Shape Customisation of Human Subjects Based on Human Parsing Technology

Shuaiyin ZHU[2], Yanghong ZHOU[2], K.P. CHAU[2], P.Y. MOK*[1,2]
[1] The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China;
[2] Institute of Textiles and Clothing, The Hong Kong Polytechnic University, hunghom, Hong Kong

DOI: 10.15221/17.288 http://dx.doi.org/10.15221/17.288

Abstract

Human modelling is an active research topic in medical, fitness and entertainment (such as game and movie) related studies. Many methods have been reported in the literature for human modeling, wjocj can be classified as scan-based, image-based and example-based approaches and all have pros and cons. In this paper, we propose a complete automatic method for human model shape customization based on the cutting-edge human parsing technology. The method allow users efficiently customize human models using orthogonal-view images in a complete automatic pipeline. We have demonstrated that our automatic customisation method is robust that generates accurate digital models for individuals using two images, and the method applies to male and female subjects, in tight-fit or arbitrary (normal and/or loose-fit) clothing, with different background conditions of the input images. Experimental results have demonstrated that our method improves accuracy, robustness, efficiency, and flexibility of human shape modelling.

Keywords: Human modelling, Automatic human shape customisation, Modelling of dressed subjects, Human parsing, Deep learning

1. Introduction

Modeling of digital humans is an important yet challenging topic that has many research applications. For example, the gaming and film entertainment industries often need to create massive 3D virtual characters, whereas accurate representations are needed in order to simulate human reactions to different environments or workplaces for health and safety research. Moreover, accurate 3D models of individuals are used in various product designs so that the resulting products are fit for the target users' needs. In the fashion industry, an accurate 3D human model is the starting point for the development of any well-fitting garments. Generally speaking, whereas research into entertainment applications mainly focuses on the realistic appearance of the resulting models, ergonomics- and fashion-related applications focus more on the accuracy of the resulting models, in terms of sizes and shapes.

Many methods have been reported in the literature regarding human modeling, including scan-based, image-based and example-based approaches. Scan-based methods directly construct the skin surface of human subjects from different sources, making it possible to construct an accurate body shape of an individual within a few seconds. However, the necessary involvement of expensive and bulky equipment limits the application of such techniques. Other methods were reported in the literature with which to reconstruct the surface model of an individual by deforming a template, based on input measurements [3] or images [1-2]. However, the resulting models obtained via these reconstructive methods have an unrealistic appearance and shape errors due to the oversimplified 2D-to-3D shape deformations [3]. Example-based methods were then proposed that involve the learning of shape deformations from example scans. Such example-based methods often use a registration process to register a specific template model to all examples contained in a dataset [4-9]. After registration, all example mesh models possess the same tessellation, enabling the transformation of every triangle face can be modelled. Dimension reduction techniques such as PCA are often used to extract the deformation control parameters. Example-based methods involve the training of a prediction model with which to deform a template, using extracted control parameters. The resulting parametric deformable model can then be deformed into various shapes with different control parameter values. As example-based methods generate models with a realistic appearance, they are widely used in graphics and vision applications. However, although resulting models exhibit realistic 'looks' as they are in principle averaged from examples, they do not capture accurately the local shape features of individuals, such as the shoulder slope or the body curve. The resulting average figures thus cannot produce accurate sizes of individuals, which restricts the application of such example-learnt parametric deformation models in the fields of fashion and ergonomics.

* tracy.mok@polyu.edu.hk; +852- 2766 4442.

Zhu et al. [10] developed novel methods that customize accurate 3D models using two orthogonal-view images of the subject. They manually extracted 2D body features from the input images, reconstructing 3D shape representations using the extracted 2D features and 2D-to-3D relationship models learned from a large database of scan models. They also developed a deformation algorithm with which to deform a high-resolution template mesh into a customized shape, based on the subject's 3D shape representation. They then applied this model customization method to obtain accurate models using images of subjects dressed in tight-fitting clothing [10] and loose-fitting clothing [11]. However, although their method is able to capture detailed local shape characteristics of individuals, tedious manual work is required to extract features from images, while the customized models are available only in standard static poses.

2. Method

2.1. Method overview

In this paper, we propose a complete automatic method for human model shape customisation, as shown in Fig. 1. The inputs to the method are front- and side-view image of a human subject, and the output is a 3D human body model in customised shape of the subject in the input images. The key novelty of our method lies in the segmentation of human subjects from complex image backgrounds using the cutting-edge human parsing technology. We trained a Convolution Neural Network (CNN) for human parsing. The outputs from the parsing step are images with pixels labeling in different body parts, such as the face, hair, clothing, shoes, and hands. We use the initial parsing results to determine the foreground (human subject) from the background in a refined segmentation step. The extracted raw contours are then used to calculate the under-the-clothes profiles of the subject, from which a 3D shape representation of the subject is constructed that later guides model customisation in the last step.

