Human Body Shape Reconstruction from Binary Image Using Convolutional Neural Network

Jiayu Chen¹, Yueqi Zhong¹²⁺ and Zhicai Yu¹²
¹ College of Textiles, Donghua University, Shanghai 201620, China
² Key Laboratory of Textile Science & Technology of Ministry of Education, College of Textiles, Donghua University, Shanghai 201620, China

*Corresponding author email: zhyq@dhu.edu.cn

Abstract. Human body modeling is an important part of virtual try-on. In order to quickly reconstruct the three-dimensional(3D) human body based on the minimum input, we propose a new method for accurate reconstruction of the human body by inputting binary images from either single view or multiple views. We first encode the shape of the human body via Principal Component Analysis (PCA) to extract the low dimensional shape descriptor. Secondly, we design a novel Body Reconstruction Convolutional Neural Network (BRCNN) with two branches, which could capture deep correlated features from different views and merge them. Given the obtained statistical shape space of the human body, we jointly train the BRCNN to learn a global mapping from the input to the shape descriptor which can be then decoded to points cloud for the reconstruction of various body shapes under neutral poses. The experimental results show that compared with the existing human reconstruction technology, the accuracy has been improved by 1.07%, and the prediction results of the two views are better than those from the single view. Further investigation also reveals that the prediction results of the weight-sharing network are better than the network without weight-sharing.

Keywords: Human body reconstruction; Binary image; Principal component analysis; Shape descriptor; Convolutional neural network.

1. Introduction

Human body modeling has a pivotal role in textile and garment field with the development of augmented reality, virtual try-on and other technologies. Having a rapid and accurate method that reconstructs the human body shape can reduce the degree of difficulty of virtual try-on and improve the effects of it. Therefore, the key to human body modeling technology is to reconstruct three-dimensional body models of users with minimum input. At present, the method of three-dimensional (3D) human reconstruction can be divided into two main categories. One is to directly calculate the 3D point cloud coordinates of the human body surface through laser or stereo vision, and the other is to recover the 3D point cloud of human body by using single or multiple images to find the mapping relationship between the two-dimensional (2D) image and the 3D human body point cloud through either machine learning or data driven algorithms.

Early works in 3D human body reconstruction mainly achieved coarse approximations geometrically [1,2], but the emergence of statistical human shape models makes it possible to reconstruct the human body shape by inputting single or multiple images [3,4]. Recently, Indri [5] et al. found a global mapping from the images to the 3D human body parameters directly by training a convolutional neural network (CNN) and reconstructed the human body using human shape parameters. Endri et al. [6] proposed a new method, using multi-images input and HKS operator to train a CNN and reconstruct human body
more precisely. In another very recent work, Ji [7] designed a CNN with two branches, respectively training different views to jointly predict the shape descriptor and decoded it to points cloud for the reconstruction. This method has a higher precision in human body reconstruction.

To establish the mapping from images to human shape parameters, CNN is a very effective method [9]. Starting from 2012, CNN [8] has greatly promoted the rapid development of the computer vision. In the field of 3D modeling, one of the main methods of CNN to deal with classification, recognition and other related tasks is to extract low-dimensional shape descriptors, which have been widely studied for target matching and search [10].

A lot of different shape descriptors have been proposed and used. Based on HKS descriptors, Endri et al. [6] encoded the human statistical model into a new embedding space and learned an inverse mapping from the embedding spaces to the human shape space by CNN. Then, they decoded it for human reconstruction. Another commonly used descriptor is the PCA descriptor, and Endri [5] and Ji [7] encoded the human statistical model by PCA dimensionality reduction and obtained the PCA descriptors. Next, they captured the features of the image through CNN and made a regression prediction. Finally, the PCA descriptors generated by the prediction are inversely mapped to complete the reconstruction of the human body.

On the other hand, the statistical human shape model is a key factor of human body reconstruction. It represents variations in human shape and pose via low-dimensional parameter spaces, and is a useful tool to solve difficult problems, e.g. in the virtual try-on application. In 2015, Pishchulin[4] et al. built the MPII Human Shape Spaces, a statistical human shape models database, from CAESAR, the largest existing commercial scan database.

