15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Two-stage Model Fitting Approach for Human Body Shape Estimation from a Single Depth Image

Mei Oyama
Aoyama Gakuin University
Kanagawa, Japan
oyama.mei@vss.it.aoyama.ac.jp

Naoshi Kaneko
Aoyama Gakuin University
Kanagawa, Japan
kaneko.naoshi@vss.it.aoyama.ac.jp

Masaki Hayashi
Liquid, Inc.
Tokyo, Japan
m.hayashi@liquidinc.asia

Kazuhiko Sumi
Aoyama Gakuin University
Kanagawa, Japan
sumi@it.aoyama.ac.jp

Takeshi Yoshida
Aoyama Gakuin University
Kanagawa, Japan
yoshida@it.aoyama.ac.jp

Abstract

Recovering an accurate 3D human body shape from a single depth image is one of the challenging problems in computer vision due to sensor noises, complexity of human body shapes, and variation of individual body shapes. In this paper, we address the problem using a two-stage model fitting approach. At the first stage, a coarse template model is fitted to the human pose of the input depth image using skeleton deformation. Then the model is fitted to the human shape by Laplacian surface editing. This fitting may corrupt the human-like shape of the template model due to the incompleteness of depth information. Then in the second stage, body shape details are recovered by fitting of a fine model to the deformed template by Stitched Puppet model fitting. Several experiments demonstrate that our approach recovers the most likely body shape of the input and deals with over a variety of input body shapes.

1 Introduction

It is a big concern for most shoppers whether interested clothes will fit them in terms of style, color, and size. Most online shopping sites have recommendation systems for predicting the preference of users, such as color, brand, styles, etc. To improve recommendation performance, a search method for a large clothes database was proposed [1]. Although this type of recommendation methods often use clothing information, a shopper's body shape information is rarely used. Recommending the best fitting shape and silhouette to the shopper as well as showing how the clothes will fit them is more important than recommending color or style. To realize such a clothes shape fitting system, shopper's body measurements are required. In terms of complexity of equipment, scanning time, and the requirement for nakedness, current body scanning is not suitable for personal users. If the body shape can be estimated from a single depth image of a dressed human, clothes recommendation becomes easier and more effective. This will result in higher sales and better user satisfaction.

Human body shape estimation has been studied extensively in computer vision. One of the popular approaches is fitting template models to inputs [2–5]. The approach of Halser et al. [2] fits a statistical human body model to an input 3D body scan using the ICP algorithm and surface deformation. Zuffi and Black proposed Stitched Puppet model fitting, which is a realistic part-based 3D human body model fitting approach [3]. Although these methods achieve high estimation accuracy, they require a full body scan or a full mesh model as an input. These full scans are taken by a full-body capture systems. These systems are expensive and require dedicated capture space.

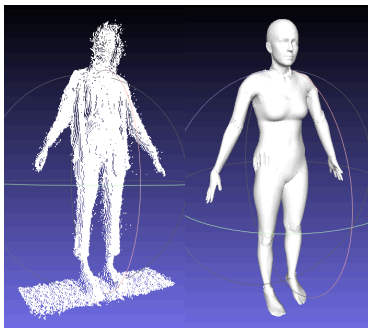

Figure 1: **Body shape estimation result.** We show the input depth (left) and the output body shape (right) of our method.

In contrast, methods using depth images as inputs have also been proposed [4, 5]. Chen et al. proposed tensor-based human body shape estimation [4]. In their method, a pseudo full scan taken from multiple depth cameras is used as an input. Perbet et al. proposed Multi-Resolution Manifold Forest [5], which contains vertical edges between tree nodes and horizontal edges between the nodes of different trees. Although depth images are comparatively easy to take with consumer depth cameras, the captured depth images are noisy and contain only one side of a human body surface. Therefore, the estimation error is large compared to the methods using a full scan.

In this paper, we propose a two-stage 3D human body shape estimation method from a single depth image. Our approach uses two types of model: coarse and fine. At the first stage, a coarse model is retrieved from our model database and is deformed according to the clothing subject's pose and body shape of the input. At the second stage, we fit a fine model to the deformed coarse model for expressing the detailed shape of the human body. Although we only use a single depth image as an input, the recovered human model is as precise as one obtained using full body scan (Figure 1).

