Capture and Stylization of Human Models

Edilson de Aguiar, Leandro Lesqueves Costalonga, Luís Otávio Rigo Júnior and Rodolfo da Silva Villaça
DCEL / CEUNES / UFES
São Mateus, ES, Brazil
Email: edilson.de.aguiar@gmail.com

(a) (b) (c)

Fig. 1. From (a) recorded subjects, we create (b) personalized models that can be (c) rendered in a game or virtual world.

***Abstract*—Capturing and creating models of human subjects is an ongoing effort in computer graphics. In this paper, we present a framework to capture, model, simplify, and stylize an human subject. We use a consumer 3D depth camera that captures the appearance, shape, and pose of the recorded person. We then fit this data with a parametric human model that is based on 3D scans available from the CAESAR database. Once we have a personalized 3D human model, we can apply stylization techniques to create a variety of outputs. We present several results obtained with our framework and verify that these models remain recognizable as their human sources.**

***Keywords*-morphable model, customized models, stylization, animation**

I. Introduction

Capturing human models has been a major effort in graphics and modeling for more than a decade. Devices, such as single and multiple-view cameras, camcorders, laser scanners, and infrared depth cameras, have been employed to acquire both photometric and geometric properties of human subjects. As these devices become cheaper and more abundant (e.g., with the advent of 3D cameras like Microsoft's Kinect) it seems as though anyone can capture a model of herself at home. However, most of these devices only provide low level readings of a scene in the form of pixels or point-samples, and there is still a gap between such outputs and an usable 3D human model that can support advanced applications.

In this paper, we present a framework to capture, model, simplify and stylize a personalized model of a subject (see Figure 1). Such models can be rendered and used as avatars in simple virtual worlds and games.

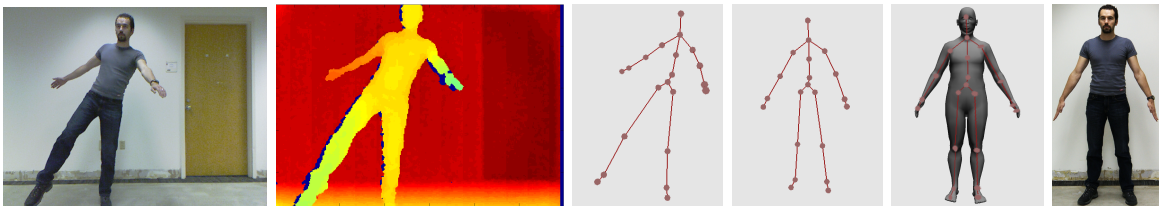
Our system starts by capturing the shape, appearance and pose of a person positioned in front of a 3D camera. Next, a parametric body shape model that is based on the CAESAR database [1] is fitted to the recorded data, to separate the pose and shape parameters such as height, hip circumference, and face length. Once these parameters have been recovered, we use them to reconstruct a simplified 3D human model that closely matches the recorded subject. We can then apply various stylization techniques to the 3D models. For example, we can interpolate between the vertex positions of the subject's model and an artistic style to achieve a stylized version of the person.

II. Related Work

Our framework builds on advances in a number of research areas, including parametric shape models, human pose estimation, human shape estimation, and stylization.

Parametric Shape Models Our technique makes use of a parametric model of human shape and pose similar to previous work [2], [3], [4], [5], [6]. Our model is learned from a large collection of 3D scans [1] and enables us to describe shape variations linearly. We follow the original work of Allen and colleagues and construct semantic parameters for body and head shapes [2].

Human Pose Estimation Pose estimation in 3D has been an important topic of research in both computer graphics and computer vision (see [7] for a review). Recovering the pose, particularly with a single camera, is a challenging problem [8]. For this reason, many methods require some form of user assistance to boost performance [9]. An alternative approach is



(a) C_C (b) D_C (c) θ_C (d) θ_R (e) Template (f) I_1

Fig. 2. **Input data to our approach** For each subject, the Kinect provides (a) an RGB image, (b) a depth map image and (c) a kinematic skeleton. The recorded person is also captured in (d) a reference pose that is used for calibration. The shape parametric models are created with the help of (e) a template model that was created by an animator. We also capture one additional high-resolution photograph of the recorded subject in the reference pose (f).

to use the additional information about depth that is available from 3D cameras to improve pose estimation [10].

