IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 5, 2018, accepted May 7, 2018, date of publication May 16, 2018, date of current version June 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2837147

Data-Driven 3-D Human Body Customization With a Mobile Device

DAN SONG[1], RUOFENG TONG[1], JIANG DU[1], YUN ZHANG[2], AND YAO JIN[3]

[1]State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou 310027, China
[2]Institute of Zhejiang Radio and TV Technology, Zhejiang University of Media and Communications, Hangzhou 310027, China
[3]College of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310027, China

Corresponding author: Ruofeng Tong (trf@zju.edu.cn)

This work was supported in part by NSFC under Grant 61572424 and in part by the Science and Technology Department of Zhejiang Province under Grant 2018C01080. The work of Y. Zhang was supported by NSFC under Grant 61602402. The work of Y. Jin was supported by the Zhejiang Provincial Natural Science Foundation of China under Grant LY17F020031.

ABSTRACT It is more convincing for users to have their own 3-D body shapes in the virtual fitting room when they shop clothes online. However, existing methods are limited for ordinary users to efficiently and conveniently access their 3-D bodies. We propose an efficient data-driven approach and develop an android application for 3-D body customization. Users stand naturally and their photos are taken from front and side views with a handy phone camera. They can wear casual clothes like a short-sleeved/long-sleeved shirt and short/long pants. First, we develop a user-friendly interface to semi-automatically segment the human body from photos. Then, the segmented human contours are scaled and translated to the ones under our virtual camera configurations. Through this way, we only need one camera to take photos of human in two views and do not need to calibrate the camera, which satisfy the convenience requirement. Finally, we learn body parameters that determine the 3-D body from dressed-human silhouettes with cascaded regressors. The regressors are trained using a database containing 3-D naked and dressed body pairs. Body parameters regression only costs 1.26 s on an android phone, which ensures the efficiency of our method. We invited 12 volunteers for tests, and the mean absolute estimation error for chest/waist/hip size is 2.89/1.93/2.22 centimeters. We additionally use 637 synthetic data to evaluate the main procedures of our approach.

INDEX TERMS Body parameters regression, data-driven application, image-based 3-D body shape estimation.

I. INTRODUCTION

Customized body shapes play an important role in the popularity of virtual fitting room for online shopping. When the 3D body model in the virtual fitting room owns the customer's shape, the buyer can get convincing visual information and size suggestions. This produces a win-win situation for both customers and sellers, saving their time and increasing the number of successful transactions.

For ordinary online customers, they want the estimation of their 3D body shapes to be convenient, fast and accurate. However, existing methods cannot satisfy all the requirements.

High-end scanners can be used to scan individuals with minimal or tight clothes for 3D body reconstruction, producing accurate results. Nevertheless, minimal or tight dressing makes customers embarrassed and scanners are costly and not widely used. Range cameras (e.g., Kinect) provide a less expensive way for 3D body reconstruction [1]–[3], but they are still not widely available. Images are more convenient to access, and some researchers take minimally dressed human images as input to constrain parametric human body models [4]–[7]. Parametric models are trained using a database of human bodies, and they represent 3D bodies by deforming a template body with a set of parameters.

Recently, a number of strategies are designed for the situation that the human body in the image is covered with clothes. First of all, dressed-human contours are extracted from photos. Zhou *et al.* [8] use dressed-human contours to constrain parametric naked body models which restrict estimated bodies to stay in the human body space. However, the clothes occlusions make the estimated bodies fatter. Balan and Black [9] take photos of human in 4 views with calibrated cameras. They detect skin parts of dressed-human contours and set higher weights for these parts. Chen *et al.* [10] propose a parametric clothed body model for several clothes types. These methods need to iteratively find the correspondences between the 2D contour points and the 3D vertices of the body model, which is time-consuming. Song *et al.* [11] construct a database consisting of 3D naked and dressed body pairs, and adopt a data-driven method to avoid the correspondence finding process. They regress 3D body landmarks using dressed-human contours in front and side views with calibrated cameras, and optimize body parameters using 3D landmarks. Camera calibration hinders the convenience for ordinary users, and their body parameters optimization costs 3.583 seconds on a PC. The time-consumption for mobile devices will be more.

We try to efficiently estimate 3D body shapes through mobile devices which are widely available. Ballester *et al.* [12] also reconstruct human bodies using a mobile phone app, but their body reconstruction is done on a remote server. We adopt the regression method proposed by Song *et al.* [11] to stay away from the time-consuming process of correspondence finding. To further improve the efficiency for the limited computation resources of mobile devices, we regress body parameters, rather than first regressing 3D landmarks and then iteratively solving body parameters like Song *et al.* [11] do. We also propose a strategy to get rid of camera calibration. The whole procedure and main techniques of our approach are introduced in the following paragraph.

Firstly, we capture the front and side views of a person holding a natural standing pose in casual clothes like a short-sleeved/long-sleeved shirt and shot/long pants using a phone camera. Dressed-human contours are obtained with several user-specified strokes. Secondly, in accordance with virtual camera configurations, the segmented human contours are scaled and positioned using height information. Thirdly, we regress body parameters from front and side silhouettes and reconstruct 3D body. We use Grabcut method [13] to segment the human and develop a user-friendly interface for segmentation. Grabcut method works relatively fast when the image size is as small as the size of the phone screen. Users can zoom in the image and are allowed to add strokes indicating the human area and the background according to current segmentation result. We use the database of 3D naked and dressed body pairs created by Song *et al.* [11] for the following training purposes: (1) We learn the relationship between the size of body contours under virtual camera configurations and human height information; (2) We define a reference point on the body to determine the positions of virtual cameras, and learn to regress the location of projected reference point from the body contour; and (3) We learn to predict body parameters from dressed-human silhouettes. 12 volunteers take part in our test, and we take photos of them and measure their height information, chest, waist and hip sizes. Height information is used as input, and the mean absolute estimation errors for chest, waist and hip sizes are 2.89, 1.93 and 2.22 centimeters respectively.

In summary, we develop a user-friendly tool for ordinary users to efficiently and conveniently obtain their 3D bodies with following contributions: (1) We implement our method on an Android device and we do not need a remote server, which is good for keeping customers' privacy; (2) Faced with the limited computation resources of mobile devices, we regress body parameters for reconstruction which only costs 1.26 seconds; and (3) We propose a strategy to get rid of camera calibration, improving the convenience for users.

