Recovery of Detailed Posture and Shape from Motion Video Images by Deforming SMPL

Yumi Ando, Fumihiko Sakaue and Jun Sato
Nagoya Institute of Technology, Japan
{*sakaue, junsato*}*@nitech.ac.jp*

Keywords: Detailed Shape Recovery from Video Images, SMPL, Shape Representation by Deformation.

Abstract: In this research, we propose a method for estimating detailed human shape and posture from video images of a person in motion. The SMPL (Skinned Multi-Person Linear Model) model can represent various body shapes with a small number of parameters, but it cannot represent detailed information such as the subject's clothing or hairstyle. In this research, we separate such detailed deformations into deformations common to all time periods and temporary deformations that appear at different times, and recover each of them to realize detailed human shape recovery from video images of people shot with various postures.

1 INTRODUCTION

In recent years, free viewpoint video technology has been attracting attention in situations such as watching sports and live performances, where the audience can enjoy the video from any viewpoint they wish. Such images are used for a variety of purposes, especially in the case of sports games, where a person is often the subject of the image. Therefore, the generation of free viewpoint video requires that a person is captured by multiple cameras and that 3D information such as shape and posture is recovered from these images. In order to achieve this, images taken from various directions by a large number of synchronized cameras are usually required. Therefore, it is difficult to recover detailed 3D information from only a common camera fixed at a certain location. To solve this problem and make it more practical, research is underway to recover the 3D shape and posture of a person from a single viewpoint image.

Let us consider the case in which a single camera is used to capture a scene of a person in motion. Although the postures of these images are different at each time, they can be treated as a group of images taken from various directions because the relative position to the camera is changing. Since these images include detailed shape information of the object, it is considered possible to analyze these images to restore the detailed shape of the object, similar to stereo reconstruction, even if the images were shot from a single viewpoint. In this study, we use these images taken from various postures to recover the shape of a person, including personal details such as clothing and hair style, as well as the person's posture.

In addition, the use of shading information is also considered in order to recover detailed shapes, including the clothing of a person. Shading in an image is generated depending on the shape of the object. Therefore, it is known that it is possible to obtain dense shape information by analyzing this shading information. When we consider the shape restoration of a human subject in this study, it is difficult to restore fine information such as wrinkles in the clothing. Therefore, in this study, we will examine the recovery of such detailed shape information by using shading information. The goal of this study is to recover detailed human figures by using time-series images taken in various postures.

2 RELATED WORKS

A method for simultaneous estimation of body shape and posture using a generalized model of the human body as the initial shape has been proposed(Bogo et al., 2016). Although this method can estimate accurate and natural postures, it is difficult to recover detailed shapes including clothing and hair because the shape parameters used for shape representation are only those embedded in the generalized model.

On the other hand, many methods have been proposed to estimate 3D human pose and shape using multi-view video(Wang et al., 2023). Multi-

Ando, Y., Sakaue, F. and Sato, J.
Recovery of Detailed Posture and Shape from Motion Video Images by Deforming SMPL.
DOI: 10.5220/0013322400003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 751-757
ISBN: 978-989-758-728-3; ISSN: 2184-4321

view video can be used to estimate pose and shape more accurately because it contains depth information directly. However, this method requires recording system by multiple cameras and synchronising the videos, which increases the recording cost. Therefore, it is desirable to recover the 3D pose and shape from the single-view video.

Furthermore, as one of the methods to directly restore shapes from images, an optimization method based on differentiable rendering has recently been proposed(Jiang et al., 2020). This method is composed so that the rendering process of the image is differentiable. This makes it possible to directly minimize image errors using gradient methods, etc., and to easily achieve shape restoration. Therefore, it is possible to estimate 3D shapes not only from images taken from multiple viewpoints, but also from images taken from only a single viewpoint. In addition, it can be easily combined with a learning-based method using a neural network, making it applicable to a variety of scenes. However, accurate restoration from a single viewpoint image requires a method that combines trained models, which causes biases in the training data to be reflected in the restoration results(Robinette et al., 2002).