Human Shape Estimation Modeling the shape of a person from images is also difficult. Recent work in this area has employed parametric models to estimate human shape and pose from single images [11] or a set of images [12] and in multiple camera setups [13]. Our framework bears some similarity to recent data-driven approaches used to capture and modify human shape and appearance in images [14] and video [13]. However, instead of working in 2D, we capture, personalize, and stylize a human model in 3D.

Stylization A number of approaches have been proposed to simplify images and video on the basis of perceptual considerations [15], [16], [17], [18]. For example, Gooch and colleagues presented a method to create black-and-white illustrations and caricatures of human faces from photographs [19]. Most of these techniques have focused on recreating the look and feel of certain media, rather than targeting specific object types. A related method [20] to abstract 3D models considers man-made shapes, such as buildings, but not human models.

III. General Methods

We will now describe the basic components that our system builds on and the general procedures followed in our approach.

Capturing Device Our system uses the Kinect to capture the image, depth and pose information of individuals. This 3D camera is a hybrid capturing device that contains an RGB camera of 640×480 pixel resolution and a depth range sensor of 320×240 pixel resolution. In addition, it also recovers an initial pose estimate for the recorded subject using a specific kinematic skeleton (Figs. 2(a)-(c)).

Capture Session Our capture session starts by capturing a subject in our reference pose (θ_R), an inverted V pose as shown in Fig. 2(d), for calibration. After this step, the person moves to a desired pose of their selection (θ_C), a candid pose as shown in Fig. 2(c), and a set of color and depth images is obtained (C_C, D_C; Figs. 2(a) & (b)). In addition, we record one photograph during the capture session with a calibrated high-resolution Sony HD camera: a frontal view of the subject in the reference pose (I_1) (Fig. 2(f)).

CAESAR Database We construct separate parametric models for the human head and the human body shape for each gender using the 3D scans, 72 landmarks for each scan, and semantic information available from the CAESAR database [1]. Before we can learn the parametric models, the scans have to be brought into semantic correspondence with each other. This correspondence is achieved by aligning and registering each scan to a template model (Fig. 2(e)) by using a set of landmarks and the method of Tena and colleagues [21]. Similar to the approach taken by Allen and colleagues [2], we create the parametric models from CAESAR scans that are all in the same standing pose, and standard vertex skinning is used to connect the models to the Kinect skeleton.

IV. Human Morphable Models

Our framework simplifies human shape modeling by using a head and a body shape parametric model created using the CAESAR database. After bringing the scans into semantic correspondence, we calculate the average model (m) and the main shape principal components (pc) using PCA. Similar to Allen and colleagues [2], we use linear regression to transform pc to meaningful components (mpc) using the semantic information associated with each scan in the database.

As a result, our parametric models are able to reconstruct mesh models (mm) for the head and body shapes by combining m and a number, N_{mpc}, of mpc as follows:

$$mm = m + \sum_{i=0}^{N_{mpc}} \alpha_i \cdot mpc_i, \tag{1}$$

where α_i is a vector with weights for each mpc_i.

A. Human Body Shape Model

Morphable models were created for the male body shape with 1234 scans and for the female body shape with 1081 scans. We used the original landmarks to align and register the scans and the semantic information associated with each scan to transform the principal components into meaningful principal components. In the CAESAR database, there are 79 semantic parameters (30 demographic parameters and 49 physical measurements) for each scanned individual.

We reduced the number of semantic parameters by excluding similar and duplicate parameters, parameters that are not visually salient and parameters that are difficult to estimate from our input data. Our final body shape mesh model, mm_{body}, is reconstructed using Eq. 1 and five semantic parameters that span the most visible shape variation: chest circumference, hip circumference, stature, thigh circumference and waist circumference. We created two gender-specific shape

