**Shape Estimation of 3D Human Body Model Based on the Geometric Generation Model**

Shuai Zhao, Jifeng Sun, Hongshi Ou and Yinggang Li

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, Guangdong, China
Email: eesz@mail.scut.edu.cn

**Abstract.** In this paper, we propose a geometric generation model for the shape estimation of 3D human body. The model decouples the two processes: (1) Encoding/decoding process maps the samples between the image space $X$ and the latent (feature) space $Z$. This step achieves the dimension deduction. (2) The process of the Probability measure transformation transforms a fixed distribution $\zeta \in \mathcal{P}(Z)$ to any given distribution $\mu \in \mathcal{P}(Z)$. Before the shape estimation, U-V transform is done with Densepose, in order to transform the 2D human body image to 3D shape pose image.

**Keywords.** Geometric generation model; human body; 3D shape estimation; U-V transform.

1. **Introduction**

The original purpose of computer vision is to restore three-dimensional geometry from two-dimensional images. In the production and life of human society, humans are usually the leading role in multimedia. Realizing their shape and gesture is very important to express the message that human beings want to convey and to interact with the world. Recent methods [1] have shown rapid progress in estimating 2D major body joints, hand joints and facial features. At the same time, recent work can also estimate the joints and rough shape of the human body from a single image. But if we want to understand human behavior, we can’t just reconstruct the joints of the human body, we also need to know more specific details of the three-dimensional human body, including the complete 3D surface of the human body composition.

First of all, we need a 3D model that can represent body composition and body gesture. Secondly, we need a method to extract this model from a single image. The development of deep learning and manually labeled image data sets has led to the rapid development of two-dimensional “pose” estimation. The term “posture” usually refers to the line shape of the human body. This is not enough to understand the information contained in human posture. OpenPose extends it to include 2D hand joints and 2D facial features [1]. Although this captures more communication intentions, it does not support reasoning about surfaces and human interaction with the 3D world [1]. To assess precision, the data we require includes the overall image of the person and the relevant 3D image of the real object on the ground. We use a scanning system to capture 3D body shapes. We find that the performance of our method is obviously better than that of the related and weak models, and the real and appropriate results are obtained.

The work of ours is an important work for extracting a three-dimensional model of the various components of the body from a single two-dimensional image. Although many researchers bear on restoring postures and harsh human body shapes from a picture [2], less people focus on restoring...
specific shapes. Recently, there are several ways to restore posture from monocular video in real-time [3, 4]. Yet those methods need constant templates for each topic that are captured in advance. Other recent works [5, 6] have repaired the human body and clothes as a substitute for the SMPL mannequin [7], or use stereo pixels to represent [8, 9]. The V-Oxel-based method [9] often produces errors in the limbs of the body and requires the installation of the model after the event [9].

we estimate 3D human shape by using the following geometric generative model [10], which is the geometric description and contains two steps instructed in Section 3.2.

![Figure 1. Geometric generative model [10].](image)

The empirical probability distribution of the original space X is shown in figure 1. V is a submanifold belonging to space X. The process of transforming the original manifold probability distribution space X to the feature space Z is the process of encoding mapping. The function of f is to transform the original empirical distribution into hidden space μ.

2. The Shape Estimation of 3D Object Based on the Geometric Generation Model

We propose an image conversion method that could convert a human body photo to complete and specific body shape. Input a two-dimensional image, and then finally show the human body in the image with a real posture through our method. Our method is mainly focus on transforming the shape regression problem to a coherent picture conversion problem. A large number of experimental results show that the method is universal and robust. The problem of automatic reconstruction of human body details from a single image is studied in this paper [11]. Human shape reconstruction is widely used in virtual reality, augmented reality, scene analysis, virtual fitting and other fields [11]. In most cases, access needs to be fast and convenient, and it should look authentic. Human shape reconstruction is an extensive research field, which is usually combined with posture reconstruction. Clothes played a key role in the reconstruction of particulars. So we gave a detailed description of the reconstruction of clothing.

