nearest contour point; (d)displacement from 2D landmark to intersection point (sampling point, if there is no intersection between sampling vector and dressed body contour); (e)displacement from sampling point to its nearest contour point.*

Table 4: *Average 3D landmarks error for different descriptors*

FD Error	(a)	(b)	(c)	(d)	(e)
X Direction	0.42cm	0.37cm	0.64cm	0.40cm	0.37cm
Y Direction	0.61cm	0.43cm	1.34cm	0.60cm	0.44cm
Z Direction	0.85cm	0.65cm	1.11cm	0.74cm	0.65cm
Euclidean Distance	1.32cm	1.03cm	2.15cm	1.25cm	1.04cm

Table 5: *Standard deviation of 3D landmarks error*

Error FD	X Direction	Y Direction	Z Direction	Euclidean Distance
(b)	0.4466cm	0.3700cm	0.6723cm	0.6850cm
(e)	0.4404cm	0.3605cm	0.6570cm	0.6574cm

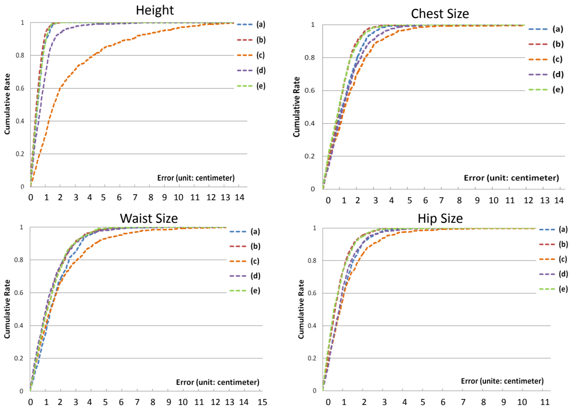

Figure 12: Performance comparison for different feature descriptors in body shapes estimation. *Cumulative error distribution in height, chest size, waist size and hip size aspects.*

9. Conclusions, Limitations and Future Work

We propose a novel automatic framework to efficiently estimate 3D body shapes from dressed-human silhouettes. We build a large database containing 3D naked and dressed body pairs, which may benefit the areas of human body estimation and cloth simulation. Critical vertices are selected as landmarks to represent body shapes, which leaves out the time-consuming vertices correspondences finding process for body reconstruction and promises efficiency of our method. We explore a novel landmark-indexed feature descriptor to combine 3D body landmarks with 2D dressed-human silhouettes. Based on our constructed database, we learn a regression function to predict 3D landmarks positions according to 2D dressed-human silhouettes with our effective feature. 3D bodies are acquired by constraining SCAPE model with regressed landmarks.

Experiments show that our approach achieves good reconstruction results in body measurements, satisfying the size tolerance of clothing industry. 3D body shapes are estimated within 4 seconds automatically while the fastest method reported previously need 1 minute. We also validate key implementation configurations for our method. Our work makes it more convenient for ordinary users to access their 3D body shapes, which will accelerate the evolution of virtual fitting industry in the future.

In the following contents, we show the limitations and future work of our work. We use silhouettes as input of our core algorithm, because clothes texture is of less use for "undressing estimation". Our method is affected by the quality of silhouettes obtained from photos, so we may explore excellent image segmentation technique in the future.

Our database depends on cloth simulation technique, which currently performs realistic simulation on some fabric properties and clothes types. With the development of cloth simulation, we could extend our method to more clothes types. Because of the tremendous manual efforts of dressing simulation for database construction, we now only have male bodies and one set of clothes in our database. We would like to explore automatic dressing simulation technique and extend our work to female and more clothes types situations.

The current feature descriptor which combines 3D landmarks with silhouettes restricts that we should know the camera configuration. The estimation of camera configuration also brings error. In the future, We would explore a new feature descriptor and a new parametric human body model that are more applicable for this problem.

References

[ACP03] ALLEN B., CURLESS B., POPOVIC Z.: The space of human body shapes: reconstruction and parameterization from range scans. *ACM Transactions on Graphics 22*, 3 (JUL 2003), 587–594. 2

[ASK*05] ANGUELOV D., SRINIVASAN P., KOLLER D., THRUN S., RODGERS J., DAVIS J.: SCAPE: Shape Completion and Animation of People. *ACM Transactions on Graphics 24*, 3 (JUL 2005), 408–416. 2, 3

[BB08] BALAN A. O., BLACK M. J.: The Naked Truth: Estimating Body Shape Under Clothing. In *Computer Vision - ECCV 2008, PT II, Proceedings* (2008), vol. 5303 of *Lecture Notes in Computer Science*, SPRINGER-VERLAG BERLIN, pp. 15–29. 1, 2, 7

[BSWX13] BOISVERT J., SHU C., WUHRER S., XI P.: Three-dimensional human shape inference from silhouettes: reconstruction and validation. *Machine Vision and Applications 24*, 1 (JAN 2013), 145–157. 2

[CC09] CHEN Y., CIPOLLA R.: Learning shape priors for single view reconstruction. In *IEEE 12th International Conference on Computer Cision Workshops(ICCV Workshops)* (2009), pp. 1425–1432. 2

[CGZZ13] CHEN X., GUO Y., ZHOU B., ZHAO Q.: Deformable model for estimating clothed and naked human shapes from a single image. *Visual Computer 29*, 11 (NOV 2013), 1187–1196. 1, 2, 7

[CTT*] CHENG K.-L., TONG R.-F., TANG M., SARKIS M., QIAN J.-Y.: Parametric human body reconstruction based on sparse key points. *IEEE Transactions on Visualization and Computer Graphics*. doi:10.1109/TVCG.2015.2511751. 1, 2, 4, 5