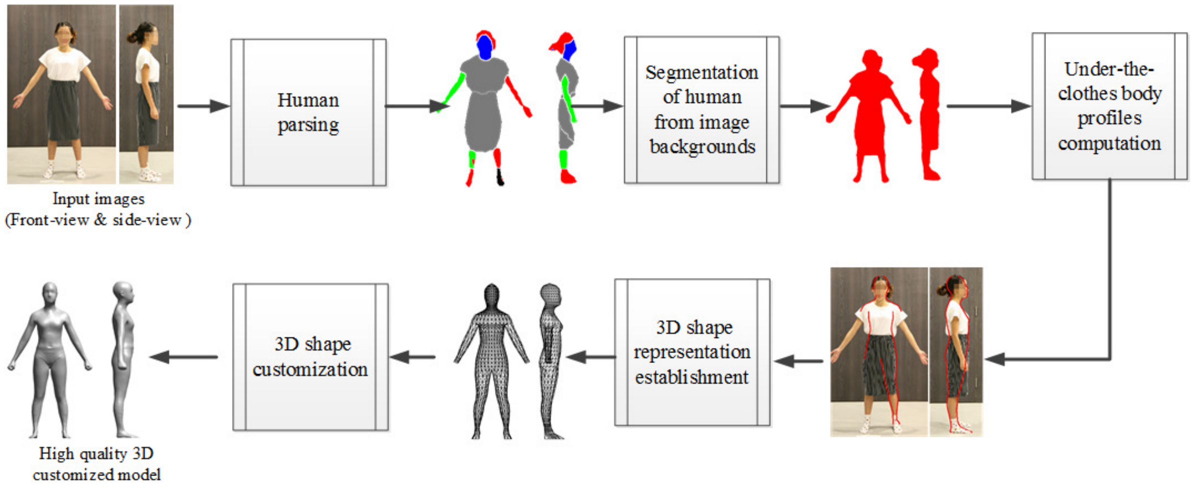
Input images
(Front-view & side-view)
Human parsing
Segmentation of human from image backgrounds
Under-the-clothes body profiles computation
3D shape representation establishment
3D shape customization
High quality 3D customized model

Fig. 1. Method overview

2.2. Human parsing-based segmentation

Image segmentation is a typical problem of computer vision. In general, segmentation refers to a typical process of partitioning an image into different regions such that each region is, but the union of any two adjacent regions is not, homogeneous [12]. Image parsing sometimes also refers to 'semantic segmentation', which is the task of partitioning the image into semantically meaningful parts and classifying each part into one of the pre-determined categories. In other words, image parsing means segmentation plus perceptual grouping plus object recognition. As a branch of image parsing, human parsing focuses on understanding human photos. There are two kinds of human parsing representations: parselet and pixel-label. The former mainly focuses on estimating human pose from images; the latter means assigning human-related pixels with different labels. Yamaguchi et al. [13] proposed solving the labeling problem of human parsing with the super-pixel concept [15] and conditional random field (CRF) model [14],[16] aligned human parts by using the parselet representation as the building blocks of a parsing model. Parselets are a group of parsable segments that can generally be obtained by low-level over-segmentation algorithms. They built a deformable mixture parsing model (DMPM) for human parsing to simultaneously handle the deformation and multimodalities of parselets.

The state-of-the-art technology of semantic segmentation was developed by Long et al. [17], who provide an end-to-end solution (pixel-to-pixel) using fully convolutional networks (FCNs).

In this paper, we adopted the architecture of the state-of-the-art 16-layer classification network model VGG-16 [18], but replaced the fully connected layers by 1x1 convolution layers [17]. Moreover, we up-sample the scores of the network according to the original image resolution for more fine-grained results. Based on this network architecture, we train the network using images annotated with semantic parts at pixel level by minimizing the average cross-entropy loss over all image pixels with Stochastic Gradient Descent (SGD). To start the training process, we initialized the network parameters using the ImageNet pre-trained VGG-16 model [18].

2.3. Profile prediction

The deep convolution neural network described above automatically annotates the input image pixels into different labels. With the parsing results, we determine the foreground (human body) and the background areas of the input images, and implement the Grab-Cut algorithm for refined segmentation to obtain a detailed raw contour.