Illuminated by the method of Ji [7], the shape of the human body is encoded via PCA to extract the low dimensional shape descriptor with the MPII Human Shape Spaces in this paper. Then a two-branched Body Reconstruction Convolutional Neural Network (BRCNN) with the brainy images of human as input and the corresponding PCA descriptors as the output is built. In contrast to previous works, the mapping is learnt from input to output with two branches in our weight sharing network. This paper finally reconstructs the human body by decoding the descriptors predict by the BRCNN.

2. Methodology and Experimental Design

2.1. Dataset

The statistical human shape model (i.e., MPII Human Shape library) extracted from CAESAR [11], one of the largest commercial human body scanning databases, has been adopted in this paper as the benchmark human shape space. It contains 4300 male and female body models in standard pose (or so-called ‘A’ pose). Each body contains 6449 vertices, and all the bodies have been calibrated to eliminate bias caused by different posture. Figure 1 enumerated several randomly selected human models.

This paper intends to use PCA to represent each body model with the PCA descriptors:

\[ H_S = \lambda \varphi_S + \overline{H} \]  

(1)

where \( H_S \in \mathbb{R}^{3N} \) is a body with N vertices and the same topology, and \( \varphi_S \in \mathbb{R}^k \) is the PCA descriptor containing k principal features of the human body obtained from PCA. The projection matrix \( \lambda \in \mathbb{R}^{3N \times k} \) can project \( \varphi_S \) into a new space, and \( \overline{H} \) is the mean body shape of the training dataset. In order to minimize the differences in human reconstruction, we set \( k=50 \) in our implementation which can objectively characterize the human body and is computationally feasible.

2.2. Network Architecture

As illustrated in Figure 2 BRCNN is proposed for accurate reconstruction of the human body by inputting binary images from either single view or multiple views. The main structure of the network is based on the DenseNet, utilizing dense block and transition blocks to capture specific features of binary mask images of the human body.
DenseNet is a CNN proposed by Huang [12] et al. in 2017. It reuses features by connecting all the previous and subsequent layers in a feed-forward fashion. This architecture has achieved excellent performance in computer vision tasks, such as image recognition and object detection. Since our purpose is to extract the global information of human body through binary mask images, we can build a CNN with the same structure as DenseNet. The information collected by the previous layers can be fed back to the subsequent layers. For each layer in the dense block, the feature-maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into all subsequent layers. This kind of network structure alleviates the vanishing-gradient problem, strengthens feature propagation, encourages feature reuse, and substantially reduces the number of parameters. In our approach, the loss function of BRCNN is defined as follows:

\[
\mathcal{L}(\phi, \phi_s) = \frac{\|\lambda(\phi - \phi_s)\|^2}{N}
\]

where \(\lambda\) is the projection matrix, \(\phi\) is the prediction descriptor, \(\phi_s\) is the PCA descriptor and \(N\) is the total number of vertices.

The BRCNN with two branches decouples the features of two binary mask images from different views of human body and gets the global information by training a separate network for each mask respectively. In the area of 3D modelling, people usually exhibit their 3D models through three views, which indicate that each view of the 3D model contains local and global features of the model. As for human reconstruction, due to the self-occlusion of body, the top view conveys limited information. Besides, unprocessed body images from different views may rise privacy issues, hence the inputs used in this paper are the binary masks without identity information.

Further considering the correlation between the front view and the lateral view of the human body, it is possible to train two branches in BRCNN together, i.e., sharing weights or the same parameters during training. This makes the local descriptors learned by each branch no longer independent but closely related. In another words, this implies that two related local descriptors can be merged together to predict the global descriptors of the human body by joint training. In order to test the flexibility of the BRCNN, the network is allowed to reconstruct the human body through a single channel. That is, except for training the BRCNN with two binary mask images from the front view and the lateral view as input, it...
is also possible to train the network by inputting only one binary mask image from either the front view or the lateral view for human reconstruction.