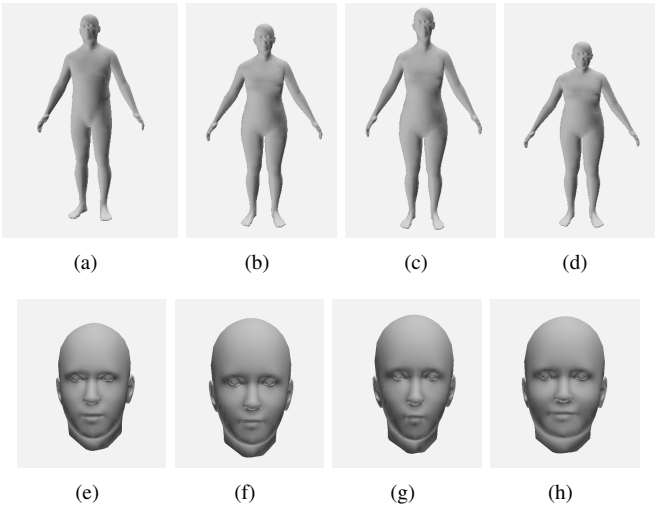


Fig. 3. (a)-(d) Average male and female body models and Stature variations for the female morphable model. (e)-(h) Average male and female head models and variations

morphable models (Figs. 3(a)-(d)) and the model is connected to the Kinect skeleton using standard vertex skinning. This way, new poses θ for mm_{body} can be generated using motion captured data.

B. Human Head Model

Given the poor quality of the head scans in the CAESAR database, we created a morphable model for the head shape from a reduced set of scans (84 male scans and 110 female scans). We ensured that the demographic distribution of scans in this subset was similar to that for the entire database. The lack of reliable landmarks in the face and head area in the original scans meant that we had to manually mark 18 landmarks around the head. These landmarks were used to align and register the scans. Only the head portion of these scans was used for learning our models.

We created a set of measurements from the 3D markings to serve as semantic information (e.g., distance between the eyes, mouth size) and to transform the head principal components into meaningful principal components in a similar process as for the body shape parameters. This semantic information is related to the facial features that we capture from the input image data using [22].

Our final head shape mesh model, mm_{head}, is then reconstructed using Eq. 1 and two parameters that span the most head shape variation based on the 18 landmarks: jawbone distance and bridge of nose to chin distance. We created two gender-specific face morphable models as show in Figs. 3(e)-(h).

V. Capturing Pose, Shape, and Appearance

After recording the person, our framework uses θ_R to adjust the pose of our generic body shape morphable model, making it more accurately fit the subject's reference pose. Next, appearance details such as clothing style and colors, hair

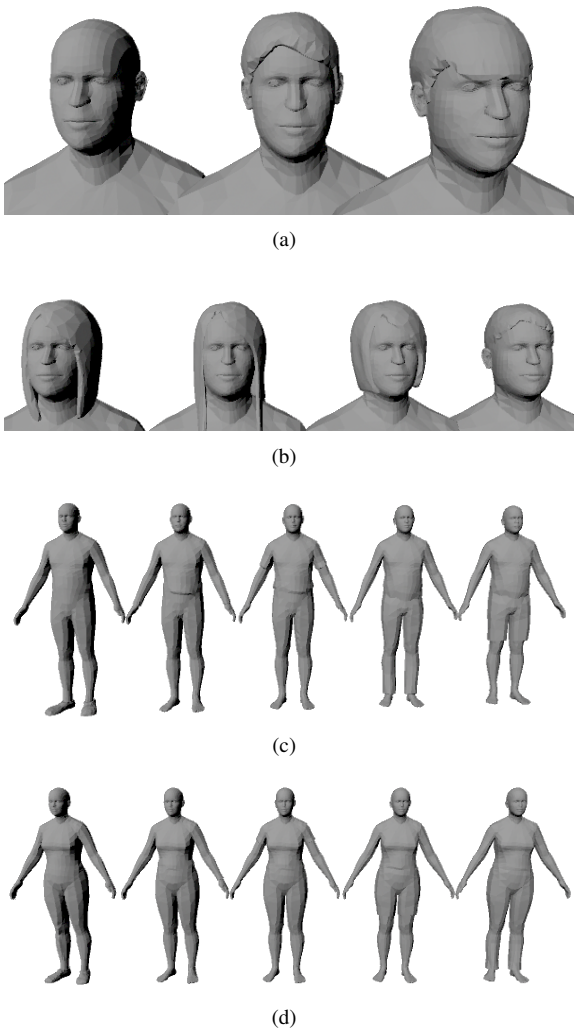


Fig. 4. Example of 3D models in our database for (a)-(b) hair and (c)-(d) clothing styles for male and female subjects that are registered to our parametric shape model.