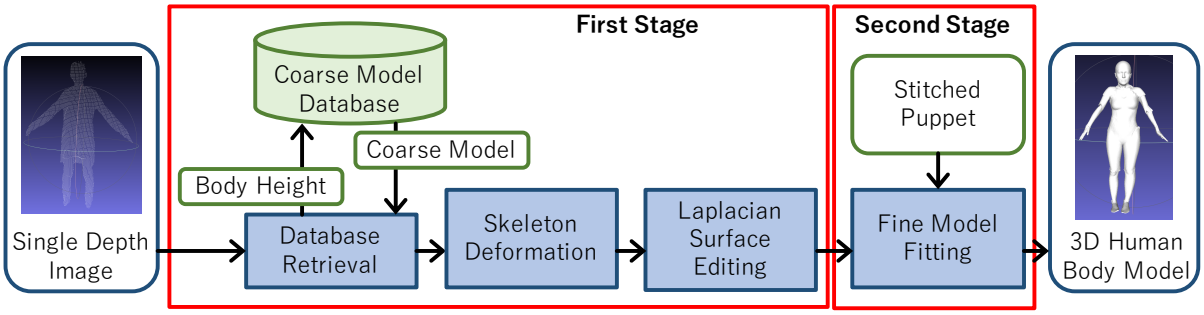

Figure 2: **Overview of our proposed method.**

2 Two-stage Model Fitting Approach

In this section, we describe our human body shape estimation method. Figure 2 shows the overview of our method. Our proposed method consists of two stages as follows:

(1) Retrieve a coarse model from the model database and deform the model according to the input subject's pose and shape.

(2) Fit a fine model to the deformed model for expressing the detailed shape of the human.

We describe the details of our method in the following subsections.

2.1 Coarse Model Database Retrieval

At the beginning of the first stage, a coarse model is retrieved from our model database.

In order to construct a coarse model database, we generate base models of males and females by MakeHuman [6]. We reduce the number of the model vertices in order to make the density of the vertices uniform. These base models have $N_M = 7,000$ vertices. Then, we change the body heights by 5 cm increments from 150 cm to 185 cm. As a result, we obtain a total of 16 coarse models for 8 body sizes and two genders. In these models, joints and skeletons are embedded.

In database retrieval, the coarse model with the height closest to the subject is retrieved. We calculate the approximate height of the subject by the distances between the body parts. We detect the body part positions from the input depth image using the method proposed by Shotton et al. [7].

2.2 Coarse Model Deformation

We deform the retrieved model pose and shape according to the human's pose and shape in the input.

First, the coarse model pose is deformed by the joint angles calculated by the detected body parts. As shown in Figure 3, left and right shoulder angles are calculated from shoulder and hand positions. Both hip angles are calculated from hip and foot positions. Then, the skeleton deformation is applied to the coarse model to align the model pose to the input. The deformation is automatically because joints and skeletons are embedded in the coarse model.

Second, the coarse model body shape is deformed. In order to align the model shape with the input, the input depth image is converted to a point cloud. The input contains the human body and the background. Since we need only the human body area, we apply depth based background subtraction to retrieve the point cloud containing only the human body.

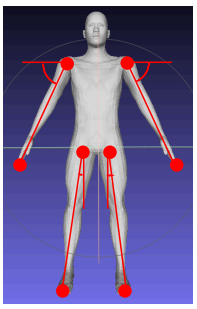

Figure 3: **Angle calculation using body parts.**

Since the point cloud only represents the human front surface, simple nearest neighbor matching results in a wrong deformation. For example, the back side of the model is collapsed with the front side. Furthermore, since the input human is wearing clothes, excessive fitting to clothes makes the coarse model loose human shape. Therefore, we introduce a constraint on the model deformation using vertex distances and normal vectors. Let $I = \{1, ..., N_M\}$ be a model vertex index set and $J = \{1, ..., N_C\}$ be an input point index set. The coarse model has vertices $M = \{m_i \,|\, i \in I\}$ and the point cloud has points $C = \{c_j \,|\, j \in J\}$. For each m_i, we choose a corresponding point p_i from C as follows:

$$p_i = \begin{cases} c_l & (d \leq d_2) \\ c_l & (d_2 < d \leq d_1 \wedge r \leq \theta) \\ null & (otherwise) \end{cases}, \tag{1}$$

where

$$l = \begin{cases} \arg\min_{k \in K}(\text{angle}(c_k, m_i)) & (d \leq d_2) \\ \arg\min_{j \in J}(\|c_j - m_i\|_2) & (d_2 < d \leq d_1 \wedge r \leq \theta), \end{cases} \tag{2}$$

$$d = \min_{j \in J}(\|c_j - m_i\|_2), \tag{3}$$

$$r = \min_{j \in J}(\text{angle}(c_j, m_i)) \tag{4}$$

The function $\text{angle}(c_j, m_i)$ is the angle between the normal vectors of c_j and m_i. Here $K = \{k \,|\, \|c_j -$

$m_i\|_2 \leq d_2\}$, and d_1, $d_2(< d_1)$ and θ are constant. Only vertices m_i that satisfy Eq. 1 are paired.

Then, we use Laplacian surface editing (LSE) [8], which is a mesh deformation method using Laplacian coordinates $\mathbf{w}_i$ as follows:

$$\mathbf{w}_i = \mathbf{v}_i - \overline{\mathbf{v}_i} \tag{5}$$

where, $\mathbf{v}_i$ is expressed in absolute coordinates, and $\overline{\mathbf{v}_i}$ is expressed in the coordinates of the gravity center of its neighbor points. LSE deforms only a portion of the model while maintaining the whole structure. We set handle vertices $H = \{m_i \,|\, i \in I \wedge p_i \neq null\}$ and their pairwise points D as destinations. The other vertices $O = M \setminus H$ are stationary. With this deformation, the coarse model is deformed according to the input body shape.

2.3 Fine Model Fitting

In the second stage, we fit a fine model to the deformed coarse model. We use the Stitched Puppet (SP) [3] as the fine model. SP is a realistic part-based 3D human body fitting model for 3D input data. SP minimizes the following energy function:

$$\sum_{i=0..15} \beta_i (D_i(\mathbf{x}_i, S) + R_{ij}(\mathbf{x}_i, S)), \tag{6}$$

where

$$D_i(\mathbf{x}_i, S) = \frac{1}{N_i} \sum_{k=1..N_i} (d_{i,k}(S)^2 + b)^\gamma, \tag{7}$$

and

$$d_{i,k}(S) = \min_{\mathbf{v}_s \in S} \|\tilde{\mathbf{p}}_{i,k}(\mathbf{x}_i) - \mathbf{m}_s\|_2 \tag{8}$$

is the distance from SP's point $\tilde{\mathbf{p}}_{i,k}(\mathbf{x}_i)$ to the coarse model $\mathbf{m}_s$, $\mathbf{x}$ are SP's variables, and N_i is the number of vertices of part i. $R_{ij}(\mathbf{x}_i, S)$ is the penalty term for normal vector miss-alignment. Refer to [3] for the complete objective function. Note that Eq. 6-8 show that SP optimization depends on the initial position of the SP model's parts. Although the original SP is randomly initialized with multiple poses, it may cause fitting failure due to wrong point correspondences. Thus, we improve the initial position using the pose obtained in the first stage. In particular, we initialize the direction of the torso to turn to the camera that captured the input depth image. This simple scheme improves estimation accuracy (See Section 3.2).

3 Experimental Results

In this section, we evaluate the performance of our approach using two different types of data:

- **Virtual data** was used for comparison with another method requiring a full mesh model as an input.
- **Noisy real data** was used to measure the estimation accuracy of our method for real data.

In the experiments, we used $d_1 = 10$ cm, $d_2 = 4$ cm and $\theta = 30$ degrees for all inputs. We measured height, arm lengths, leg lengths, bust, waist, and hip sizes. The ground truth of virtual data was measured using Maya [9], and real measurements were taken over clothes.

3.1 Virtual Data

In this experiment, we verify that our approach achieves accuracy comparable with the method requiring a full body mesh as an input. We created 5 male and 5 female models as test models using MakeHuman [6]. We randomly changed model parameters such as height, weight and amount of muscle. These test models were dressed in various clothes prepared with MakeHuman.