II. RELATED WORK

A. PARAMETRIC HUMAN BODY MODELS

Parametric body models represent human body through deforming a template body mesh with a set of parameters. SCAPE model [14] captures the variations of body mesh as edge deformations. Each edge goes through rigid rotation deformation, pose-induced non-rigid deformation and non-rigid shape deformation. Pose-induced non-rigid deformation is linear to the relative partition rotation of neighbouring body partitions, and a shape space is learned to explain the shape deformation. Hasler *et al.* [15] treat the edge deformation as affine transformation, and they factorize the affine matrix to rotation part and stretch part. Then they learn the deformation space by performing PCA (Principal Component Analysis). Later in 2010, they [16] learn a multilinear pose and shape model to separate pose and shape deformations. Zhu *et al.* [17] design the template mesh in low and high quality, and transfer the deformation of low-resolution template mesh to the deformation of high-resolution mesh via the displacements of vertices. Tsoli *et al.* [18] augment SCAPE model by adding parameters of breathing model to animate human breathing. Pons-Moll *et al.* [19] propose Dyna to capture soft-tissue deformations induced by motions. SMPL model [20] is a skinning vertex-based model. Apart from the classic skinned structure, it adds displacement caused by individual pose and shape to body vertex.

B. IMAGE-BASED BODY RECONSTRUCTION

Bălan and Black [9] set 4 calibrated cameras around the human to obtain dressed-human silhouettes in 4 views. They detect skin parts of the dressed-human contour and set higher weights for them to constrain SCAPE model. Given the subject's height and a few clicked points for body joints, Guan *et al.* [6] initialize the pose and shape of body. They use contour and shading information as clues to optimize the body parameters of the height-preserving SCAPE model. Zhou *et al.* [8] integrate SCAPE model into the reshaping of human bodies in the images. Chen *et al.* [10] extend SCAPE model to a dressed human model. They use deformation transfer [21] to construct a database of naked and dressed bodies, and learn clothes-related coefficients. They reconstruct both naked and dressed bodies from image by optimizing parameters. All of the methods stressed above need to iteratively find the correspondences between contour

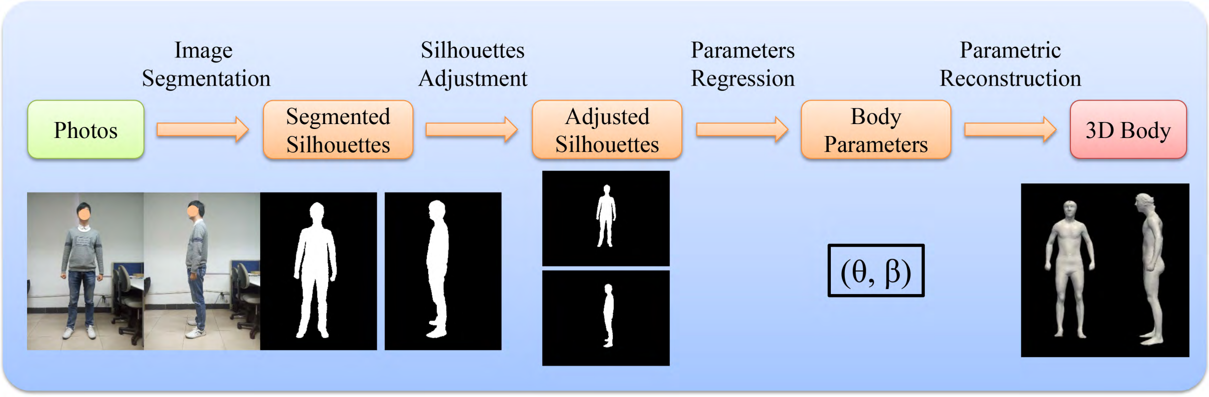


FIGURE 1. ***Overview.*** **We take photos of human body from front view and side view with a phone camera. Human areas are segmented from photos with several user strokes. Then segmented human contours are scaled and translated to the ones under virtual camera configuration. Body parameters which determine the 3-D body are regressed from the adjusted silhouettes.**

points in the input image and 3D vertices on the body mesh (or 2D projected body points), which is time-consuming.

Pre-defined landmarks are proposed to avoid correspondence finding process. Zhu and Mok [22] make users to estimate the locations of body landmarks under clothes in the image. Corresponding 3D vertices are predicted from 2D landmarks with a neural network. They construct a coarse mesh with these 3D vertices, and compute the deformation from the template coarse mesh to the constructed coarse mesh. Finally they transfer the deformation to a refined template mesh. Song *et al.* [11] construct a database of naked and dressed bodies with physically based cloth simulation, and learn to predict 3D landmarks from dressed-human silhouettes. With 3D landmarks as constraint, they iteratively optimize the body parameters of SCAPE model.

C. LANDMARKS REGRESSION

Our parameters regression approach is inspired by landmarks regression methods applied in face alignment and body landmarks detection. Dollár *et al.* [23] propose CPR (cascaded pose regression) method to efficiently and effectively detect landmarks which describe the object's attributes in image. They introduce a landmarks-indexed feature descriptor, and use it to describe the relationship between current landmarks and input image. They use random ferns approach for reducing the dimensionality of feature, which randomly selects several bits from the feature descriptor vector. Cao *et al.* [24] further improve CPR method. They propose a two-level regression framework, and select the bits which are most correlated with regression target. Cao *et al.* [25] use similar method to regress 3D facial landmarks. Cheng *et al.* [1] are the first to use such a method to regress body landmarks, accelerating the process of 3D body reconstruction. The feature descriptors of the methods mentioned above are all based on the pixel values of landmarks. When the input information lack valuable color information, Song *et al.* [11] propose a feature descriptor defined as the displacements between landmarks and corresponding nearest contour points, and they regress body landmarks from dressed-human silhouettes.

III. OVERVIEW AND TRAINING DATA PREPARATION

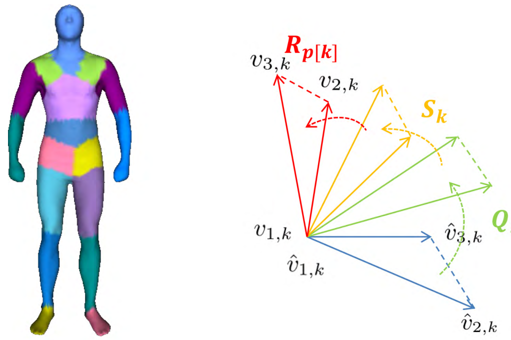
A. OVERVIEW

Fig. 1 illustrates the framework of our approach. We first segment dressed-human areas and get human contours (section IV-A). Then the contours are scaled and translated approximately to the ones under the configurations of virtual cameras used in the training phase (section IV-B). Finally, we regress body parameters which determine 3D body from dressed-human silhouettes with a series of trained regressors (section IV-C).