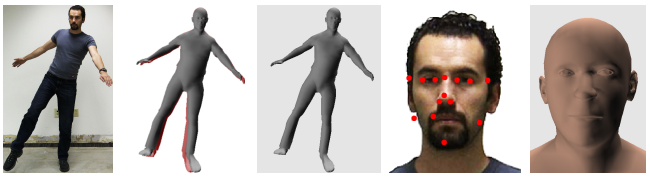
style and color are automatically estimated from I_1, and their closest matches are selected from a database of 3D hair and 3D clothing models using a simple voting-based algorithm. After this step, the input candid pose, θ_C, is refined using our underlying human shape representation. Semantic parameters, mpc_{body}, are also optimized making the shape of the 3D model closer to the recorded subject's body shape. Facial features are extracted from I_1 and are used to optimize the semantic parameters, mpc_{head}, yielding an improved head shape model that better matches the recorded person.

A. Capturing Appearance Details

Before we can capture appearance details, we conduct two pre-processing steps: i) our template model is segmented into five distinct regions (hair, upper body, lower body, skin, and shoes) and ii) all the 3D models for hair and clothing (four types of garments and seven hair styles) are registered to our parametric shape model (Figs. 4(a)-(d)). These pre-processing

(a) (b) (c) (d)

Fig. 5. **Capturing appearance and pose** An (a) improved average body shape model, $m_{bodycloth}$, is obtained after matching to a subject's clothing style. The improved model is used by an optimization cost that considers the silhouette as specified by (b) a Chamfer distance map and the (c)-(d) 3D point cloud generated from the depth information to refine pose estimates.



(a) (b) (c) (d) (e)

Fig. 6. **Intermediate results of our system** The (a) recorded subject and the (b) optimized pose reconstructed by our approach is overlaid on the initial pose (in red). The (c) optimized body shape for the recorded person in (a). (d) Facial features (marked with red dots) are extracted and used for improving our (e) estimated head model for the individual in (d).

steps only have to be carried out once.

We start by recovering the skin color, hair color, and garment colors from I_1. We use I_1 rather than C_C for this step due to its higher resolution. The vertices of our parametric shape model are then projected onto I_1 in segments that correspond to the upper and lower body garments. Specific garment models for each subject (e.g, shorts vs. pants for lower body, shirt vs. t-shirt for upper body) are selected automatically based on the degree of overlap between the projected 3D garment and the detected garments in the image I_1. Once 3D garment models have been selected, they are assigned the same colors as the detected garments in I_1. The 3D models for hair are selected in the same way and are then connected to m_{body} via mean value coordinates [23]. The selected clothing models are used to deform the vertices of m_{body}, yielding $m_{bodycloth}$, Fig. 5(a). Because we explicitly model clothing layers and hair, we are able to match the recorded person more closely using $m_{bodycloth}$ than simply using m_{body} in Eq. 1.

B. Improving the Candid Pose

We use our improved body shape morphable model, $mm_{bodycloth}$, that includes clothing to refine the initial pose, θ_C, in a hierarchical pose estimation framework, Fig. 5(a). The shape model is connected via skinning to the Kinect skeleton, which comprises 24 joints. For each recorded person, we process the input image data by first segmenting C_C using D_C in order to create a silhouette image. Then, a Chamfer distance map which is zero inside the silhouette is generated, Fig. 5(b). We also process the depth information D_C to create a 3D point cloud, Fig. 5(c)-(d). Our hybrid analysis-by-synthesis hierarchical scheme divides the kinematic structure into 5 regions: torso, left arm, right arm, left leg, and right leg. We optimize the pose parameters of each sub-kinematic chain separately starting at the torso, and continuing down the kinematic hierarchy. The energy function to be optimized contains two terms: a silhouette-based term (V_{Sil}) and a 3D point cloud-based term (V_{3D}). The first term, V_{Sil}, tries to minimize the distance between each projected vertex from $mm_{bodycloth}(\theta)$ and the input silhouette boundary using the Chamfer distance map. The second term, V_{3D}, tries to minimize the distance between each 3D point and the closest vertex in the model $mm_{bodycloth}(\theta)$. Our final energy function, V_{TP}, is a weighted sum of V_{3D} and V_{Sil}, with a higher weight associated to V_{3D}. We use a Quasi-Newton LBFGS-B method [24] in order to optimize V_{TP} to obtain a refined pose estimate ($\theta_{C'}$) that more accurately reproduces the subject's candid pose, Fig. 6(c).