We compared our proposed method with SP fitting [3]. Because SP requires a full body mesh as an input, we used the test models themselves as the input. Therefore, we simulated a virtual time-of-flight camera to obtain depth images using BlenSor [10]. The depth image was acquired from only one viewpoint. The body joint positions of the depth image were manually given.

Figure 4 shows the estimation errors and Figure 5 shows examples of our results. The difference in estimation errors between our method and SP fitting is less than 1 cm for almost all items. Most of the precision of our method is the same as that of SP, except for the female hip. Compared to SP, our method is more practical because it does not require a full mesh model but only a single depth image as an input.

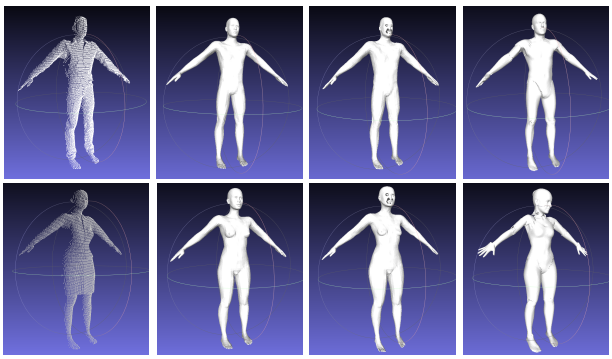

Figure 5: **Visual results for virtual data.** (a) Input point cloud. (b) Pose deformed coarse model. (c) Shape deformed coarse model. (d) Final output model.

3.2 Noisy Real Data

The noisy real test data were taken with a Microsoft Kinect v2 for 10 subjects, 5 males and 5 females, standing in A-pose. We compared the following three methods: our proposed two-stage method, our method without the coarse model deformation, and our method without initialization of the fine model.

Figure 6 shows the estimation errors and Figure 7 shows examples of our results. The result shows that body shape deformation improves estimation accuracy. No initialization of fine model causes large torso rotation in both genders. Our two-stage approach can prevent this problem by performing the first stage as initialization step for the second stage. Our method can estimate accurate body shape features from a single depth image. However, the estimated time per subject is about 11 minutes. Most of the time is LSE and SP fitting, and their speeding up is a future work.

