ShapeMate: A Virtual Tape Measure

Alexandros NEOPHYTOU*, Qizhi YU, Adrian HILTON
Centre for Vision, Speech and Signal Processing, University of Surrey, Surrey, UK

http://dx.doi.org/10.15221/13.127

Fig. 1. Body shape estimation from a single image.

Abstract

We introduce ShapeMate, a framework for human body shape estimation and classification for on-line fashion applications. Given a single image of a subject our framework is able to simultaneously estimate detailed 3D human body shape and compute foreground segmentation with minimal user input. Once the body shape has been estimated, various semantic parameters are extracted for garment size and style recommendation. Preliminary results demonstrate that a single image holds enough information for accurate shape classification.

Keywords: body shape estimation, classification, on-line fashion

1. Introduction

Obtaining the 3D body shape of an individual is a long-standing and active research topic with applications varying from entertainment and e-commerce to medical analysis. Recent work has demonstrated that detailed 3D human shape can be accurately estimated from multiple images or depth maps. In this work, we address the problem of robustly and accurately estimating 3D human body shape from a single image for use in on-line fashion applications. Our goal is to develop a cheap and easy to use system that will allow users to accurately extract body measurements using a mobile phone or tablet.

Traditionally, shape reconstruction techniques require foreground segmentation to be computed separately. Unlike existing methods [1, 2, 3, 4, 9], our approach is able to perform integrated segmentation and shape estimation using a novel image cost function which incorporates a strong human shape prior. This allows for minimal user interaction which enhances user experience by hiding the complexity of system from the user. Our method is based on a statistical model of the human body shape which is learned from a range of 3D scans of different individuals.

We demonstrate our results by estimating the shape of known subjects from frontal view photographs and comparing with ground truth measurements obtained from a 3D scanner.

*a.neophytou@surrey.ac.uk

2. Related Work

Estimating human shape from a single image is a challenging under constrained problem. To make the problem tractable, prior knowledge can be used to sufficiently constrain the problem. Statistical human models are used that parameterize a database of registered body scans to generate a low-dimensional space which encapsulates the statistical variation within the database. The parameters of such models can then be optimized to find the best fit to a single image. If the same subject is shown in multiple images, then human body parameters can be optimized across multiple images.

Balan *et al*. [2] used the popular SCAPE model to estimate detailed human body shape from silhouettes by formulating the problem as an optimization over the SCAPE model parameters. This approach however requires a manual initialization step to bootstrap estimation. In order to avoid this problem, Sigal *et al*. [9] proposed a direct probabilistic mapping between monocular silhouette contour features and the SCAPE parameters to automatically initialize the stochastic optimization.

In order to further improve shape estimation, other cues from an image may be considered. Balan *et al*. [1] extend their model to consider cast shadows on the ground plane and a light source which act as additional constraints especially in the case of monocular scenes. Similarly, Guan *et al*. [4] estimate detailed human shape from silhouette by incorporating shading cues and internal edges. The disadvantage of both approaches is that their use is restricted to carefully controlled in-door environments and under the assumptions that only one light source is applied to the model. Furthermore, the shading cues restrict them to naked subjects.

In a different approach [3] a probabilistic framework was proposed for inferring 3D shape parameters from silhouettes. The shape and pose variation are modeled as two separate Gaussian Process Latent Variable Models (GPLVMs).

The methods above either require a manually segmented image or sparse correspondence between image points and a template 3D model. Our framework tries to simultaneously produce an estimate of the 3D human shape and compute the foreground segmentation by minimizing a single objective function.

3. Framework pipeline

To make use of our framework possible on hand-held devices we have developed a client-server based system with the basic work flow shown in figure 2. In order to assist the shape reconstruction process, the user is asked to provide the subject's height and weight. Furthermore, increased accuracy can be achieved by manually specifying the positions of the hands and feet. We require the subject in the photo to be facing towards the camera in a relaxed pose with straight arms and legs. The system can cope with skirts and tight dresses provided the arms and legs are visible.