The detailed implementation is as follows. For efficient computation, the input images to the deep neural network for human parsing are set to a predefined small resolution. However, for model customization application, high-resolution input images should be used for refined raw contours. Therefore, we scaled the coarse parsing results to obtain both an inner mask and outer mask, corresponding to the foreground and background of the images, respectively. Besides, small discontinued pixels of the scaled human-parsing results are filtered to remove the noises. Next, we used the Grab-Cut algorithm to segment on the overlapped possible zone of the original high-resolution images for the raw contours. We called this process a refined segmentation. Fig. 2(b) and (c) show the parsing result and the raw contour after refined segmentation for the input image shown in Fig. 2(a).

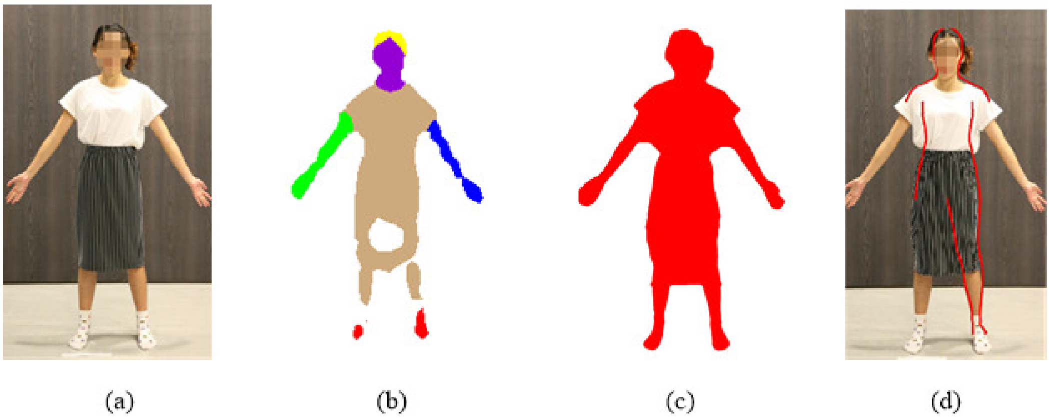

Fig. 2. An example of human parsing and refined segmentation

2.4. Computation of under-the-clothes profiles

Raw contours segmented from images are clothing edges. We develop an algorithm to compute the under-the-clothes body profiles (or body contours) of the subject as a 2D body shape representation. To this end, we first establish a under-the-clothes profile database $\{\mathbf{P}_1,\ldots,\mathbf{P}_{N^p}\}$ from scan data, with method similar to [11], which each body profile $\mathbf{P}_i$ is define in standard parameterization composed of n data points with even vertical spacing. We next transform the raw contours to the same parameterization as profile $\mathbf{P}_i$, and the transformed contour is denoted as **B**. We sort out a small number (χ) of profiles from the under-the-clothes profile database $\{\mathbf{P}_1,\ldots,\mathbf{P}_{N^p}\}$, each profile is selected by:

$$\underset{\mathbf{P}}{\operatorname{argmin}} \sum \left\| \mathbf{f}_k - \mathbf{v}_k^{\mathbf{P}_i} \right\|^2 + \varphi(\mathbf{P}_i) \quad (1)$$

where f_k is a point feature $\mathbf{f}_k \in \{\mathbf{f}_1,\ldots,\mathbf{f}_{N^f}\}$ extracted from the raw contours **B**, $\mathbf{v}_k^{\mathbf{P}_i}$ represents the corresponding point on profile $\mathbf{P}_i$, and $\varphi(\mathbf{P}_i)$ is a penalty function to filter profiles of different gender as follows

$$\varphi(\mathbf{P}_i) = \begin{cases} 0 & \text{gender of } \mathbf{P}_i \text{ is the same as the target;} \\ \infty & \text{gender of } \mathbf{P}_i \text{ is different than the target.} \end{cases} \quad (2)$$

We used Equations (1) and (2) to search in the profile database for under-the -clothes profiles with similar shape characteristics, defined by feature set $\{\mathbf{f}_1,\ldots,\mathbf{f}_{N^f}\}$, as the extracted contours.