3. Results and Discussion
In order to fully evaluate the performance of our proposed approach, several tests with different inputs were carried. The nomenclature for our tests is summarized in Table 1. All the tests in this table use the same data set, which includes a training dataset with 4000 human body models and a test dataset with 300 human models. To evaluate the accuracy of the reconstructed body by BRCNN objectively, we used the Root Mean Squared Error (RMSE) as the metric to calculate the arithmetic square root of the difference between the reconstructed body and the ground truth, namely the original body. The result is shown in Figure 3.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (H_i - H_s)^2}$$

Table 1. Nomenclature for various tests.

| Name       | Input                        | Architecture                               |
|------------|------------------------------|--------------------------------------------|
| M-BRCNN-1  | Front-view Binary Mask       | Single Channel                             |
| M-BRCNN-2  | Lateral-view Binary Mask     | Single Channel                             |
| M-BRCNN-3  | Two-view Binary Masks        | Two-branches, Non-weight-sharing            |
| M-BRCNN-4  | Two-view Binary Masks        | Two-branches, Weight-sharing                |
| C-BRCNN-1  | Front-view Silhouette Sketch | Single Channel                             |
| C-BRCNN-2  | Lateral-view Silhouette Sketch | Single Channel                          |
| C-BRCNN-3  | Two-view Silhouette Sketches | Two-branches, Non-weight-sharing            |
| C-BRCNN-4  | Two-view Silhouette Sketches | Two-branches, Weight-sharing                |

Figure 3. The average RMSE of the test dataset and the RMSE of one selected sample. RMSE is very sensitive to the abnormal experimental error. Lower value of RMSE indicates better accuracy. Meanwhile, we also used RMSE to analyze the error between the reconstructed body and obtained reconstructed bodies rendered with an error map which can clearly observe the details of reconstruction. In this paper, we analyze a selected sample as an example and compare the pros and cons of each experiment in terms of its reconstruction result.

3.1. One View Reconstruction vs. Two View Reconstruction
Figure 4 gives one sample to illustrate the effectiveness of M-BRCNN-1, which has been trained with a front-view binary mask image as the input. Intuitively, the shape of the reconstructed model in Figure 4b is basically the same as the shape of the ground truth in Figure 4a. We visualized the error map in Figure 4a and Figure 4b. According to the chromatography, the closer the colour is to green, the smaller the error is. The closer the colour is to red and blue, the bigger the error is. From the front view of the model shown in Figure 4c, it is obvious that bigger error often occurs in the areas including knees, chest, and ankles. However, from the back view of the model shown in Figure 4d, we can see that in addition
to knees, the shoulders and the hip also possess relative bigger errors. Since the front-view binary mask image only contains the frontal contour information of the body and lacks the depth information, it is highly possible the generated human model lacks the ability to reconstruct accurate 3D topology in side view accordingly. That is why the error in lateral view (such as the chest, the knee and the hip) is significantly greater than that of the front view.

Figure 4. Illustration of M-BRCNN-1 result. (a) Ground truth; (b) result of M-BRCNN-1; (c) front view error map; (d) back view error map.

Figure 5. Illustration of M-BRCNN-2 result. (a) Ground truth; (b) result of M-BRCNN-2; (c) front view error map; (d) back view error map.

Figure 6. Illustration of M-BRCNN-3. (a) Ground truth; (b) result of M-BRCNN-3; (c) The front view of overlap of (a) and (b); (d) The back view of overlap of (a) and (b).
As is shown in Figure 5, the reconstructed human body model generated by M-BRCNN-2 from a lateral-view binary mask image is also plausible. In order to improve the accuracy, we used both front-view binary mask and lateral-view binary mask as the input to train the M-BRCNN-3. Each branch of M-BRCNN-3 does not share the weight. And the generated model is shown in Figure 6. The error map in Figure 6c and Figure 6d show that the error distribution of the model is relatively uniform (i.e., approximately ± 4.4057mm), which is less than the error of M-BRCNN-1 and M-BRCNN-2.

As is shown in Figure 7, this is also demonstrated by the C-BRCNN-1, C-BRCNN-2, and C-BRCNN-3, which use the contour sketches from a single view and two views as the input. The RMSE value shown in Figure 3 also objectively supports the above conclusions. Taking the sample mentioned above as an example, the RMSE of M-BRCNN-3 and C-BRCNN-3 is 1.8599mm and 2.2482mm, which are significantly smaller than M-BRCNN-1 (2.9294mm), M-BRCNN-2 (2.0014mm), C-BRCNN-1 (2.513mm) and C-BRCNN-2 (2.9991mm). The average RMSE of the test dataset also shows the same trend.

These results show that the BRCNN can successfully approximate human body from either a single view or two views, and the accuracy of the 3D human body reconstructed from two views are higher than that from the single view. In addition, the results from the front view are more accurate than that from the lateral view.