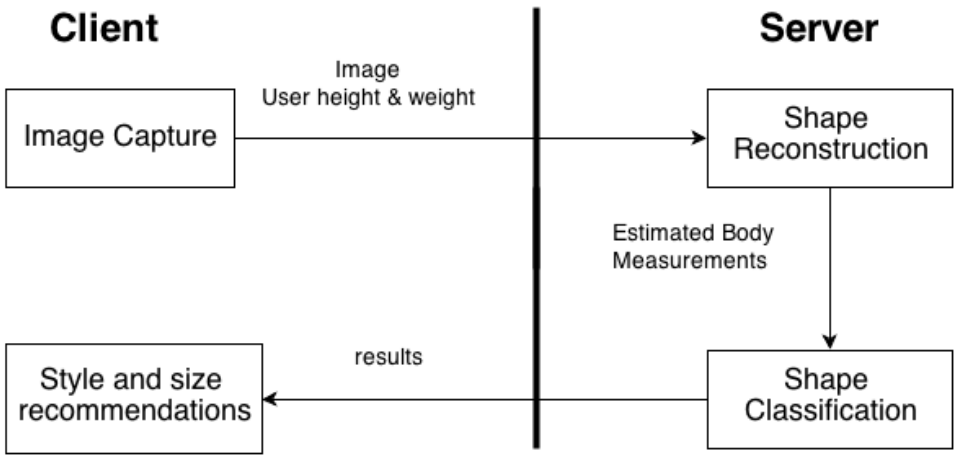

Fig. 2. Basic work-flow diagram.

4. Human Shape Model

In this work, a shape model is learnt from a set of registered 3D scans of 110 male and female subjects in a standard pose [5]. We learn a lower dimensional subspace of human shape variation by performing principal component analysis (PCA) on the vertex coordinates of example scans.

A PCA model allows us to generate arbitrary human models by moving in different directions along the principal variation axes. These axes successfully capture variations due to height, weight, gender etc. as shown in figure 3.

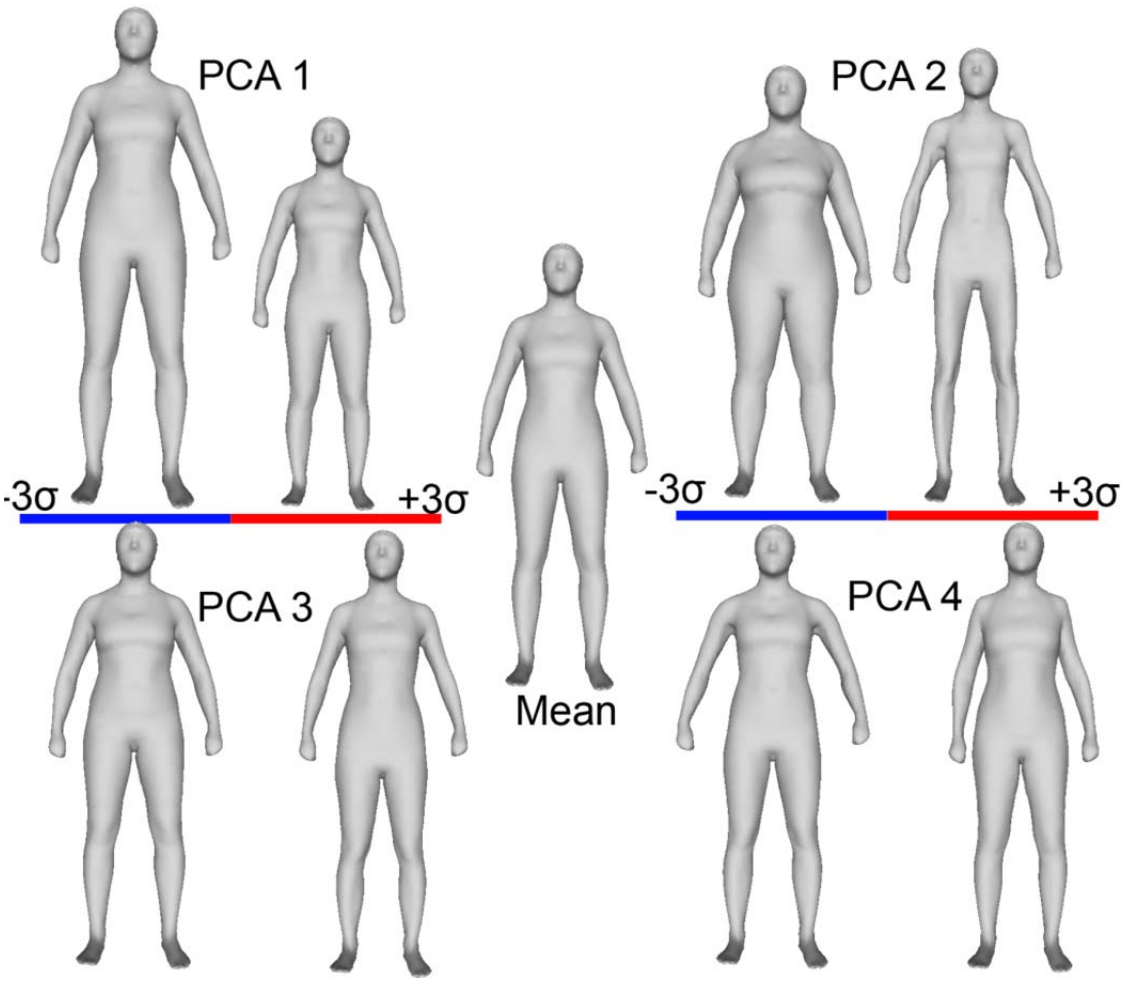

Fig. 3. Shape variation modes.