After identifying a number χ of the most similar profiles from $\{\mathbf{P}_1,\ldots,\mathbf{P}_{N^p}\}$, we synthesized under-the-clothes profile $\mathbf{P}'$ for the subject by interpolating these χ most similar profiles:

$$\mathbf{P}' = \frac{\sum_{k=1}^{\chi} \beta_k \cdot \mathbf{P}_k}{\sum_{k=1}^{\chi} \beta_k} \tag{3}$$

where β_k is the weight for profile $\mathbf{P}_k$, and β_k is defined by the similarity of features $\{\mathbf{f}_1,\ldots,\mathbf{f}_{N^f}\}$ of $\mathbf{B}$ and the corresponding features of profile $\mathbf{P}_k$:

$$\beta_k = \sum_{j=1}^{N^f} \left\| \mathbf{f}_j - \mathbf{v}_j^{\mathbf{P}_k} \right\| \tag{4}$$

With reference to Equations (1) and (3), we can conclude that the effective identification of similar profiles and the later profile synthesis is determined by the selection of feature points $\{\mathbf{f}_1,\ldots,\mathbf{f}_{N^f}\}$. Zhu and Mok [11] manually selected feature points to construct under-the-clothes profiles, and thus the resulting profiles vary depending on the selected features. In this paper, we proposed an algorithm (Table 1) to automatically select feature points $\{\mathbf{f}_1,\ldots,\mathbf{f}_{N^f}\}$ from the extracted raw contours. The rationale of the algorithm is to define some sample profiles $\{\mathbf{S}_1,\ldots,\mathbf{S}_{N^S}\}$ from the profile database, according to Body Mass Index (BMI). Next, we look for profiles in the database with similar shape characteristics, in form of characteristic vector $\mathbf{c}_j = [c_{j1} \ldots c_{jm}]^T$ of the subject.

Table 1. Feature point selection algorithm

1. Calculate the area of the raw contour region, $A(\mathbf{B})$
2. Select two profile samples $\mathbf{S}_a$ and $\mathbf{S}_b$ from the sample set $\{\mathbf{S}_1,\ldots,\mathbf{S}_{N^S}\}$ as follows:

$$\arg\min_{\mathbf{S}} \left\| A(\mathbf{S}_a) - A(\mathbf{B}) \right\|^2, \text{ if } A(\mathbf{S}_a) - A(\mathbf{B}) > \varepsilon \tag{5a}$$

$$\arg\min_{\mathbf{S}} \left\| A(\mathbf{S}_b) - A(\mathbf{B}) \right\|^2, \text{ if } A(\mathbf{S}_b) - A(\mathbf{B}) < -\varepsilon \tag{5b}$$

where ε is a small positive area threshold.

3. For each point k (($1 \le k \le n$) of raw contour $\mathbf{B}$, calculate the selection function $\mathbf{I}()$ based on the characteristic matrix $\mathbf{C}$:

$$I(\mathbf{c}_k^{\mathbf{B}}) = \begin{cases} 1 & \text{if} \min(\mathbf{c}_k^{\mathbf{S}_a}, \mathbf{c}_k^{\mathbf{S}_b}) < \mathbf{c}_k^{\mathbf{B}} < \max(\mathbf{c}_k^{\mathbf{S}_a}, \mathbf{c}_k^{\mathbf{S}_b}); \\ 0 & \text{else.} \end{cases} \tag{6}$$

4. Select points on the raw contours $\mathbf{B}$ with selection function $\mathbf{I}(\cdot) = 1$, and record it as feature set $\{\mathbf{f}_1,\ldots,\mathbf{f}_{N^f}\}$.

The above algorithms automatically identify the sections of the raw contours that closely follow the shape/contours of the human figure as feature points $\{\mathbf{f}_1,\ldots,\mathbf{f}_{N^f}\}$, excluding the points of the raw contours that are not close to the body, for example the clothing edge. The selected $\{\mathbf{f}_1,\ldots,\mathbf{f}_{N^f}\}$ can be where the clothing hangs on the body, such as the shoulder, hip, and belly, and these sections of raw contours follow the body silhouette.

2.5. 3D shape representation establishment and shape customisation

The predicted under-the-clothes profiles cover the entire body from floor to head. We define a new 3D shape representation covering the feet and head (see Fig. 1) with the method similar to [11]. The new shape representation is aligned with the 2D body profiles and describes the body shape in a more comprehensive way. The features of the ankle and neck can be accurately recognized from full body profiles using anthropometric knowledge. The new 3D shape representation covers all 30 feature layers of the previous definition, plus 10 more layers for the head and 2 more layers for the feet, as shown in Fig. 1.

We trained relationship models for all feature layers of the 3D shape representation using a large database of over 10000 scan models. These models were used to predict the 3D shape of each feature layer using 2D information provided by the under-the-clothes profiles. These are resampled as a unique shape representation of the subject in the input images.