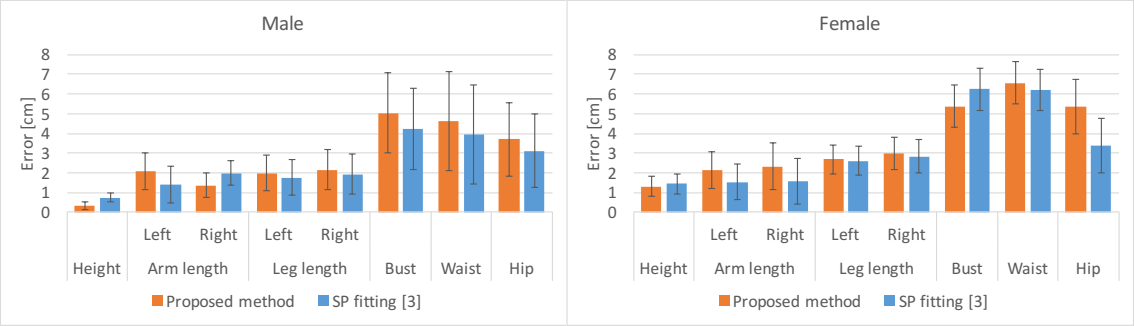

Figure 4: **Mean errors for virtual data.**

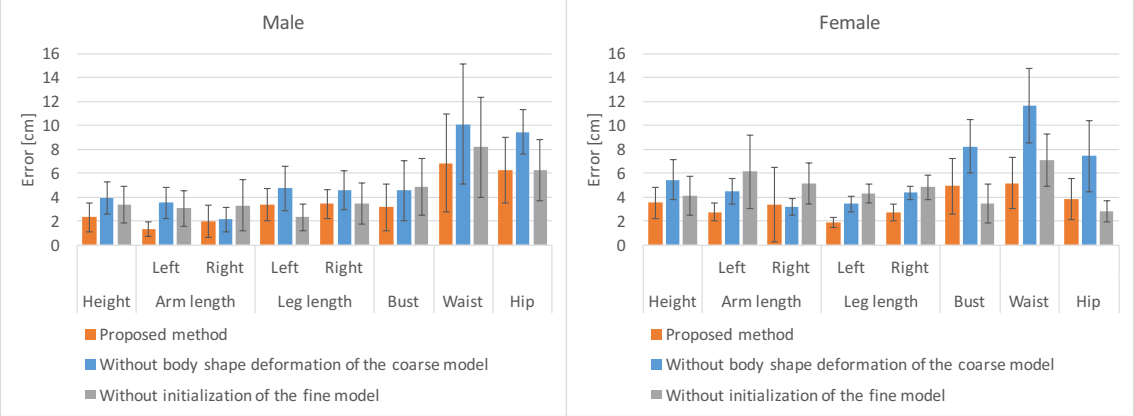

Figure 6: **Mean errors for noisy real data.**

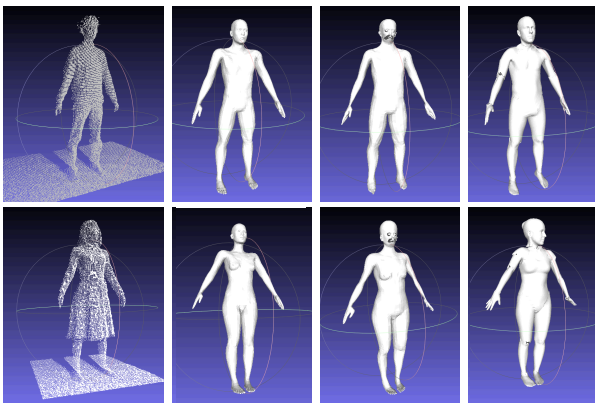

Figure 7: **Visual results for noisy real data.** (a) Input depth. (b) Pose deformed coarse model. (c) Shape deformed coarse model. (d) Final output model.

4 Conclusion

We proposed a two-stage model fitting approach for human body shape estimation. We constructed a coarse model database and performed coarse pose and shape fitting before fine model fitting. Experimental results have shown that our proposed method can deal with noisy depth images. In the future, we wish to improve the estimation time and accuracy by incorporating clothing segmentation and category classification.

References

[1] K. Lin, H.-F. Yang, K.-H. Liu, J.-H. Hsiao, and C.-S. Chen, "Rapid Clothing Retrieval via Deep Learning of Binary Codes and Hierarchical Search," in *ICMR*, pp. 499–502, ACM, 2015.

[2] N. Hasler, C. Stoll, B. Rosenhahn, T. Thormählen, and H.-P. Seidel, "Estimating Body Shape of Dressed Humans," *Computers and Graphics*, vol. 33, no. 3, pp. 211–216, 2009.

[3] S. Zuffi and M. J. Black, "The Stitched Puppet: A Graphical Model of 3D Human Shape and Pose," in *CVPR*, pp. 3537–3546, IEEE, 2015.

[4] Y. Chen, Z. Liu, and Z. Zhang, "Tensor-Based Human Body Modeling," in *CVPR*, pp. 105–112, IEEE, 2013.

[5] F. Perbet, S. Johnson, M.-T. Pham, and B. Stenger, "Human Body Shape Estimation Using a Multi-Resolution Manifold Forest," in *CVPR*, pp. 668–675, IEEE, 2014.

[6] MakeHuman. `http://www.makehuman.org`.

[7] J. Shotton, T. Sharp, A. Kipman, A. b. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. b. Moore, "Real-Time Human Pose Recognition in Parts from Single Depth Images," in *CVPR*, pp. 1297–1304, IEEE, 2011.

[8] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel, "Laplacian Surface Editing," in *EUROGRAPHICS*, pp. 175–184, ACM, 2004.

[9] Maya. `http://www.autodesk.co.jp/products/maya/`.

[10] M. Gschwandtner, R. Kwitt, A. Uhl, and W. Pree, "BlenSor: Blender Sensor Simulation Toolbox," in *ISVC*, pp. 199–208, Springer, 2011.