B. TRAINING DATA PREPARATION

We use the training database containing 5405 × 3 pairs of naked and dressed 3D bodies, which is created by Song et al. [11], to prepare training data. Song *et al.* [11] manually dress 3D bodies with three sets of clothes, namely L/L, S/L and S/S, which are respectively abbreviated for long-sleeved shirt and long pants, short-sleeved shirt and long pants, and short-sleeved shirt and short pants. They design the clothes types by a software based on cloth simulation and dress bodies one by one with suitable clothes size, which needs a lot of efforts. The clothes types currently chosen are most commonly used in daily life and well simulated. Our approach can be applied to more clothes types once the database prepares corresponding naked and dressed body pairs.

Dressed-human silhouettes are obtained by projecting 3D dressed bodies in front view and side view with virtual camera configuration as shown in Fig. 4. For the process of silhouettes adjustment introduced in section IV-B, we predict the size of silhouette (i.e. BBH, the bounding box height of human contour in image) using human height information, and we prepare each training data as (human height, BBH). Besides the size, we need to adjust the positions of human contours in front and side silhouettes, where each training



(a) (b)

FIGURE 2. ***Human partition and triangle-based deformation.*** **(a) We divide the human body into 17 partitions. (b) SCAPE parametric model deforms each edge of body mesh with 3 × 3 matrix Q_k (pose-induced non-rigid deformation), S_k (non-rigid shape deformation) and $R_{p[k]}$ (pose-induced rigid deformation).**

data is set as (silhouette, 2D landmarks) for both front and side conditions. For the process of parameters regression illustrated in section IV-C, we prepare training data as (front and side silhouettes, body parameters). Human body parameters are introduced in the next subsection.

C. HUMAN BODY PARAMETERS

We use SCAPE model [14] to represent the 3D human body. The template body mesh is deformed according to a set of parameters $(\boldsymbol{\theta}, \boldsymbol{\beta})$. As Fig. 2 shows, human body is divided into 17 partitions. $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \ldots, \boldsymbol{\theta}_{17}^T]^T$ is 4×17 dimensional and $\boldsymbol{\theta}_i$ $(i = 1, 2, \ldots, 17)$ is the quaternion representation for the relative rotation of part i with respect to the corresponding part of template mesh. $\boldsymbol{\beta}$ determines the shape, and we make it 30 dimensional in our implementation.

Given the low-dimensional body parameters $(\boldsymbol{\theta}, \boldsymbol{\beta})$, the high-dimensional 3D positions of V vertices $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_V\}$ are solved by minimizing the least square error:

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

where, K represents the number of triangles on the template triangular mesh. $p[k]$ denotes that triangle k belongs to partition p. The 3-dimensional vector $\boldsymbol{y}_{1,k}$, $\boldsymbol{y}_{2,k}$ and $\boldsymbol{y}_{3,k}$ are the positions of vertices in the triangle k, and $\hat{\boldsymbol{e}}_{d,k}$ is the corresponding edge of template mesh. The 3×3 matrix $\boldsymbol{R}_{p[k]}(\boldsymbol{\theta})$ represents the relative rotation for part p, and the 3×3 matrix $\boldsymbol{Q}_k(\boldsymbol{\theta})$ illustrates non-rigid deformation caused by pose. $\boldsymbol{S}_k(\boldsymbol{\beta})$ is a 3×3 matrix explaining non-rigid shape deformation.

$\boldsymbol{R}_{p[k]}(\boldsymbol{\theta})$ transforms quaternion representation to rotation matrix. We put the 3×3 matrix $\boldsymbol{Q}_k(\boldsymbol{\theta})$ in a 9-dimensional vector form $\boldsymbol{q}_k(\boldsymbol{\theta})$, and $\boldsymbol{q}_k(\boldsymbol{\theta})$ is linear to the neighbouring joint angles with coefficient α_k (equation (2)). $\Delta A_{p[k]}(\boldsymbol{\theta})$ is the axis-angle representation for the relative rotation of partition $p[k]$ with respect to its neighbouring partition(s). It is a 3-dimensional or 6-dimensional vector. $\boldsymbol{\alpha}_k$ is trained using the pose database of MPI database [26]. We put the 3×3 matrix $\boldsymbol{S}_k(\boldsymbol{\beta})$ in a 9-dimensional vector form $\boldsymbol{s}_k(\boldsymbol{\beta})$, and $\boldsymbol{s}_k(\boldsymbol{\beta})$ is the linear combination of the bases of a body shape space (equation (3)). The shape database of MPI database [26] is used to train a body shape space. μ is a $9K$-dimensional vector, denoting the mean shape deformation. U is the first 30 eigenvectors given by PCA (Principal Component Analysis). More details can be found in [14].

$$\boldsymbol{q}_k(\boldsymbol{\theta}) = \boldsymbol{\alpha}_k \begin{pmatrix} \Delta A_{p[k]}(\boldsymbol{\theta}) \\ 1 \end{pmatrix} \qquad (2)$$

$$\boldsymbol{s}_k(\boldsymbol{\beta}) = (\boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\mu})_k \qquad (3)$$

IV. METHOD

A. IMAGE SEGMENTATION

Our image segmentation is based on the Grabcut method [13], and we develop a user-friendly interface for segmentation. As Fig. 3 shows, users firstly point out two corners of a rectangle to surround the human body. All the area outside of the rectangle is background. They can add several strokes indicating the foreground area and the background area both before and after Grabcut. When the Grabcut method is completed, we set the background areas darken. Due to the limited screen size of a phone, we design zoom function for more convenient strokes input. What we are concerned about is the perfect dressed-human contour, instead of the perfect segmentation. We compute the contours of the mask image of current segmentation and select the biggest one as human contour. Take the second "Grabcut Result" subimage in Fig. 3 for example, despite some background areas are not darken, we can get the desired result.

B. SILHOUETTES ADJUSTMENT

We cannot regress body parameters directly using segmented human contours on account of unknown camera configurations. When we use the configurations of virtual cameras which are set during training phase, we need to adjust the sizes and locations of human contours in the silhouettes. This section illustrates how we approximately transform our segmented contours to the ones under virtual camera configurations.