With the customized 3D shape representation, we deformed a high-resolution 3D template model using the ct-FFD deformation algorithm [10] for model customization. Fig. 1 shows a customized model and the input images. It can be seen that the appearance of the customized model is realistic. Moreover, the pose and body curve of the customized model follow the body silhouette of the images closely. Such shape and posture customization technology can be applied in different disciplines. For example, in the clothing industry, the standing posture and body curves affect the pattern design in bespoke or tailor-made garments. In ergonomic applications, products can be designed to fit individual users' or patients' needs, for example chairs design or spine correction.

3. Experimental results

We evaluated the effectiveness of the model customization method by recruiting 40 human subjects, including 18 male and 22 female subjects. Customized models were created using the front-view and side-view photos of these 40 clothed subjects. Fig. 3 and Fig. 4 show some examples of customized models for male subjects and female subjects, respectively. In these figures, columns (a) and (e) are the front-view and side-view input images; columns (b) and (f) are the segmented raw contours overlain with computed under-the-clothes profiles in red; columns (c) and (g) are the customized models, which are compared against the corresponding scans shown in columns (d) and (h).

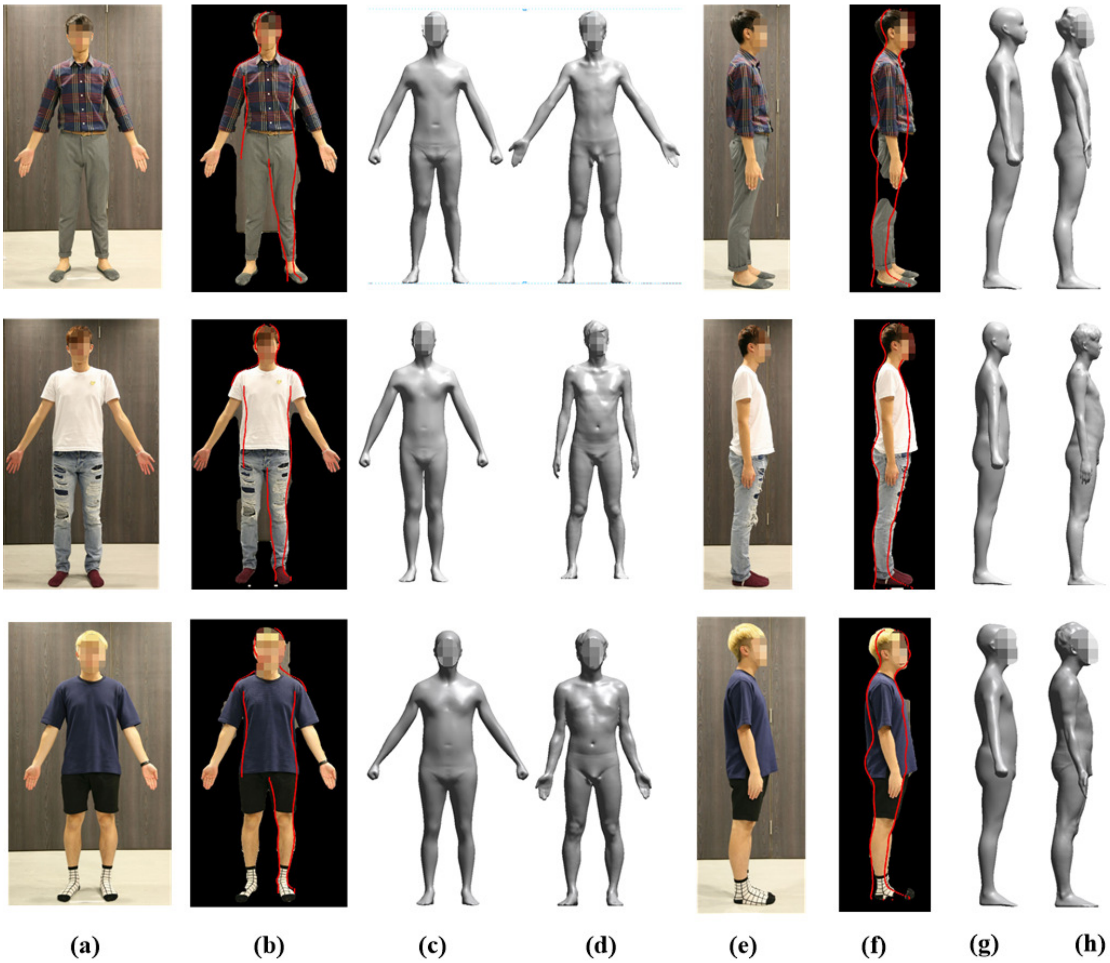

Fig. 3. Some customised model examples of male subjects: (a) and (e) input images, (b) and (f) predicted under-the-clothes body profiles, (c) and (g) front and side view of customised models, (d) and (h) front and side view of scanned models for comparison.