C. Capturing Body Shape

Next, we improve the overall body shape of $mm_{bodycloth}(\theta_{C'})$ by optimizing the semantic parameters mpc_{body}. Our shape optimization is based on the silhouette, V_{Sil}, and on the 3D points, V_{3D}. Our final energy function V_{TS} is a weighted sum of V_{3D} and V_{Sil}. We found out that enhanced results are achieved when V_{3D} has a higher weight. The optimal vector with weights α is found with a Quasi-Newton LBFGS-B method [24] and the refined shape model is reconstructed with Eq. 1.

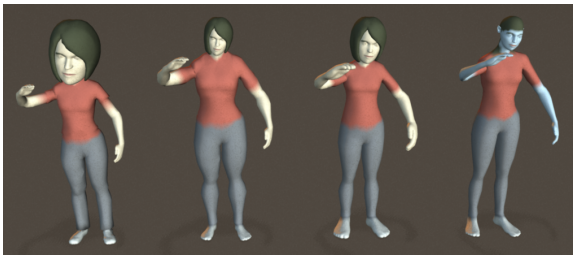
D. Capturing Head Shape

In order to increase the realism and accuracy of our final 3D model, we use the technique described in [22] to extract facial features from I_1, like eye and mouth position (Fig. 6(d)). We only optimize the two parameters for the parametric head model. The optimal values for α_1 and α_2 are found by minimizing the distances between the projected 3D markers from mm_{head} and the corresponding detected image features. We optimize α_1 and α_2 using the Quasi-Newton LBFGS-B method [25] and reconstruct a more accurate head model using Eq. 1, as shown in Fig. 6(e). We connect the refined head model mm_{head} to the optimized body shape mesh model, $mm_{bodycloth}$, using a set of correspondences calculated automatically in a pre-processing step. We use some additional face features extracted from I_1 to automatically determine the colors of the eyes, eyebrows, and lips for each recorded subject. Further personalization is achieved for the male head models by automatically detecting facial hair using a voting scheme similar to the one presented in Section V-A.

VI. Stylizing the Captured Model

We have used three different schemes to stylize our reconstructed 3D model ($hmodel$): morphing, caricatures, and props. Morphing styles are created by combining $hmodel$ with one of four stylized models ($mstyle_i$) specified by an animator. These gender-specific stylized models are derived from the average models (m_{body}) in the reference pose (θ_R). Therefore, the semantic correspondence between the average

(a)



(b)

(c) (d)

Fig. 7. **Stylization results** We can stylize our 3D models in many different ways. For example, (a)-(b) by varying the type or the influence of a given style, (b) by creating caricatures that emphasize the most salient aspect of the person, or (c) by adding props.

and stylized models is preserved. Stylized models can also be created using any modeling tool and semantic correspondence to m_{body} can be achieved with a few manually selected correspondences by using existing methods [21]. A stylized model is generated by adding the difference between the vertex positions of $mstyle_i$ and m_{body} to the final reconstructed model $hmodel$ through a blending factor β as follows:

$$hmodel_{style} = hmodel + \beta \cdot (mstyle_i - m_{body}). \quad (2)$$

Fig. 7(a) shows the morphing stylization results as a function of the blending factor β. As can be seen, the body shape of the recorded subject is preserved in this stylization transformation. Fig. 7(b) shows examples of our four morphing styles: Bobblehead, Strong, Cartoon and Avatar. We can also caricaturize models by changing the optimized weighting vector α in Eq. 1 for reconstructing $mm_{bodycloth}$. One common technique for creating caricatures is to emphasize a single meaningful component mpc_{body} and increase or decrease its effect. Fig. 7(c) shows the results of multiplying the semantic component with the largest variation from the average model for this subject, waist circumference, by a factor of 0.5 and 2. Another possible type of stylization is the addition of props or external components to the models. To exemplify this type of style, we created some props and added them to the models depicted in Fig. 7(d).