Figure 2 shows the processing the shape estimation based on U-V transform and geometric generative model:

![Figure 2. The shape estimation based on U-V transform and geometric generative model.](image)
2.1. U-V Transform
The first step of system of figure 2 is U-V transform which can be described as follows:

The surface algorithm visible in the lighting algorithm uses Lambert's formula \( i = s(L \cdot N) \) to get the light intensity. Among them, \( i \) refers to the light intensity, \( s \) refers to the surface shadow, \( L \) refers to the direction vector of the light, \( N \) refers to the surface normal vector, and "\( \cdot \)" refers to the inner product of the directivity.

\[
i = s(L \cdot N) ** n, \text{ for } n > 1
\]  

The formula is shown above, and the variable \( i \) is used to give the impression of a shiny surface. However, the adjustable range of this function is limited, and physical adjustments cannot be made. For the first time, BuiTuong Phong did research on more real lighting models. His reflection model is about producing highlights on a certain part of the surface of an object. Among them, the normal on this surface falls between the direction of the light source and the direction of the person's viewing angle. This is because actual surfaces tend to reflect more light in directions that form the same incident angle and reflection angle as the face normal [12]. This operation is achieved by simulating a virtual light source between the light source and the direction of the human eye, by increasing the energy of the light source to make the highlights more clear and eye-catching:

\[
i = s(L \cdot N) + g(L' \cdot N) ** n
\]  

In the above formula, \( L' \) is the direction of the virtual light source, and \( g \) is the glossiness of the surface of the object, ranging from 0 to 1. This method works well on smooth surfaces, but lacks authenticity on highly polished surfaces. This is mainly due to the fact that surrounding objects and distributed light sources have no real reflections [12, 13]. The simulation of the reflection in the curved surface requires accurate modeling of the characteristics of the surface, and it is necessary to obtain an accurate normal vector at all points on the surface.

However, the subdivision algorithm can give the appropriate appearance normal for each pixel. This is the first algorithm to provide proper information for simulating curved mirrors. Combine the vector from the object to the spotter and the vector from the appearance to the normal. These vectors belong to the images. This method is used to decide which parts of the circumstance are reflexed by the surface. For the surface normal vector \((X_n, Y_n, Z_n)\), observation point \((1, 0, 0)\) and reflection direction \((X_r, Y_r, Z_r)\), the relationship between the three is \(X_r = 2 * X_n * Z_n\), \(Y_r = 2 * Y_n * Z_n\), \(Z_r = 2 * Z_n * Z_n - 1\). And it determines the direction of reflection into the eyes. We also need to find out which part of the environment produced this ray. Therefore, an environment model is needed to simulate the neighbouring light. Obviously, the results obtained from different angles on the surface are different. However, if the environment is made up of nearby neighbors and light rays, the environment can be modeled as a stereo projection around two target objects. In other words, if you think of the environment as a spherical space, then the object is at its center, and a picture of the environment has been drawn inside the sphere. These simplifications allow the environment to be modeled as a 2D intensity map, which is indexed by the polar angle of the reflected rays \((X_r, Y_r, Z_r)\).

2.2. Geometric Generation Model [10]
we estimate 3D human shape by using the geometric generative model [10], which contains two steps shown in figure 1.

The first step is to map the samples from the image space \(X\) to the feature space \(Z\) by using the deep neural networks. We can analyze the dimension through this way.

The second step is to transform the probability measure. We all know that the information contained in image obey a certain distribution. And we’d like to switch the original distribution to a given distribution by mapping. We will achieve this step by geometric methods.

2.2.1. Pose and Shape Reconstruction. In the early work, the 2D pose was generated by full or partial manual clicks [14], and later this process was automatically generated by 2D landmark detection from
the depth neural network [15]. Recently, the SMPL [7] model has become part of the network framework [2]. This further realizes the automation and robustness of the process. Clothes and hair can be obtained through optimization-based methods [6, 11, 16]. Another recent study estimates poses and shapes in the form of stereo pixels [8, 17, 18], and this method increases the level of detail of the clothes while reducing details. In [19], the author mitigates this drawback by using the predicted normal map to enhance the visible part.

2.2.2. Face Reconstruction. Some recent facial reconstruction methods use shadow-based thinning methods for geometric enhancement. For example, in fitting [20] or refinement by composite analysis, or in a trained neural network [21, 22]. Recent work is also related to our method, which embeds a differential face renderer in the neural network to evaluate the geometry and albedo correction relative to the basic model example [23]. Or learn the basics of distinguishing geometry and albedo from the faint details in the video [13].