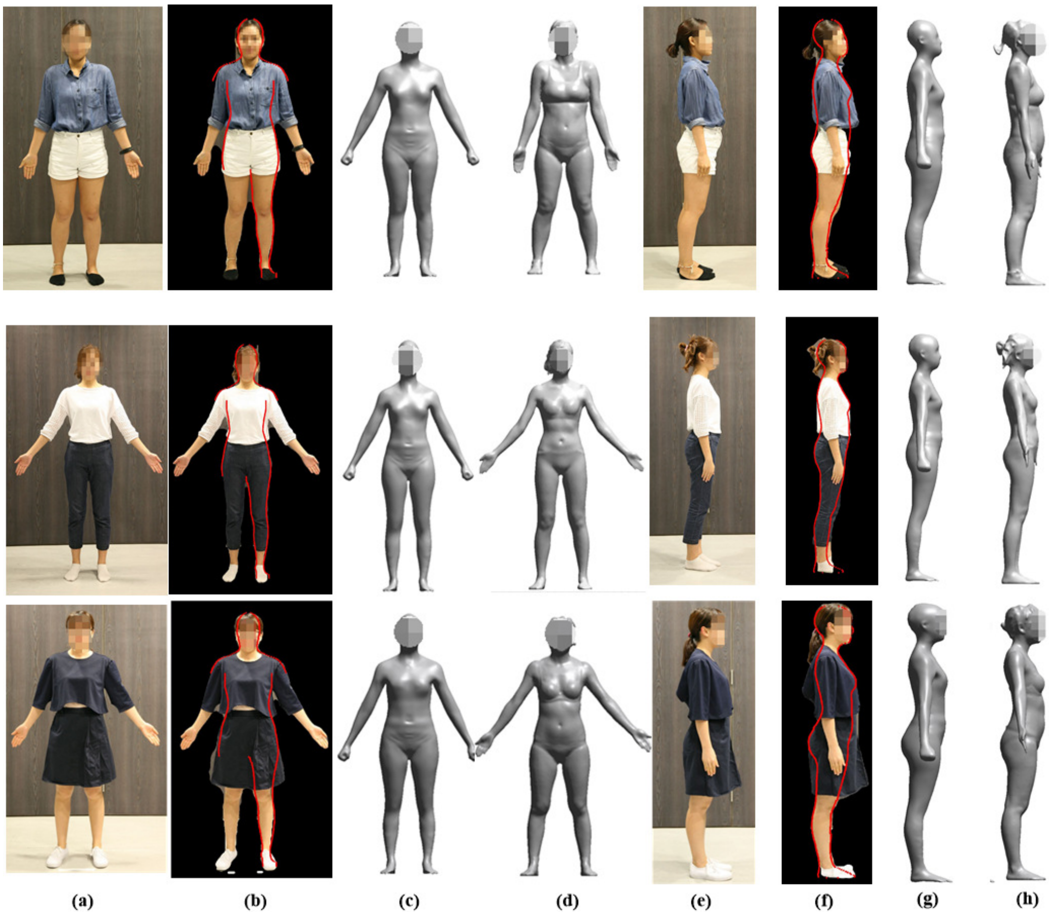

Fig. 4. Some customised model examples of female subjects: (a) and (e) input images, (b) and (f) predicted under-the-clothes body profiles, (c) and (g) front and side view of customised models, (d) and (h) front and side view of scanned models for comparison.

It is found that, in general, the raw contours have been segmented accurately, except for the first male subject of Fig. 4. The incomplete side-view raw contours of this case are probably because of the color similarity between the clothing and the background. Moreover, by comparing the customized models and the scans, we can conclude that the resulting customized models have similar shapes to the subjects in the images. However, it is important to note that it may not be possible to recover local shape characteristics from images if the body silhouette has been covered too much. For example the second male subject of Fig. 4, since his back waist was completely covered by his shirt in the side-view photo, the waist shape of the customized model is different from his scanned model. Therefore, we will provide users with guidelines for suitable clothing to take photographs based on the different accuracy criteria of the customized models. Examples include wearing single-layer clothing and tidying up the clothing at the waist level before taking photos for more accurate customized models.