5. Estimating Shape from a single image

Given a photo $\boldsymbol{I}$, our task is to estimate the PCA parameters and camera positions by minimizing an energy function. For a given set of shape parameter $\boldsymbol{\Theta_S}$, the corresponding mesh is projected onto the input image $\boldsymbol{I}$ with camera translation $\boldsymbol{\Theta_c} = [\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}]$.

The energy function $\mathbf{E}(\boldsymbol{\Theta_S}, \boldsymbol{\Theta_C}, \boldsymbol{I})$, to be minimized, measures how well the projected image fits the human figure of the input image. Our energy function includes four components:

$$\mathbf{E} = \mathbf{E}_{\boldsymbol{colour}} + \boldsymbol{E}_{\boldsymbol{edge}} + \boldsymbol{E}_{\boldsymbol{joint}} + \boldsymbol{E}_{\boldsymbol{semantic}} \qquad (1)$$

Color distribution term

Good shape estimation should correspond to a good image segmentation defined by the projected silhouette. Therefore, the quality of the segmentation can be evaluated by comparing the color distribution of the regions inside and outside the silhouette using the Bhattacharyya distance [6].

Edge matching term

In addition to the color distribution error, the quality of the segmentation can be evaluated by comparing the boundary of the projected silhouette and the observed body boundary. To evaluate this matching we use a smoothness term as introduced in [8].

Joint position term

We use a joint position term in the energy function to encourage the hypothesized joints, face and hands, to fall close to corresponding joints inferred from the input image. Our experiments show that this term is very important to avoid premature termination of our optimization process. While the face is automatically detected, our system requires the user to manually select the rough position of the hands and feet in the input image.

Semantic term

To eliminate single view ambiguities we have added a semantic term which constrains the hypothesized height and weight to be close to the given height and weight. Additional measurements can be added for increased accuracy such as waist girth, bust girth, etc.

Optimization

The energy function (1) to be minimized is non-linear and complex. As a compromise between speed, robustness and accuracy we have chosen the Nelder-Mead [7] method which is a commonly used non-linear optimization technique.

6. Shape Classification

Shape classification framework is based on the technique presented in [10]. Given the ratios of selected body measurements (waist, bust, hips, high hips, abdomen and stomach), a classifier outputs a body type from a set of nine. The nine different body types are: Rounded, Pear, Triangle, Inverted Triangle, Rectangle, Diamond, Hourglass, Bottom Hourglass and Top Hourglass (see figure 4).

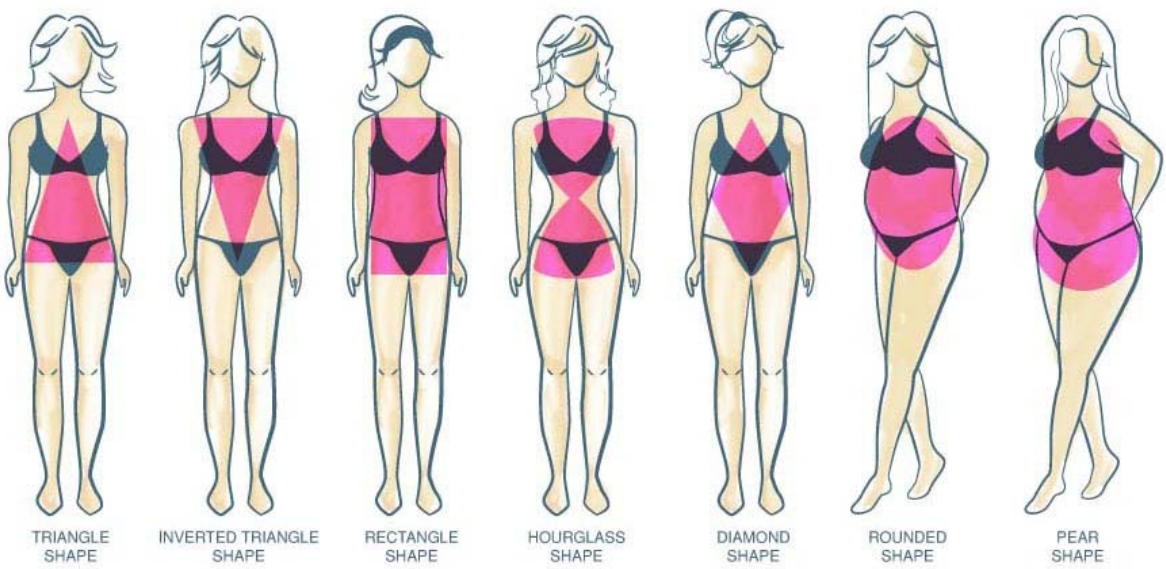

Fig. 4. Female body types. Image courtesy freelancer.com.