As Song *et al.* [27] point out, camera configuration (intrinsic and extrinsic parameters) has an effect on the size and shape of human contours. Nevertheless, perspective projection has little effects on the shape of body contour when the following conditions are satisfied: (1) the camera points approximately to the center of human, (2) the direction of camera view is nearly orthogonal to human plane, and (3) the distance between camera and human is relatively far, compared with which, the thickness of human body can be ignored. These three conditions are easy to be achieved when we take photos with a mobile device.

We follow previous work [27] to compute the size of body contour under the virtual camera configurations. They define

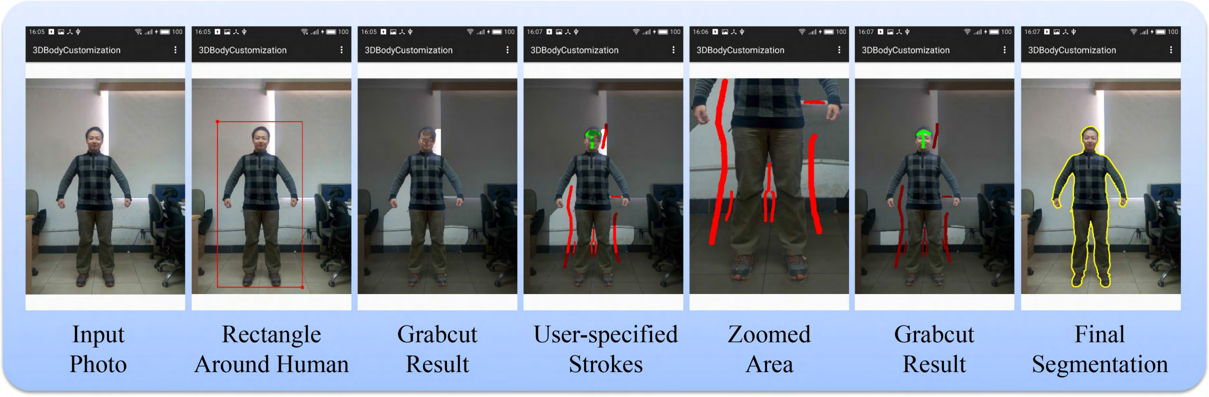


FIGURE 3. ***Image segmentation*. We use Grabcut method to segment the human. Users firstly click the top left and bottom right corners of a rectangle for the human area. Then we get a segmentation result, the darken area of which is the background of current result. According to this result, users can add strokes indicating the human area (green strokes) and the background (red strokes). With both the rectangle and the strokes as constraints, a new result is generated. Finally, we can see the segmented dressed-human contour.**

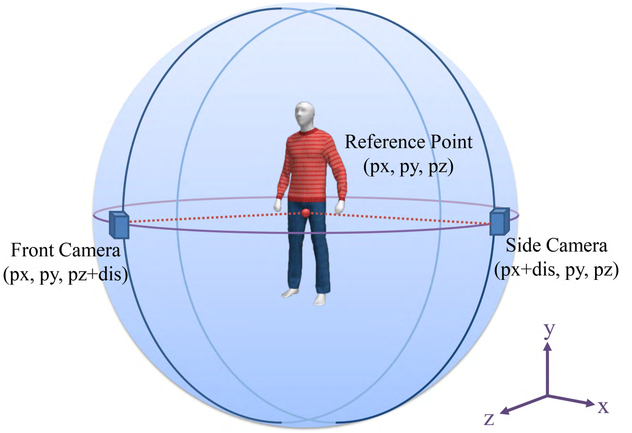


FIGURE 4. ***Virtual camera configuration*. Suppose the position of the reference point is (px, py, pz), the positions of the front and the side virtual cameras are ($px, py, pz + dis$) and ($px + dis, py, pz$). The directions of the front and side cameras both point from camera to the reference point.**

the size of body contour as the height of the bounding box (denoted as BBH). BBW represents the width of the bounding box, and HWR equals to BBH/BBW. They assume that users already know their height information and suppose BBH value is linear to $(h, HWR)^T$. They learn the linear coefficients with ridge regression [28].

Different from [27], we should adjust the position of contour in the silhouette besides the size. Song *et al.* [27] separately predict 2D landmarks from front or side silhouette, whereas we regress body parameters using both front and side silhouettes. Since the configurations of the front and side cameras are fixed and the 3D human body is placed at a fixed position, the positions and the relative relationship of body contours in the front and side silhouettes are determined. In another word, the positions of body contours in the front and side silhouettes should be consistent with each other.

We propose a reference point to solve this problem. As shown in Fig. 4, we select one vertex of human body as the reference point (marked as the red point in the figure). For all the registered bodies in the database, the reference point locates at the same location of the body. Both the front and the side cameras point to this point at a fixed distance. Subsequently, the corresponding projected points in both the front and the side silhouettes locate at the center. We adopt the 2D landmarks regression method [27] to locate the positions of projected reference point in the front and side silhouettes. We adjust the positions of body contours by moving the points to the center.

C. BODY PARAMETERS REGRESSION

In this section, we introduce how we regress body parameters $(\boldsymbol{\theta}, \boldsymbol{\beta})$ which determine a 3D body from front and side silhouettes. The training data are introduced in section III-B, and each training sample for training parameters regression consists of initial body parameters, target body parameters and a pair of silhouettes (in front and side views). We prepare 9 sets of initial parameters whose corresponding heights range from 150cm to 190cm with an interval of 5 centimeters. Initial parameters for each sample are decided by the corresponding height.

We use the boosting tree regression method [11] to regress the residual between initial body parameters and target body parameters according to front and side silhouettes. The residual is decreased little by little through a series of cascaded regressors G_i $(i = 1, 2, \cdots, m$, where m denotes the total number of regressors). The i^{th} regressor G_i takes input silhouettes and the body parameters updated by last regressor G_{i-1} as input, and the relative relationship between current body parameters and silhouettes is used to guide the change of body parameters. We train each regressor with training samples, and the goal of the i^{th} regressor G_i is formulated as:

$$\arg\min_{G_i} \sum_{j=1}^{N} \| \boldsymbol{P}_T^j - (\boldsymbol{P}_{i-1}^j + G_i(\boldsymbol{I}^j, \boldsymbol{P}_{i-1}^j)) \|^2 \tag{4}$$

where, N is the number of training samples. $\boldsymbol{P}_T^j$ denotes the target parameters of training sample j, and $\boldsymbol{P}_{i-1}^j$ is the parameters updated by last regressor. $G_i(\boldsymbol{I}^j, \boldsymbol{P}^j_{i-1})$ regress the residual between current parameters and target parameters according to silhouettes $\boldsymbol{I}^j$ and $\boldsymbol{P}^j_{i-1}$.