We also examined the customization results for subjects with special body shapes. Some of the resulting models are shown in Fig. 5, in which the computed under-the-clothes profiles are overlain on the segmented raw contours. As shown, the under-the-clothes profiles closely follow the body silhouettes of the subjects in the input images, correctly projecting the detailed shape characteristics of the subjects, such as the shoulder slope, the curvature of the back, abdomen and hips. The customized models have a natural and realistic shape appearance, with correct global and local shape characteristics of the subjects. For the global shape characteristics, we refer to the overall shape and proportion of the resulting models, e.g. the leg/torso or waist/height ratio, both of which are similar to those of the scanned models. The local shape characteristics refer to local shape features of the customized models, such as the shoulder slope, back curvature and hip curvature. For example, the first (female) and the third (male) subjects have a flat shoulder slope, as shown in the front-view photos. The second (male) subject has a humpback. The customized models reveal the flat shoulder of the first (female) subject and the third (male) subject with reference to the front-view photos, and the humpback of the second (male) subject with reference to the side-view photo, and the distended bellies of the third and fourth (male) subjects with reference to the side-view photos. The ability to model the local shape characteristics of individuals is fundamentally important, especially for clothing-related applications.

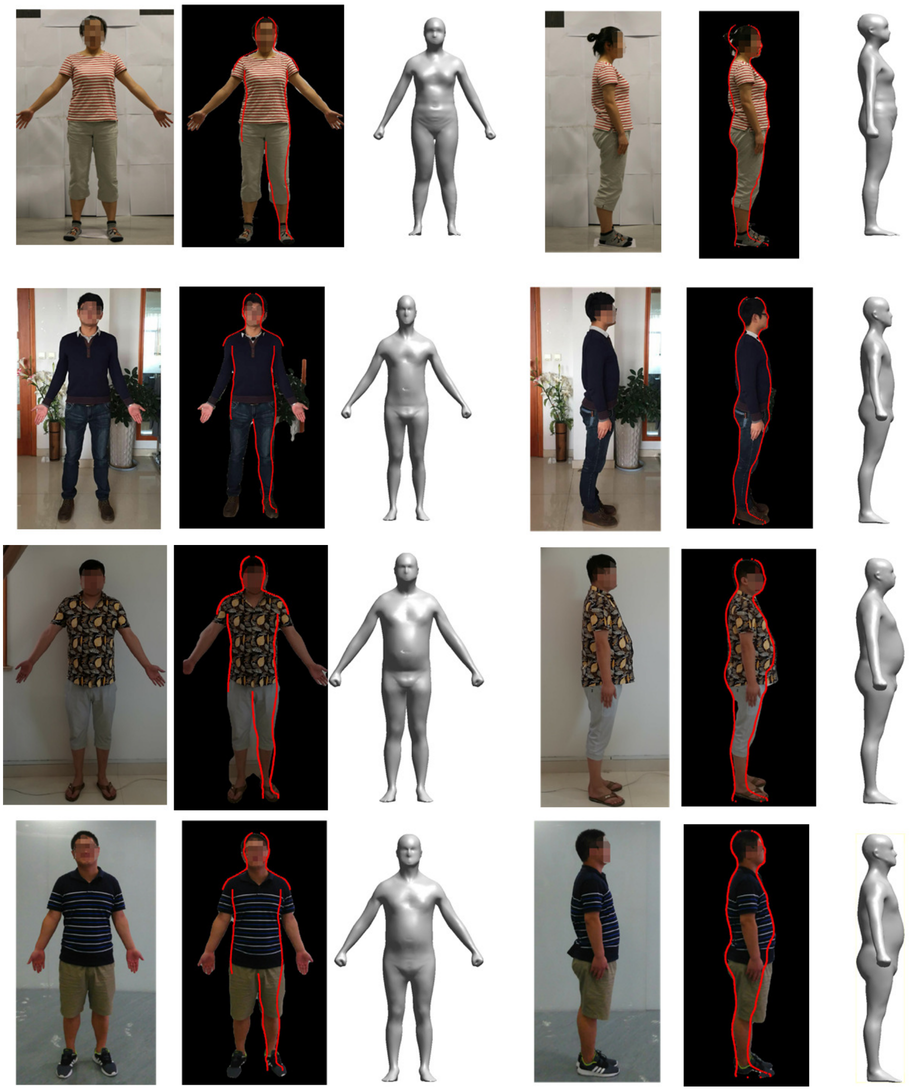

Fig. 5. Four examples of customised models with special body shapes: (a) and (e) input images, (b) and (f) predicted under-the-clothes body profiles, (c) and (g) front and side view of customised models, (d) and (h) front and side view of scanned models for comparison.