7. Experimental Results

In order to evaluate the accuracy of our system, we compared selected body measurements estimated with our approach against ground truth measurements obtained from a 3D scanner. The images contain a variety of backgrounds and subjects are wearing normal clothing. Our results show that selected body dimensions can be estimated and classified within an average accuracy of <3cm even with subjects wearing tight dresses. Table 1 shows the average and maximum errors obtained when comparing our approach against ground truth measurements.

The waist and abdomen measurements appear to have the highest average error. This occurs due to the significant depth ambiguity in the 2D image around the subjects' upper torso. More specifically, this area exhibits significant shape variation amongst individuals and due to the depth ambiguity, a variety of shapes can satisfy the silhouette overlap constraint. A practical solution to this problem would be to ask users to specify their waist measurement and add a waist error term in the optimization process.

The shape classification heavily depends on accurate shape estimation. In our tests, 8 out of 10 subjects were classified successfully. In the two cases that the classification failed, a high waist error measurement was observed (Figure 5 row 3).

As illustrated in figure 5, the joint segmentation and reconstruction framework allows for shape estimation and classification in cluttered scenes. This is achieved from images captured on mobile phone device with unknown camera parameters and unknown distance to the subject.).

Table 1. Estimated measurements vs. ground truth.

Measurement	Average error (cm)	Max Error (cm)
Hips	1.49	2.88
High Hips	1.42	4.30
Abdomen	1.88	5.52
Waist	2.16	4.96
Stomach	2.97	6.11
Bust	2.88	6.907

8. Conclusion and Future Work

In this paper, we have presented a practical system for estimating 3D human shape from a single image captured from a hand held device.

Our system has various limitations. Firstly, it is important to note that our system was trained only on a small dataset thereby limiting its capabilities. Secondly, the subjects are required to be in a fairly standard pose and wearing relatively tight clothing. In addition, since only one image is being used the shape estimation may not perform well for specific measurements. This can be compensated by requiring the user to input more information on their body dimensions for increased accuracy. Our method performs best when the color of the subjects' clothing is easily distinguishable from the background as shown in figure 5.

Future work will be aimed at further increasing the accuracy of the system and improving user interaction and practicality.

Acknowledgements

We wish to thank the Fashion Digital Studio and the London College of Fashion, University of the Arts London for providing the dataset used for the evaluation of the system and Hasler *et al.* [5] for kindly providing the dataset of registered scans for training the shape model. Current research project was supported by an EPSRC grant entitled: Body Shape Recognition for Online Fashion (Project reference: EP/1032169/1).