We adopt the feature descriptor proposed by previous work [11], which is the displacements from projected landmarks to their nearest dressed-human contour points, to classify training samples. Projected landmarks can be acquired by body parameters, so the feature descriptor also describes the relative relationship between body parameters and silhouettes. G_i in formula (4) is approximately represented by a piecewise function with classification conditions. In the training phase, training samples with similar feature descriptor are classified into the same classification. Take classification Ω_c for example, G_i is learned as:

$$G_i = \frac{\sum_{j\in\Omega_c} \delta \boldsymbol{P}^j_i}{|\Omega_c|} \tag{5}$$

$\delta \boldsymbol{P}^j_i = \boldsymbol{P}^j_T - \boldsymbol{P}^j_{i-1}$. $|\Omega_c|$ is the number of training samples in classification Ω_c.

More details about classification can be found in [11]. Different from [11], we regress body parameters instead of 3D landmarks. We do not need to iteratively optimize body parameters, which improves the efficiency of our method and makes our method possible for mobile devices with limited computation resources.

V. RESULTS AND APPLICATION

A. TIME CONSUMPTION AND SHAPE ESTIMATION ERROR

Our testing device is MEIZU PRO 5 with Android version 5.1. The mobile phone's CPU is Samsung Exynos 7420 and RAM is 4GB. It appeared to the market with price as 2800 RMB since September, 2015. Our parameters regression process costs only 1.26 seconds on this device. We scale the size of photo to fit the size of phone screen, and Grabcut method consumes about 5 seconds on this device. Segmentation usually takes two rounds of Grabcut with added strokes.

12 volunteers including European and Asian with shape variations took part in our experiments. After being captured the front view, the volunteers turn to their left for another shot. The distance between human and camera when taking photos is not restricted to a fixed number. We measure participants' height as input and chest/waist/hip size as ground truth. We use traditional anthropometric method [29] to measure participants. Fig. 5 shows how we compute the chest, waist and hip girth of reconstructed 3D body. The shape variations and mean absolute errors of 12 volunteers are shown in Table. 1. We visually show several examples of our results in Fig. 6.

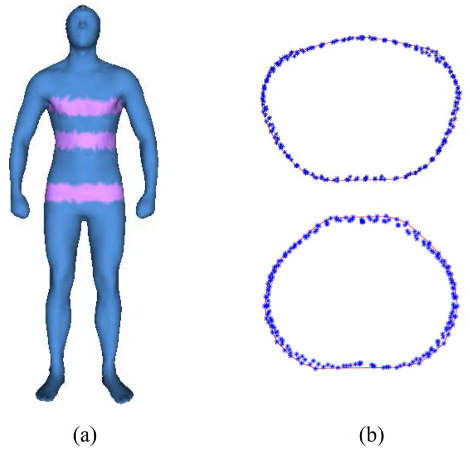


FIGURE 5. ***Girth calculation.*** **(a) Pink points show the vertices used for calculating chest size, waist size and hip size. (b) For each body size, we firstly use related 3-D vertices to fit a plane, and then project these vertices to the plane. We compute the convex hull of projected points and compute the circumference of the convex hull as body size.**

TABLE 1. **Shape variations and mean absolute estimation errors of participants.**

Item	Chest Size	Waist Size	Hip Size
Shape Variation	(86, 116) cm	(75, 112) cm	(95, 112) cm
Mean Absolute Error	2.89 cm	1.93 cm	2.22 cm

In Table. 2, we compare our methods with previous methods proposed for estimating body shapes under clothes. "N.A." in the table is abbreviated for "Not Available". We do not implement their methods on Android devices, so we illustrate the hardware they described in their paper. Due to different conditions of input, we also do not use the same input for the comparison. The results in accuracy of [11] and [22] are better than ours. However, Song *et al.* [11] calibrate camera and segment human using Photoshop whereas Zhu and Mok [22] make users to specify several body points under clothes. We make a trade off between the estimation accuracy and the practicality for ordinary users. Our method is efficient, because we directly regress body parameters, without finding correspondences between the input information and the target body or iteratively solving parameters with constraints.

B. SILHOUETTE ADJUSTMENT ERROR

As we stressed in section IV-B, we adjust the size and position of human contour in the silhouette. We use synthetic data (637 pairs of naked and dressed 3D bodies, which are separate from training data) created by Song *et al.* [11] to test the process of silhouette adjustment. We also use the camera configurations as shown in Fig. 4 to generate testing silhouettes.

As reported in [27], the resolution of testing silhouettes is 800 × 600. The average *BBH* (i.e., height of human bounding box) value of front silhouettes is 362.98 pixels and the average *BBH* value of side silhouettes is 372.26 pixels. The mean absolute estimation errors of front and side contour sizes are 1.58 pixels and 1.63 pixels respectively.

The position of human contour is determined by the location of the projected reference point. We adopt the previous method [27] to regress 2D landmarks which include the