4. Conclusions

In this paper, we have developed an automatic 3D human model customization method based on human parsing technology. We trained a VGG16 neural network for parsing human images, in which each input human image can be parsed into 18 labels. For the specific application of human model customization, we combined some labels and used the parsing results to segment the human subject (foreground) from the image background. The parsing-based segmentation approach can robustly and accurately segment the human from the image background.

Based on the parsing results, we have a refined segmentation by Grab-Cut algorithm to obtain detailed raw contour. Next, we automatically determine key feature points along the raw contours. Under-the-clothes profiles of the subject are computed using the raw contour. A 3D shape representation of the subject is constructed from the computed under-the-clothes profiles, and this shape representation drives the deformation of a template model into a customized shape.

A systematic experiment has been conducted to evaluate the proposed human modelling method. The experimental results show that both the size and overall shape accuracy of the resulting models are close to the ground truths. The size measurements critical to the clothing industry have less than 3% difference with the scanned models, satisfying the size tolerance of the clothing industry. In conclusion, the proposed parsing-based human modelling provides an efficient and robust solution to customize human models using orthogonal-view images.

Acknowledgments

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 5218/13E). The partial support of this work by The Hong Kong Polytechnic Univeristy (Project No. G-UA9L), Guangdong Provincial Department of Science and Technology (Project No. R2015A015) and Shenzhen Science and Technology Innovation Commission (Project No. R2017A001), are gratefully acknowledged.

References

[1] Hilton A., Beresford D., Gentils T., Smith R., Sun W., & Illingworth J. (2000). Whole-body modelling of people from multiview images to populate virtual worlds. *Visual Computer*, 16(7): 411-36.

[2] Wang, C.C.L., Wang Y., Cheng T.K.K., & Yuen M.M.F., (2003a). Virtual human modeling from photographs for garment industry, *Computer-Aided Design*, 35(6), 577-589.

[3] Wang, C.C.L. (2005) Parameterization and parametric design of mannequins, *Computer Aided Design*, 37(1), 83–98.

[4] Sloan, P.-P. J., Rose, C.F. and Cohen, M.F. (2001). Shape by example, i3D - Interactive 3D Graphics and Games, pp. 135–143.

[5] Wang, X.C. & Phillips, C. (2002) 'Multi-weight enveloping', Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation - SCA'02, p. 129.

[6] Mohr, A. & Gleicher, M. (2003). Building efficient, accurate character skins from examples, ACM Transactions on Graphics, 22(3), p. 562.

[7] Allen, B., Curless, B. & Popović, Z. (2003). The space of human body shapes: reconstruction and parameterization from range scans, ACM Transactions on Graphics TOG, 22(3), pp. 587–594.

[8] Seo, H. & Magnenat-Thalmann, N. (2004) An example-based approach to human body manipulation, Graphical Models, 66(1), 1–23.

[9] Anguelov, D., Srinivasan, P., Pang, H. & Koller, D. (2004) The Correlated Correspondence Algorithm for Unsupervised Registration of Nonrigid Surfaces, Nips, pp. 33–40.

[10] Zhu, S., Mok, P. Y., & Kwok, Y. L. (2013). An efficient human model customization method based on orthogonal-view monocular photos, CAD Computer Aided Design, 45(11), pp. 1314–1332.

[11] Zhu, S. & Mok, P. Y. (2015). Predicting Realistic and Precise Human Body Models Under Clothing Based on Orthogonal-view Photos, Procedia Manufacturing. Elsevier B.V., 3(Ahfe), pp. 3812–3819.

[12] Cheng, H.D., Jiang, X.H., Sun, Y., & Wang, J. (2001). Color image segmentation: advances and prospects. Pattern recognition, 34(12): 2259-2281.

[13] Yamaguchi, K., Kiapour, M. H., Ortiz, L. E., & Berg, T. L. (2012). Parsing clothing in fashion photographs. Paper presented at the Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.

[14] Boykov, Y., & Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE transactions on pattern analysis and machine intelligence, 26(9), 1124-1137.

[15] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence, 34(11), 2274-2282.

[16] Dong, J., Chen, Q., Xia, W., Huang, Z. & Yan, S. (2013), A deformable mixture parsing model with parselets, Proceedings of the IEEE International Conference on Computer Vision, pp. 3408–3415.

[17] Long J, Shelhamer E, & Darrell T. (2015). Fully convolutional networks for semantic segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3431-3440.

[18] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.