3.2. Weight-sharing Network and Non-Weight-sharing Network

The correlation between the front view and the lateral view was considered when capturing the features from different views in this paper. The BRCNN uses a weight-sharing mechanism, which means all parameters in the convolution core of two branches remain unchanged during training. As is shown in Table 1, both M-BRCNN-3 and M-BRCNN-4 take two binary masks from two views as input, but the former adopts the non-weight-sharing mechanism while the latter adopts the weight-sharing mechanism. According to the error map from front and back view of the human body generated by M-BRCNN-3 in Figure 8c and Figure 8d respectively, it is clear that the errors are mainly distributed from -4.4058mm to 4.4057mm. As for M-BRCNN-4 shown in Figure 8e and Figure 8f, the error is mainly distributed from -1.0686mm to 4.4058mm, which is obviously decreased.
This phenomenon has also been proved by C-BRCNN-3 and C-BRCNN-4 with the contour sketches from two views as the input as illustrated in Figure 9. The RMSE value shown in Figure 3 can be seen as an evidence to objectively support this conclusion.

From our observation, the weight-sharing BRCNN are better than the non-weight-sharing BRCNN, and the accuracy of the weight-sharing BRCNN is improved by 1.07% compared with the non-weight-sharing BRCNN with two views binary mask images as the input.

4. Conclusions
We presented a novel method for reconstructing 3D human body accurately from one or two binary mask images. It is achieved by building a weight sharing regression network (BRCNN) via CNN and it is used to learn a global mapping between the binary images from views and the PCA descriptors. At last, the descriptors were decoded to points cloud for the reconstruction of human body. The experimental results showed that compared with the existing human reconstruction technology, the accuracy increased by 1.07%, and the prediction results of the two views are better than those of a single view. It further reveals that the prediction results of the weight-sharing network are better than that of the network without weight-sharing.

In conclusion, the BRCNN with two branches is capable of capturing the main features from the front view and the lateral view respectively, while the weight-sharing mechanism can keep the correlation between two views and make the reconstructed human body more accurate.

Acknowledgement
This project was supported by the National Natural Science Foundation of China (Grant No. 61572124)

References
[1] Kakadiaris I A, Metaxas D N. Three-Dimensional Human Body Model Acquisition from Multiple Views[J]. International Journal of Computer Vision, 1998, 30(3): 191-218.
[2] Laurentini A. The visual hull concept for silhouette-based image understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(2): 150-162.
[3] Loper M, Mahmood N, Romero J, et al. SMPL: A Skinned Multi-Person Linear Model[C]// Acm Siggraph Asia Conference. 2015.

[4] Pishchulin L, Wquirer S, Helten T, et al. Building statistical shape spaces for 3D human modeling[J]. Pattern Recognition, 2017, 67(C):276-286.

[5] Dibra E, Jain H, Öztireli C, et al. HS-Nets: Estimating Human Body Shape from Silhouettes with Convolutional Neural Networks[C]// International Conference on 3d Vision. 2016.

[6] Dibra E, Jain H, Öztireli C, et al. Human Shape from Silhouettes Using Generative HKS Descriptors and Cross-Modal Neural Networks[C]// Computer Vision & Pattern Recognition. 2017.

[7] Ji Z, Qi X, Wang Y, et al. Shape-from-Mask: A Deep Learning Based Human Body Shape Reconstruction from Binary Mask Images[J]. 2018. arXiv:1806.08485 [cs.GR]

[8] Lecon Y, Boser B E, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.

[9] V. Mateev, I. Marinova, "Machine Learning in Magnetic Field Calculations," 2019 19th International Symposium on Electromagnetic Fields in Mechatronics, Electrical and Electronic Engineering (ISEF), Nancy, France, 2019, pp. 1-2.

[10] I. Marinova, V. Mateev, 3D model building of anatomical objects, 2009 XVI-th International Symposium on Electrical Apparatus and Technologies, SIELA 2009, Proceedings, 2, pp. 172-177.

[11] Robinette K M, Daanen H, Fubini E, et al. The Caesar Project: A 3-D Surface Anthropometry Survey[C]// International Conference on 3-d Digital Imaging & Modeling. 2002.

[12] Huang G, Liu Z, Maaten L V D, et al. Densely Connected Convolutional Networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2017.