[CWLZ13] CAO C., WENG Y., LIN S., ZHOU K.: 3D Shape Regression for Real-time Facial Animation. *ACM Transactions on Graphics 32*, 4 (JUL 2013). 5

[CWWS14] CAO X., WEI Y., WEN F., SUN J.: Face Alignment by Explicit Shape Regression. *International Journal of Computer Vision 107*, 2, SI (APR 2014), 177–190. 5

[DWP10] DOLLAR P., WELINDER P., PERONA P.: Cascaded Pose Regression. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2010), IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer soc, pp. 1078–1085. 5

[GWBB09] GUAN P., WEISS A., BALAN A. O., BLACK M. J.: Estimating Human Shape and Pose from a Single Image. In *2009 IEEE 12th International Conference on Computer Vision (ICCV)* (2009), IEEE International Conference on Computer Vision, IEEE, pp. 1381–1388. 2

[HSR*09] HASLER N., STOLL C., ROSENHAHN B., THORMAEHLEN T., SEIDEL H.-P.: Estimating body shape of dressed humans. *Computers & Graphics-uk 33*, 3, SI (JUN 2009), 211–216. 1, 2, 7

[HSS*09] HASLER N., STOLL C., SUNKEL M., ROSENHAHN B., SEIDEL H. P.: A Statistical Model of Human Pose and Body Shape. *Computer Graphics Forum 28*, 2 (2009), 337–346. 2, 3, 4

[JTST10] JAIN A., THORMAEHLEN T., SEIDEL H.-P., THEOBALT C.: MovieReshape: Tracking and Reshaping of Humans in Videos. *ACM Transactions on Graphics 29*, 6 (DEC 2010). 2

[KSvdP09] KRAEVOY V., SHEFFER A., VAN DE PANNE M.: Modeling from contour drawings. In *Proceedings of the 6th Eurographics Symposium on Sketch-Based Interfaces and Modeling* (2009), pp. 37–44. 2

[LVG*13] LI H., VOUGA E., GUDYM A., LUO L., BARRON J. T., GUSEV G.: 3D Self-Portraits. *ACM Transactions on Graphics 32*, 6 (NOV 2013). 2

[NH14] NEOPHYTOU A., HILTON A.: A layered model of human body and garment deformation. In *2014 2nd International Conference on 3D Vision (3DV)* (2014), vol. 1, pp. 171–178. 1, 2, 7

[OCLF10] OEZUYSAL M., CALONDER M., LEPETIT V., FUA P.: Fast Keypoint Recognition Using Random Ferns. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 3 (MAR 2010), 448–461. 6

[RCWS14] REN S., CAO X., WEI Y., SUN J.: Face Alignment at 3000 FPS via Regressing Local Binary Features. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 1685–1692. 5

[SI03] SIMMONS K. P., ISTOOK C. L.: Body measurement techniques: Comparing 3d body-scanning and anthropometric methods for apparel applications. *Journal of Fashion Marketing and Management: An International Journal 7*, 3 (2003), 306–332. 5

[SP04] SUMNER R., POPOVIC J.: Deformation transfer for triangle meshes. *ACM Transactions on Graphics 23*, 3 (AUG 2004), 399–405. 2

[SYW06] SEO H., YEO Y. I., WOHN K.: 3D body reconstruction from photos based on range scan. In *Technologies for E-learning and Digital Entertainment, Proceedings* (2006), vol. 3942 of *Lecture Notes in Computer Science*, Springer-verlag Berlin, pp. 849–860. 2

[TCL*13] TAM G. K. L., CHENG Z.-Q., LAI Y.-K., LANGBEIN F. C., LIU Y., MARSHALL D., MARTIN R. R., SUN X.-F., ROSIN P. L.: Registration of 3D Point Clouds and Meshes: A Survey from Rigid to Nonrigid. *IEEE Transactions on Visualization and Computer Graphics 19*, 7 (JUL 2013), 1199–1217. 2

[TZL*12] TONG J., ZHOU J., LIU L., PAN Z., YAN H.: Scanning 3D Full Human Bodies Using Kinects. *IEEE Transactions on Visualization and Computer Graphics 18*, 4 (APR 2012), 643–650. 2

[WHB11] WEISS A., HIRSHBERG D., BLACK M. J.: Home 3D Body Scans from Noisy Image and Range Data. In *2011 IEEE International Conference on Computer Vision (ICCV)* (2011), IEEE, pp. 1951–1958. 2

[WPB*14] WUHRER S., PISHCHULIN L., BRUNTON A., SHU C., LANG J.: Estimation of human body shape and posture under clothing. *Computer Vision and Image Understanding 127* (OCT 2014), 31–42. 1, 2, 7

[WS13] WUHRER S., SHU C.: Estimating 3D human shapes from measurements. *Machine Vision and Applications 24*, 6 (AUG 2013), 1133–1147. 2

[ZCD*15] ZENG M., CAO L., DONG H., LIN K., WANG M., TONG J.: Estimation of human body shape and cloth field in front of a kinect. *Neurocomputing 151*, 2 (MAR 5 2015), 626–631. 1, 2, 7

[ZFL*10] ZHOU S., FU H., LIU L., COHEN-OR D., HAN X.: Parametric Reshaping of Human Bodies in Images. *ACM Transactions on Graphics 29*, 4 (JUL 2010). 2

[ZM15] ZHU S., MOK P.: Predicting realistic and precise human body models under clothing based on orthogonal-view photos. *Procedia Manufacturing 3* (2015), 3812–3819. 1, 2, 7

[ZMK13] ZHU S., MOK P. Y., KWOK Y. L.: An efficient human model customization method based on orthogonal-view monocular photos. *Computer-aided Design 45*, 11 (NOV 2013), 1314–1332. 5