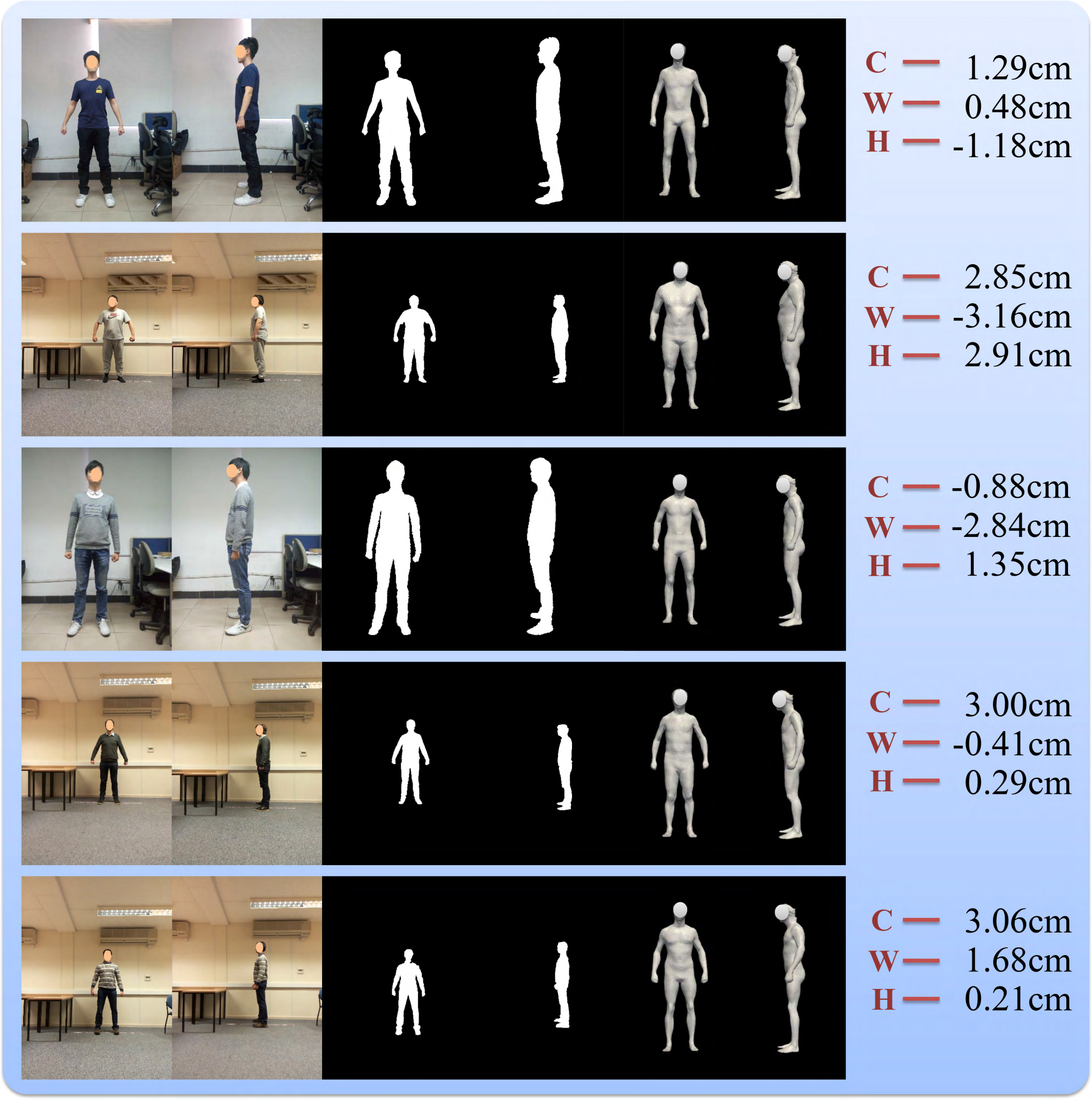


FIGURE 6. ***Several examples of the results.*** **From left to right: photos, segmented silhouettes, estimated 3-D body, estimation errors for chest size(C), waist size(W), and hip size(H).**

projected reference point. We also try some other methods to estimate the location of the projected reference point. In Table. 3, we compare the mean absolute error of projected reference point using three methods: (1) Regressing 39 2D landmarks for front silhouette and 24 2D landmarks for side silhouette as proposed in [27]; (2) Regressing only 1 2D landmark (i.e., the projected reference point); and (3) Predicting the translation from the center of the bounding box to the location of the projected reference point with linear regression. The comparison in Table. 3 show that method (1) performs best, and we suppose it is because more landmarks absorb the global information of the body contour and help locating the location of the projected reference point.

C. BODY RECONSTRUCTION ERROR

We also use the synthetic data to test the 3D body reconstruction error induced by the process of parameters regression. We use the same 637 testing samples generated by Song *et al.* [11], and compute the mean absolute vertex error between estimated body mesh and ground truth body mesh. Fig. 7 visually shows the mean absolute vertex error for 3 clothes types, and we find that they achieve similar performance. The errors of vertices located at calves for clothes type 3 are slightly less than that for the other two types, because clothes type 3 does not cover calves. However, the exposure of arms (clothes type 2 and 3) does not decrease corresponding errors on account of arm pose ambiguity.

TABLE 2. Comparison with previous methods.

Method	Input	Clothes	Camera Calibration	Chest/Waist/ Hip Size Error	Reconstruction Time	Hardware
Balan and Black [9]	4 silhouettes	6-10 kinds	Yes	1.03/4.65/4.73 cm	40minutes	PC (2GHzCPU)
Hasler et al. [15]	1 scan	Casual clothes	3D scanner	N.A./1/N.A. cm (1 person)	11.5 minutes	PC
Chen et al. [10]	1 photo, height	7 kinds	Orthogonal Projection	N.A.	1 minute	PC (3GHzCPU, 2GBRAM)
Wuhrer et al. [30]	1 or more scans	Casual office clothes	Kinect Fusion	17.26/17.62/ N.A. cm	N.A.	PC
Zhu and Mok [22]	2 photos, height, user-specified body points	loose-fit clothes	N.A.	1.06/0.83/0.95 cm	less than 2.8 seconds	PC(dual core 1.86GHzCPUs, 4GBRAM)
Song et al. [11]	2 silhouettes	3 sets	Yes	1.83/1.90/1.85 cm	3.611 seconds	PC (4.0GHzCPU, 32GBRAM)
Ours	2 photos, height	3 sets	No	2.89/1.93/2.22 cm	1.26 seconds	Android Phone

TABLE 3. Mean absolute error of projected reference point.

View Method	Front	Side
(1)	1.66 pixels	0.98 pixels
(2)	2.67 pixels	4.95 pixels
(3)	2.57 pixels	3.80 pixels

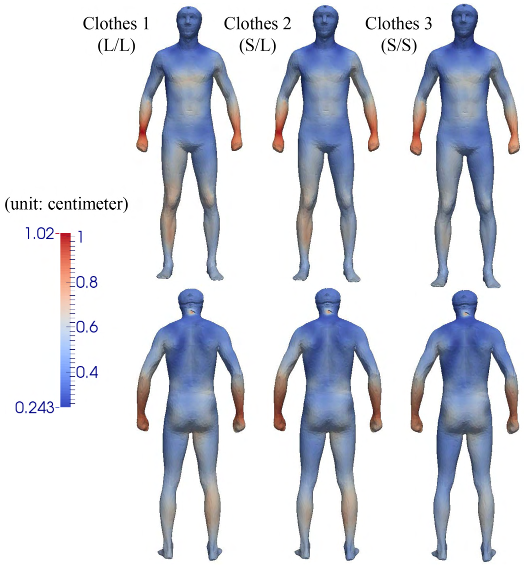


FIGURE 7. ***Mean absolute vertex error.*** The colors of points illustrate the mean absolute vertex error of 637 testing samples.