2.2.3. Clothing Reconstruction and Styling. According to 3D scanning and RGB-D [24], the body shape under the clothing has been estimated by combining a separate clothing layer [25] without [26]. The method in [27] can fit clothes details to nude offsets. Reference [28] introduces a model that encodes shapes, clothing sketches and clothing models in a potential coding form, which makes interactive clothing design possible. CNN is used in UV space or data-driven optimization method is used to predict high-frequency wrinkles directly in 3D space. All of these methods [27, 29, 30] are realistic animations for clothing, and can only predict clothing in isolation [29, 30]. Normal and depth recovery based on learning [31] or grid [32] has been demonstrated, but it is also applicable to individual clothing. Instead, by mapping from image, our method rebuilds the detailed shape of the whole body from a single image.

2.2.4. Human Body Pose Estimation Implementation Scheme. The generative network $G$ in the generative confrontation network can be equivalent to the optimal transmission mapping in the optimal batch transmission (OMT). The transformation parameter $g_\theta$ from the image space to the feature space generated by generative network corresponds to the most optimal batch transmission (OMT). Optimal transmission mapping, the mapping contains coding $g_\theta : Z \rightarrow X$ and a uniform transformation of the probability distribution $T : Z \rightarrow Z'$, $T\#\zeta = \mu$, where $Z$ is the image manifold and $Z'$ is the image of the uniformized sum of the discrete sampling probability distribution and Data manifold. $X$ is the dimensionality reduction parameter manifold, $\zeta$ is the probability distribution of $Z'$, $\mu$ is the distribution after the probability distribution is homogenized, $\#$ is the probability distribution transformation under a certain mapping (such as $T$) and a certain distribution (such as $\zeta$).

Twindom DBA 3D human body data set will be used to do the experiments on U-V mapping tests with Densepose and the shape estimation of 3D human body in future work.

3. Conclusion
This paper mainly studies the 3D shape reconstruction of the human body. A method of geometrically generating models is proposed to reconstruct the shape of three-dimensional human body. The geometric model includes the encoding and decoding process of the image space and the conversion process of the probability measure of the image distribution. The conversion process from image space to feature space is accomplished by using deep neural networks. Our method focuses on transforming the shape regression problem into a coherent picture conversion problem. The expected effect is to convert the input two-dimensional image into a complete and specific body shape, and show the human body in the image in a real posture. The experiment in this article is to be completed and will be carried out in the follow-up project research work. The main research direction of the subsequent work is to convert the branch fragments of the human body after U-V mapping into three-dimensional space through probability measurement.
Acknowledgement
We thank the inspiring discussions with Dr. Jifeng Sun from Tsinghua University, all members of the Second Laboratory of Deep Learning and Computer Vision and so on. The geometric generative model experiment is conducted by Mr. Shuai Zhao from the South China University of Technology. The project is supported by NSFC No. 62071183.

References
[1] Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman A A, Tzionas D and Black M J 2019 Expressive body capture: 3D hands, face, and body from a single image Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp 10975-10985.
[2] Kanazawa A, Black MJ, Jacobs DW and Malik J 2018 End-to-end recovery of human shape and pose Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 7122-7131.
[3] Habermann M, Xu W, Zollhofer M, Pons-Moll G and Theobalt C 2019 Livecap: Real-time human performance capture from monocular video ACM Transactions on Graphics (ToG) 38 (2) 1-17.
[4] Xu W, Chatterjee A, Zollhöfer M, Rhodin H, Mehta D, Seidel H P and Theobalt C 2018 Monoperfcap: Human performance capture from monocular video ACM Transactions on Graphics (ToG) 37 (2) 1-15.
[5] Alldieck T, Magnor M, Bhatnagar B L, Theobalt C and Pons-Moll G 2019 Learning to reconstruct people in clothing from a single RGB camera Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp 1175-1186.
[6] Alldieck T, Magnor M, Xu W, Theobalt C and Pons-Moll G 2018 Video based reconstruction of 3D people models Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 8387-8397.
[7] Loper M, Mahmood N, Romero J, Pons-Moll G and Black M J 2015 SMPL: A skinned multi-person linear model ACM Transactions on Graphics (ToG) 34 (6) 1-16.
[8] Natsume R, Saito S, Huang Z, Chen W, Ma C, Li H and Morishima S 2019 Siclope: Silhouette-based clothed people Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp 4480-4490.
[9] Varol G, Romero J, Martin X, Mahmood N, Black M J, Laptev I and Schmid C 2017 Learning from synthetic humans Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 109-117.
[10] Lei N, Su K, Cui L, Yau S T and Gu X D 2019 A geometric view of optimal transportation and generative model Computer Aided Geometric Design 68 1-21.
[11] Alldieck T, Pons-Moll G, Theobalt C and Magnor M 2019 Tex2shape: Detailed full human body geometry from a single image Proceedings of the IEEE/CVF International Conference on Computer Vision pp 2293-2303.
[12] Blinn J F and Newell M E 1976 Texture and reflection in computer generated images Communications of the ACM 19 (10) 542-547.
[13] Tewari A, Bernard F, Garrido P, Bharaj G, Elgarhib M, Seidel H P and Theobalt C. 2019 FML: Face model learning from videos Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp 10812-10822.
[14] Guan P, Weiss A, Balan A O and Black M J 2009 Estimating human shape and pose from a single image 2009 IEEE 12th International Conference on Computer Vision pp 1381-1388.
[15] Pishchulin L, Insafutdinov E, Tang S, Andres B, Andriluka M, Gehler P V and Schiele B 2016 DeePCut: Joint subset partition and labeling for multi person pose estimation Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 4929-4937.
[16] Alldieck T, Magnor M, Xu W, Theobalt C and Pons-Moll G 2018 Detailed human avatars from monocular video 2018 International Conference on 3D Vision (3DV) pp 98-109.
[17] Jackson A S, Manafas C and Tzimiropoulos G 2018 3D human body reconstruction from a single image via volumetric regression *Proceedings of the European Conference on Computer Vision (ECCV)*.