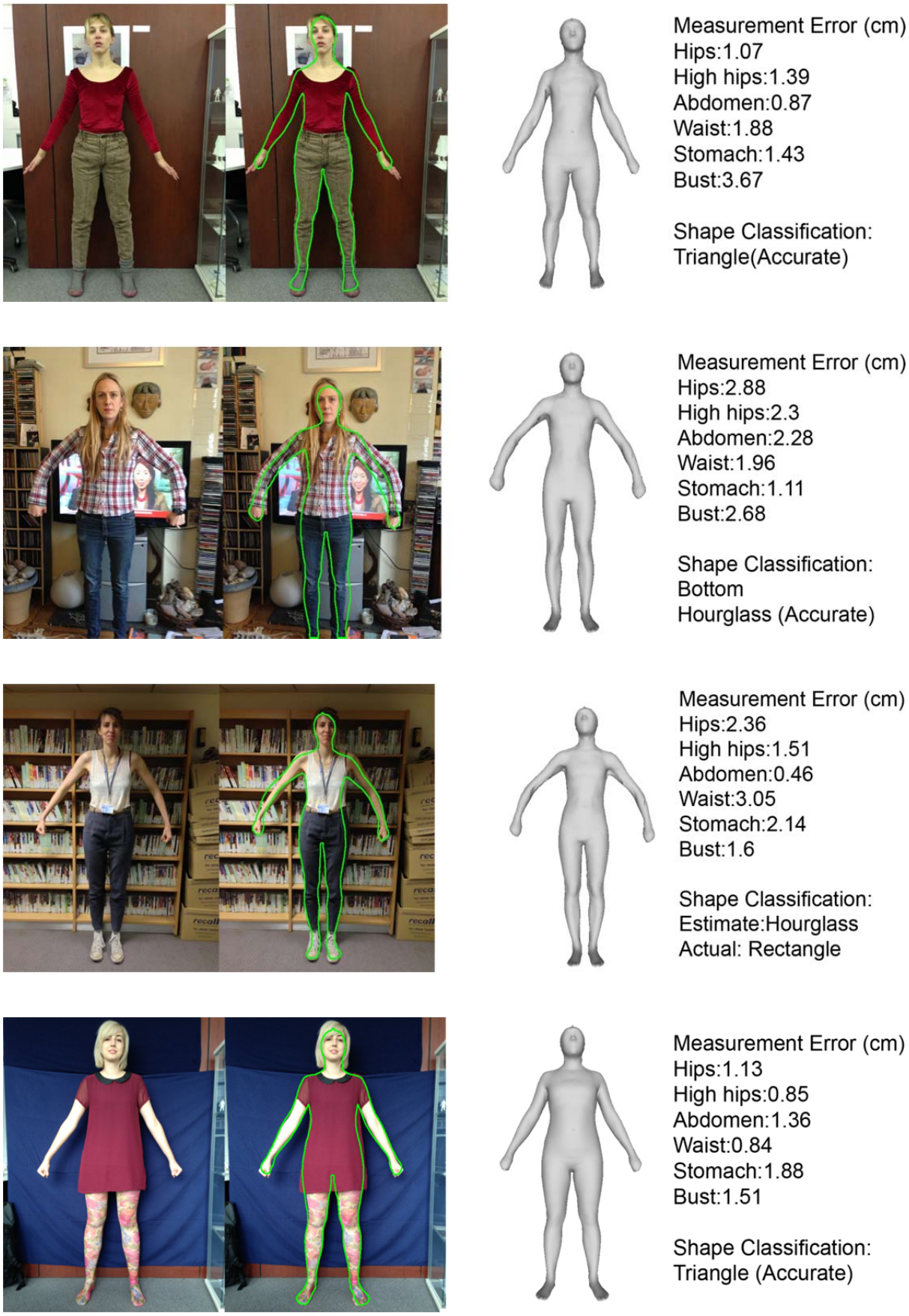

Fig. 5. Body shape estimation in a mixture of indoor backgrounds.

References

1. BALAN, A. O., BLACK, M. J., HAUSSECKER, H., AND SIGAL, L., 2007. Shining a Light on Human Pose: On Shadows, Shading and the Estimation of Pose and Shape.
2. BALAN, A. O., SIGAL, L., BLACK, M. J., DAVIS, J. E., ANDHAUSSECKER, H. W. 2007. Detailed Human Shape and Pose from Images. 2007 IEEE Conference on Computer Vision and Pattern Recognition, July, 1–8.
3. CHEN, Y., KIM, T.-K., AND CIPOLLA, R. 2010. Inferring 3D Shapes and Deformations from Single Views. ECCV 6313, 300–313.
4. GUAN, P., WEISS, A., AND BLACK, M. J. 2009. Estimating Human Shape and Pose from a Single Image. Work 2, ICCV, 1381–1388.
5. HASLER, N., STOLL, C., SUNKEL, M., ROSENHAHN, B., AND SEIDEL, H. P. 2009. A Statistical Model of Human Pose and Body Shape. Computer Graphics Forum 28, 2, 337–346.
6. LEE, M. W. L. M. W., AND COHEN, I., 2006. A model-based approach for estimating human 3D poses in static images.
7. POWELL, M. J. D. 1973. On search directions for minimization algorithms. Mathematical Programming 4, 1, 193–201.
8. ROTHER, C., KOLMOGOROV, V., AND BLAKE, A. 2004. ”GrabCut”: interactive foreground extraction using iterated graph cuts. Computer 23, 3, 309–314.
9. SIGAL, L., BALAN, A., AND BLACK, M. 2007. Combined discriminative and generative articulated pose and non-rigid shape estimation. Advances in Neural Information Processing Systems 20 20, 1337–1344.
10. SIMMONS, K., ISTOOK, C. L., AND DEVARAJAN, P. 2004. Female Figure Identification Technique for apparel. Part I: Describing Female Shapes. Journal Of Textile And Apparel Technology And Management 4, 1.