In addition, 3D posture and body shape estimation from single-view images using a learning model may produce incorrect estimation results owing to factors like the race of the person to be reconstructed(Zengin et al., 2016; Robinette et al., 2002). Our study aims to alleviate this problem and bring the estimation results closer to the correct answer by treating a group of time-series images taken from a single viewpoint as a group of images taken from practically different directions.

3 3D RECONSTRUCTION BASED ON DIFFERENTIABLE RENDERING

First, we describe an overview of the differentiable rendering (Kato et al., 2020) that is used in this research for shape restoration. Rendering in computer graphics refers to the process of converting information such as shape, illumination, and camera information that compose a 3D space into a 2D image. Because rendering usually involves discrete processing, the image obtained by rendering cannot be directly differentiated by the information to be estimated, such as shape. Differentiable rendering, on the other hand, replaces the inherently discrete operation of rendering with a differentiable operation in the 3D representation using a mesh. This makes it possible to differentiate the error between the input image and the rendered image in the same way, thus enabling direct minimization of image errors using gradient descent and other similar methods.

This method can be combined with various methods of optimization using gradient descent methods. Therefore, 3D restoration methods using differentiable rendering are often used in combination with 3D shape representation methods using neural networks. In this study, the posture and shape of a person are represented based on the SMPL model, which is then optimized using differentiable rendering to estimate the detailed shape of the object.

4 REPRESENTATION OF HUMAN POSTURE AND SHAPE USING SMPL

Next, we describe the SMPL, the posture and shape representation model of the human body used in this study. The Skinned Multi-Person Linearmodel (SMPL)(Loper et al., 2015; Pavlakos et al., 2019) can represent various body shapes by adding shape changes that can be expressed with a few parameters to a standard template mesh. By assigning posture parameters to the shape, models with a variety of postures can be easily generated. Let $\bar{T}$ be a template mesh, and consider the possibility of representing 3D models of various shapes and postures by changing this mesh. In this case, let β be the shape deformation parameter and θ be the posture parameter, and let $M(\beta,\theta)$ be the shape represented by these parameters.

$$M(\beta,\theta) = W(\bar{T} + B_s(\beta),\theta), \tag{1}$$

where M is a function that represents the deformation of the shape represented by β, and represents the shift of each vertex in the template mesh. Various human body shapes can be created by using M and β as inputs. In the SMPL model, each mesh representing a shape is associated with a human body, so that various postures can be represented by changing the posture parameter θ, which indicates the angle of each joint. This makes it possible to create arbitrary postures for deformed shapes, and to create human body shapes with various body shapes and postures.

The differentiable rendering described above can be combined with SMPL to estimate the appropriate shape and orientation for a given image. Let I be the input image and R be the rendering function by the differentiable renderer, and define the image error ε as follows:

$$\varepsilon = |I - R(W(\bar{T} + B_s(\beta),\theta))|, \tag{2}$$

where W is a function that parametrically deforms the mesh by θ, so it is differentiable by θ. Furthermore, R is differentiable as well. This makes it easy to compute the derivative of the deformation parameter β and the orientation parameter θ with respect to the rendered image and minimize ε by gradient descent-based optimization methods. The obtained β and θ represent the pose and shape of the input image I.

5 ESTIMATION OF SMPL INCLUDING DETAILED SHAPE

5.1 Detailed Human Shape Representation Using SMPL

The SMPL model described above can represent various body shapes by changing the parameter β that represents the shape of the person. However, the deformable shapes are limited to some extent by the parameter β included in SMPL alone. Therefore, it is not possible to represent detailed shape deformations such as clothing and hair. In addition, although the SMPL is composed by statistically processing a large number of human shapes, it is known that the dataset used to compose the SMPL has some biases such as racial bias, which may prevent appropriate representation. Therefore, in this study, in addition to the shape representation in the SMPL model using β and θ to represent posture and body shape, a more detailed shape representation is achieved by deforming the template mesh more directly. For this purpose, we introduce ΔT and ΔT_i that represent the displacement of each vertex in the template mesh $\bar{T}$. The ΔT represents a common shape at all times and is used to represent the target person's detailed shape. The ΔT_i represents the shape that changes at each time, and represents temporary fluctuations such as the twisting of clothes or hair due to movement. Figure1 shows an overview of the method. In this figure, the left image is represented by SMPL only, and the center image is obtained by adding ΔT. The image on the right shows the time-specific transform, ΔT_i, added. The combination of these two methods provides detailed shape estimation that cannot be achieved with SMPL.