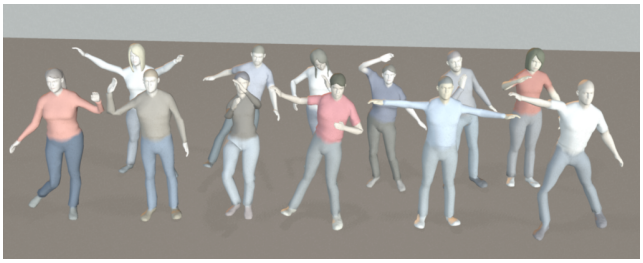


Fig. 8. Example of 3D models created by our approach for 12 different subjects.

VII. Experiments and Results

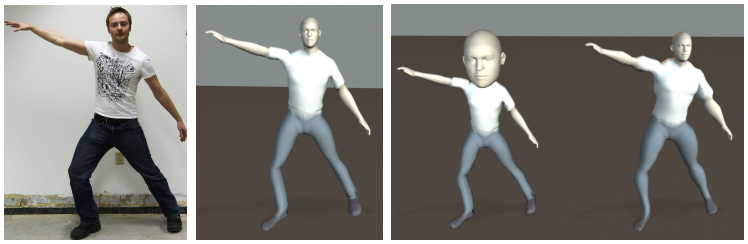
The main goal of this paper has been to create recognizable models using the coarse pose estimates provided in the input data by the Kinect. By customizing the body shapes, face shapes and hair, our models become more recognizable as the person they were based on. Figs. 8 and 9 show the 3D models created for 12 different subjects wearing everyday attire in a variety of candid poses.

Our system is able to preserve individual characteristics of recorded subjects. In terms of speed, our system is reasonably fast: capturing the appearance takes 1s, refining the input candid pose takes $\approx$ 20s, optimizing the body shape parameters takes $\approx$ 10s and capturing head shape and facial features takes $\approx$ 2s. These timings were obtained with unoptimized single-threaded code running on a laptop (Intel Core2 Duo, 2.4 GHz). Future work will include improving the time performance using GPUs and parallelization and will therefore allow interactive applications of our system.

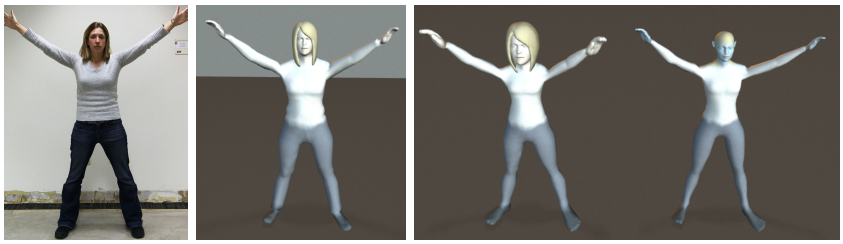
Although our framework has focused on rendering of 3D models, we are able to recover the life-sized dimensions of our recorded subjects. For example, we compared the physical measurements from one male and one female subject to the estimates of the five body shape parameters that are used by our system, and found differences on the order of a few centimeters. This accuracy is similar to that reported by [11] and demonstrates that our framework can be used for many other applications as well.

VIII. Conclusions

Our current framework delivers recognizable 3D models and their stylizations automatically. However, as with any other system, there are a few limitations to be considered. Although we rely on the Kinect for input data, as we have pointed out in Section III, this device provides low resolution RGB and depth images which makes an additional high-quality camera necessary. Noisy or missing data due to occlusions is also common with depth range equipments. Our 3D camera suffers from these issues and, as a result, some poses are not captured properly. As seen in Fig. 9, the correct orientation of hands, feet, and head cannot always be captured mostly due to the lack of accurate depth information for these regions (i.e. 3D points). It is for this reason that we perform our



(a)



(b)

Fig. 9. Results of our system in the form of renderings. From left to right, recorded person, reconstructed model and two different styles for (a) male and (b) female subjects.

appearance capture and head shape optimization in 2D using high-resolution images from the Sony HD camera. Although we are able to, in most cases, detect the correct clothing style of a subject sometimes manual input was necessary to correct the detected garments. This happened when garments were of a color similar to a subject's skin color or when the type of garment worn by the person was not represented in our 3D clothing dataset (e.g. a skirt or a 3/4-sleeve shirt).