We further compute the mean absolute errors of chest, waist and hip sizes, and compare our results with [11] in cumulative error distribution. Since three clothes types achieve similar performance, Fig. 8 only shows the results for clothes type 3 for simplification. For our method, about 90% of testing synthetic bodies' chest/waist/hip size error is less than 2/2.2/2.6 centimeters. We directly regress body parameters which determine 3D body, rather than firstly regressing landmarks and then using landmarks as constraints to optimize body parameters as Song *et al.* [11] do. With the same PC, parameters regression only costs 0.509 second whereas previous parameters optimization with landmarks consumes 3.611 seconds (landmarks regression costs 0.028 second and optimization needs 3.583 seconds).

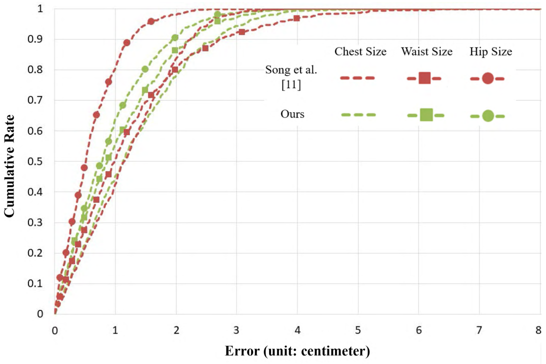


FIGURE 8. ***Cumulative error distribution in body measurements.*** Cumulative error distribution in chest size, waist size and hip size.

D. APPLICATION

In Fig. 9, we show how ordinary users get their customized 3D bodies with our approach and the application of the customized body shape. Users only need a mobile device which is widely available. In the view of users, firstly they are captured the front and side views. Then they input their height information and choose the type of clothes in the photos. Afterwards, they specify several strokes for segmentation. The customized body shapes come out in about 1.26 seconds. With the development of cloth simulation and the determination of the property of various cloth, our customized body shapes can be used for virtual try-on. In the future, when we shop clothes online, we can get valuable visual information and size suggestions. For example, we can see our virtual avatars wearing different clothes and choose the favorite clothes type efficiently. For clothes size suggestions, the color

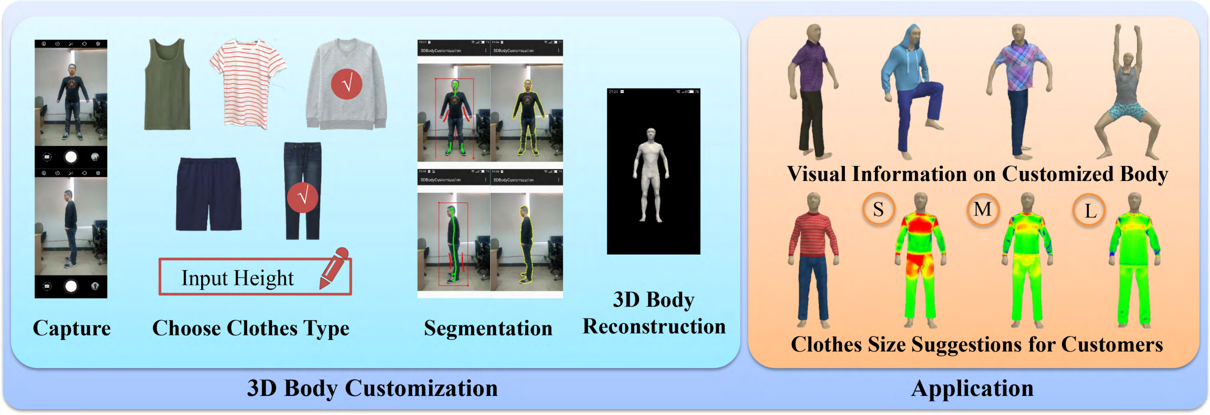


FIGURE 9. ***The process of 3D body customization and the application.*** **Customers only need a mobile device to get their body shapes. They need to input height, choose the type of clothes in the photos and segment human with several strokes. The orange area shows the application of the customized body for virtual try-on. For clothes size suggestions, the color map shows the strain of cloth where red color denotes "tight" and blue color indicates "loose".**

map in the figure indicates the strain of cloth where red color represents "tight" and blue color denotes "loose". It helps users to decide the clothes size and increases the number of successful transactions for online shopping.

VI. CONCLUSIONS, LIMITATIONS AND FUTURE WORK

We develop a data-driven application for ordinary customers to efficiently and conveniently access their 3D body shapes, in order to meet the need of virtual fitting room in online shopping. With photos captured by phone camera, we use Grabcut method to segment human. To achieve good segmentation, we allow users to zoom in and out the image and add strokes indicating human and background. The segmented human contours are scaled according to human height and translated using the regressed reference points in the silhouettes. Then they are transformed to the ones under our virtual camera configurations. We avoid camera calibration and make our approach more convenient. We train a series of regressors to predict body parameters from the adjusted silhouettes. We neither iteratively find correspondences between the 2D points in silhouettes and the 3D vertices on target body mesh, nor iteratively solve least square problems for parameters optimization. Body parameters regression only costs about 1.26 seconds on an android phone. 12 volunteers took part in our testing experiments, and the mean absolute estimation error for chest/waist/hip size is 2.89/1.93/2.22 cm. We make a trade off between the estimation accuracy and the availability for ordinary users.

In the future, we should work out a way to compare our approximate calibration approach with the traditional camera calibration and the automatic calibration using phone camera information and gravity sensor data [12]. Further assessment for the corresponding impacts on body shape estimation of different calibration methods should also be tried. To provide more comfortable experiences for ordinary users, we still have a lot of work to do. More efficient and convenient segmentation technique is in demand. We should also extend our work to more poses, clothes types and camera views so that users can use their photos captured in daily life. Last but not least, the shape and texture of human face play an important role in the reality of virtual fitting [31], so they are also the issues that we should pay attention to in the future.

REFERENCES

[1] K.-L. Cheng, R.-F. Tong, M. Tang, J.-Y. Qian, and M. Sarkis, "Parametric human body reconstruction based on sparse key points," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 11, pp. 2467–2479, Nov. 2016.

[2] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev, "3D self-portraits," *ACM Trans. Graph.*, vol. 32, no. 6, Nov. 2013, Art. no. 187.

[3] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan, "Scanning 3D full human bodies using Kinects," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 4, pp. 643–650, Apr. 2012.

[4] H. Seo, Y. I. Yeo, and K. Wohn, "3D body reconstruction from photos based on range scan," in *Technologies for E-Learning and Digital Entertainment* (Lecture Notes in Computer Science), vol. 3942. Berlin, Germany: Springer-Verlag, 2006, pp. 849–860. [Online]. Available: https://link.springer.com/chapter/10.1007/11736639_105#citeas

[5] Y. Chen and R. Cipolla, "Learning shape priors for single view reconstruction," in *Proc. IEEE 12th Int. Conf. Comput. Vision Workshops (ICCV Workshops)*, Sep./Oct. 2009, pp. 1425–1432.