[18] Varol G, Ceylan D, Russell B, Yang J, Yumer E, Laptev I and Schmid C 2018 Bodynet: Volumetric inference of 3D human body shapes *Proceedings of the European Conference on Computer Vision (ECCV)* pp 20-36.

[19] Zheng Z, Yu T, Wei Y, Dai Q and Liu Y 2019 Deephuman: 3D human reconstruction from a single image *Proceedings of the IEEE/CVF International Conference on Computer Vision* pp 7739-7749.

[20] Sela M, Richardson E and Kimmel R 2017 Unrestricted facial geometry reconstruction using image-to-image translation *Proceedings of the IEEE International Conference on Computer Vision* pp 1576-1585.

[21] Huynh L, Chen W, Saito S, Xing J, Nagano K, Jones A and Li H 2019 Mesoscopic facial geometry inference using deep neural networks *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 8407-8416.

[22] Sengupta S, Kanazawa A, Castillo C D and Jacobs D W 2018 SfSNet: Learning shape, reflectance and illuminance of faces in the wild *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 6296-6305.

[23] Tewari A, Zollhöfer M, Garrido P, Bernard F, Kim H, Pérez P and Theobalt C 2018 Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 2549-2559.

[24] Yu T, Zheng Z, Zhong Y, Zhao J, Dai Q, Pons-Moll G and Liu Y 2019 Simulcap: Single-view human performance capture with cloth simulation *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp 5504-5514.

[25] Pons-Moll G, Pujades S, Hu S and Black M J 2017 ClothCap: Seamless 4D clothing capture and retargeting *ACM Transactions on Graphics (ToG)* 36 (4) 1-15.

[26] Zhang C, Pujades S, Black M J and Pons-Moll G 2017 Detailed, accurate, human shape estimation from clothed 3D scan sequences *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 4191-4200.

[27] Yang J, Franco J S, Hétroy-Wheeler F and Wuhrer S 2018 Analyzing clothing layer deformation statistics of 3D human motions *Proceedings of the European Conference on Computer Vision (ECCV)* pp 237-253.

[28] Wang TY, Ceylan D, Popovic J and Mitra N J 2018 Learning a shared shape space for multimodal garment design arXiv preprint arXiv:1806.11335.

[29] Jin N, Zhu Y, Geng Z and Fedkiw R 2020 A pixel-based framework for data-driven clothing *Computer Graphics Forum* 39 (8) 135-144.

[30] Lahner Z, Cremers D and Tung T 2018 Deepwrinkles: Accurate and realistic clothing modeling *Proceedings of the European Conference on Computer Vision (ECCV)* pp 667-684.

[31] Bednarik J, Fua P and Salzmann M. 2018 Learning to reconstruct texture-less deformable surfaces from a single view *2018 International Conference on 3D Vision (3DV)* pp 606-615.

[32] Daněřek R, Dibra E, Öztireli C, Ziegler R and Gross M 2017 Deepgarment: 3D garment shape estimation from a single image *Computer Graphics Forum* 269-280.