As our method only captures the coarse shape of the recorded subject, details such as facial expressions, wrinkles in garments or skin, curly hair, glasses, or watches cannot be reproduced. We leave the improvement of our system to include these features as future work.

ACKNOWLEDGMENT

The authors would like to thank FAPES (Espirito Santo Research Foundation Agency) for the financial support of this research.

REFERENCES

[1] CAESAR, "Civilian american and european surface anthropometry resource project, http://store.sae.org/caesar/," 2002.

[2] B. Allen, B. Curless, and Z. Popović, "The space of human body shapes: reconstruction and parameterization from range scans," *ACM Trans. Graph.*, vol. 22, 2003.

[3] H. Seo and N. Magnenat-Thalmann, "An example-based approach to human body manipulation," *Graph. Models*, vol. 66, pp. 1–23, 2004.

[4] 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, 2005.

[5] B. Allen, B. Curless, Z. Popović, and A. Hertzmann, "Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis," in *Proceedings of SCA*, 2006, pp. 147–156.

[6] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel, "A statistical model of human pose and body shape," in *Computer Graphics Forum (Proc. Eurographics 2008)*, vol. 2, no. 28, 2009.

[7] R. Poppe, "Vision-based human motion analysis: An overview," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 4–18, 2007.

[8] A. Agarwal and B. Triggs, "Recovering 3d human pose from monocular images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, pp. 44–58, 2006.

[9] X. Wei and J. Chai, "Videomocap: modeling physically realistic human motion from monocular video sequences," *ACM Trans. Graph.*, vol. 29, 2010.

[10] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Proceedings of IEEE CVPR*, 2010.

[11] P. Guan, A. Weiss, A. O. B. lan, and M. J. Black, "Estimating human shape and pose from a single image," in *ICCV*, 2009.

[12] 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 *Proceedings of IEEE CVPR*, 2010, pp. 1823–1830.

[13] A. Jain, T. Thormählen, H.-P. Seidel, and C. Theobalt, "Moviereshape: tracking and reshaping of humans in videos," *ACM Trans. Graph.*, vol. 29, 2010.

[14] S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han, "Parametric reshaping of human bodies in images," *ACM Trans. Graph.*, vol. 29, 2010.

[15] D. DeCarlo and A. Santella, "Stylization and abstraction of photographs," in *ACM Trans. Graph.*, vol. 21, 2002, pp. 769–776.

[16] J. Wang, Y. Xu, H.-Y. Shum, and M. F. Cohen, "Video tooning," *ACM Trans. Graph.*, vol. 23, pp. 574–583, August 2004.

[17] H. Winnemöller, S. C. Olsen, and B. Gooch, "Real-time video abstraction," *ACM Trans. Graph.*, vol. 25, pp. 1221–1226, 2006.

[18] C. Hong, Z. Yang, J. Bu, Y. Liu, and C. Chen, "Cartoon-like stylization of video for real-time applications," in *Proceedings of the IEEE ICMExpo*, 2008, pp. 985 – 988.

[19] B. Gooch, E. Reinhard, and A. Gooch, "Human facial illustrations: Creation and psychophysical evaluation," *ACM Trans. Graph.*, vol. 23, pp. 27–44, 2004.

[20] R. Mehra, Q. Zhou, J. Long, A. Sheffer, A. Gooch, and N. J. Mitra, "Abstraction of man-made shapes," *ACM Trans. Graph.*, vol. 28, pp. 137:1–137:10, 2009.

[21] J. Tena, M. Hamouz, A. Hilton, and J. Illingworth, "A validated method for dense non-rigid 3d face registration," *Proceedings of IEEE AVSBS*, vol. 0, 2006.

[22] J. Saragih, S. Lucey, and J. Cohn, "Deformable model fitting by regularized landmark mean-shifts," *IJCV (in press)*, 2011.

[23] T. Ju, S. Schaefer, and J. Warren, "Mean value coordinates for closed triangular meshes," *ACM Trans. Graph.*, vol. 24, 2005.

[24] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM J. Sci. Comput.*, vol. 16, 1995.

[25] R. Fletcher, *Practical Methods of Optimization, Unconstrained Optimization*. New York: Wiley, 1980.