[6] P. Guan, A. Weiss, A. O. Balan, and M. J. Black, "Estimating human shape and pose from a single image," in *Proc. IEEE 12th Int. Conf. Comput. Vis. (ICCV)*, Oct. 2009, pp. 1381–1388.

[7] J. Boisvert, C. Shu, S. Wuhrer, and P. Xi, "Three-dimensional human shape inference from silhouettes: Reconstruction and validation," *Mach. Vis. Appl.*, vol. 24, no. 1, pp. 145–157, Jan. 2013.

[8] S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han, "Parametric reshaping of human bodies in images," *ACM Trans. Graph.*, vol. 29, no. 4, p. 126, Jul. 2010.

[9] A. O. Bălan and M. J. Black, "The naked truth: Estimating body shape under clothing," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 5303. Berlin, Germany: Springer-Verlag, 2008, pp. 15–29. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-88688-4_2#citeas

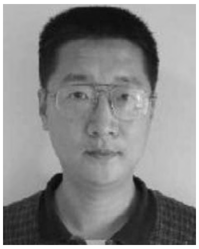
[10] X. Chen, Y. Guo, B. Zhou, and Q. Zhao, "Deformable model for estimating clothed and naked human shapes from a single image," *Vis. Comput.*, vol. 29, no. 11, pp. 1187–1196, Nov. 2013.

[11] D. Song, R. Tong, J. Chang, X. Yang, M. Tang, and J. J. Zhang, "3D body shapes estimation from dressed-human silhouettes," *Comput. Graph. Forum*, vol. 35, no. 7, pp. 147–156, 2016.

[12] A. Ballester *et al.*, "Data-driven three-dimensional reconstruction of human bodies using a mobile phone app," *Int. J. Digit. Hum.*, vol. 1, no. 4, pp. 361–388, 2016.

[13] C. Rother, V. Kolmogorov, and A. Blake, "grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.
[14] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "SCAPE: Shape completion and animation of people," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 408–416, Jul. 2005.
[15] N. Hasler, C. Stoll, B. Rosenhahn, T. Thormählen, and H.-P. Seidel, "Estimating body shape of dressed humans," *Comput. Graph.*, vol. 33, no. 3, pp. 211–216, Jun. 2009.
[16] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormählen, and H.-P. Seidel, "Multilinear pose and body shape estimation of dressed subjects from image sets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 1823–1830.
[17] S. Zhu, P. Y. Mok, and Y. L. Kwok, "An efficient human model customization method based on orthogonal-view monocular photos," *Comput.-Aided Des.*, vol. 45, no. 11, pp. 1314–1332, Nov. 2013.
[18] A. Tsoli, N. Mahmood, and M. J. Black, "Breathing life into shape: Capturing, modeling and animating 3D human breathing," *ACM Trans. Graph.*, vol. 33, no. 4, 2014, Art. no. 52.
[19] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black, "DYNA: A model of dynamic human shape in motion," *ACM Trans. Graph.*, vol. 34, no. 4, 2015, Art. no. 120.
[20] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, 2015, Art. no. 248.
[21] R. W. Sumner and J. Popović, "Deformation transfer for triangle meshes," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 399–405, Aug. 2004.
[22] S. Zhu and P. Y. Mok, "Predicting realistic and precise human body models under clothing based on orthogonal-view photos," *Procedia Manuf.*, vol. 3, pp. 3812–3819, Jul. 2015.
[23] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 1078–1085.
[24] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *Int. J. Comput. Vis.*, vol. 107, no. 2, pp. 177–190, 2014.
[25] C. Cao, Y. Weng, S. Lin, and K. Zhou, "3D shape regression for real-time facial animation," *ACM Trans. Graph.*, vol. 32, no. 4, 2013, Art. no. 41.
[26] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel, "A statistical model of human pose and body shape," *Comput. Graph. Forum*, vol. 28, no. 2, pp. 337–346, 2009.
[27] D. Song *et al.*, "Clothes size prediction from dressed-human silhouettes," in *Next Generation Computer Animation Techniques*. Cham, Switzerland: Springer, 2017, pp. 86–98.
[28] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.
[29] M. Kouchi, "Anthropometric methods for apparel design: Body measurement devices and techniques," in *Anthropometry, Apparel Sizing and Design*. Amsterdam, The Netherlands: Elsevier, 2014, pp. 67–94.
[30] S. Wuhrer, L. Pishchulin, A. Brunton, C. Shu, and J. Lang, "Estimation of human body shape and posture under clothing," *Comput. Vis. Image Understanding*, vol. 127, pp. 31–42, Oct. 2014.
[31] C. Cao, H. Wu, Y. Weng, T. Shao, and K. Zhou, "Real-time facial animation with image-based dynamic avatars," *ACM Trans. Graph.*, vol. 35, no. 4, 2016, Art. no. 126.

DAN SONG received the B.S. degree from Shandong University, China, in 2013. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Zhejiang University, China. Her research interests include 3-D human body models, parametric human body reconstruction, virtual fitting, and image processing.



RUOFENG TONG received the B.S. degree from Fudan University, China, in 1991, and the Ph.D. degree from Zhejiang University, China, in 1996. He is currently a Professor with the Department of Computer Science and Technology, Zhejiang University. His research interests include image and video processing, computer graphics, and computer animation.

JIANG DU received the B.S. and M.S. degrees from Dalian Polytechnic University, China, in 2009 and 2012, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Zhejiang University, China. His research interests include RGBD-based 3-D face reconstruction, editing, and 3-D face recognition.

YUN ZHANG received the B.S. and M.S. degrees from Hangzhou Dianzi University, China, in 2006 and 2009, respectively, and the Ph.D. degree from Zhejiang University, China, in 2013. He is currently an Associate Professor with the Zhejiang University of Media and Communications. His research interests include image/video editing and computer graphics.

YAO JIN received the B.S. degree in apparel engineering and the M.S. degree in computer science from Zhejiang Sci-Tech University, China, in 2007 and 2010, respectively, and the Ph.D. degree from Zhejiang University, China, in 2015. He is currently a Lecturer with the Department of Digital Media Technology, Zhejiang Sci-Tech University. His research interests include computer graphics, digital geometry process, and digitalization technology in apparel and textile.