5.2 Shape Estimation Using Differentiable Rendering

Now, given n time-series silhouette images $I_i^b (i = 1, \cdots, n)$, consider estimating the shape of a person from these images. The error ε_b between all input images and the SMPL silhouette rendering images is

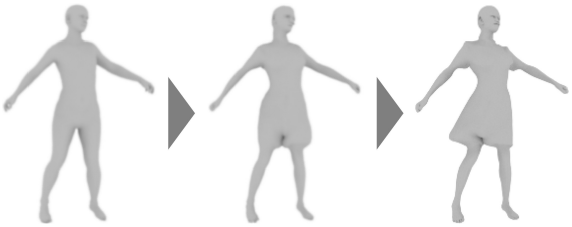


Figure 1: Detailed shape representation by SMPL deformation: the left image is represented by SMPL only, the center image by SMPL+ΔT, and the right image by SMPL+ΔT + ΔT_i.

defined as follows:

$$\varepsilon_b = \sum_{i=1}^{n} |I_i^b - R^b(W(\bar{T} + \Delta T + \Delta T_i + B_s(\beta), \theta_i))| \quad (3)$$

where R^b is a differentiable silhouette renderer. By minimizing this error, we can estimate ΔT that matches all images and also estimate the detailed shape deformation ΔT_i at each time that cannot be represented by ΔT alone.

To estimate the color and detailed shading information of a person, $I_i^c (i = 1, \cdots, n)$ of n time-series RGB images are used. The error ε_c between all input images and SMPL rendered images is defined as follows:

$$\varepsilon_c = \sum_{i=1}^{n} |I_i^c - R^c(W(\bar{T} + \Delta T + \Delta T_i + B_s(\beta), \theta_i), \rho, S| \quad (4)$$

where R^c is a differentiable color image renderer that takes as input the reflectance ρ of each mesh and the light source parameter S in addition to shape information. It is assumed that the object surface can be represented by a diffuse reflection model. Under this assumption, the observed color does not change when the viewpoint changes. Therefore, it is determined by the reflectance ρ of each mesh and the lighting environment. The lighting environment is assumed to be unchanged at all times, and is a constant parameter across all time periods. The scene is assumed to be illuminated by ambient light and one light source, and the light parameter S includes the intensity of the ambient light and the position and intensity of one light source. By minimizing the errors expressed by ε_b and ε_c using the gradient descent method, the parameters of attitude, shape, color, and lighting environment can be estimated.

However, the large number of meshes in the SMPL provides a high degree of freedom in estimation, and overfitting to the image occurs when ΔT and ΔT_i are estimated at the same time. Therefore, the Laplacian regularization (Nicolet et al., 2021) is used for the mesh shape to suppress excessive deformation.

Let L be the term related to this Laplacian regularization, the error ε to be minimized is expressed by the following equation.

$$\begin{aligned}\varepsilon = \sum_{i=1}^{n} & |I_i^b - R^b(W(\bar{T}+\Delta T+\Delta T_i + B_s(\beta),\theta_i))| \\ & + |I_i^c - R^c(W(\bar{T}+\Delta T+\Delta T_i + B_s(\beta),\theta_i),\rho,S| \\ & + L(T+\Delta T+\Delta T_i)\end{aligned} \quad (5)$$

By minimizing it, we can estimate a smooth shape that fits all images.

5.3 Regularization for Detailed Shapes

As mentioned in the previous section, ΔT and ΔT_i introduced for this study have a high degree of freedom because all vertices of the template mesh can be moved. In addition, since the sum of the two transformations represents the shape at each time, it is possible to estimate an image that matches each time even if only ΔT_i is used. In addition, the aforementioned Laplacian regularization can smooth the shape when deforming the mesh, but it does not have the constraints of the human body, such as the positions of body parts, and may generate unnatural shapes for the human body. To prevent this, regularization is introduced to prevent significant deformation from the initial shape of SMPL. Specifically, the distance between the initial SMPL shape $\bar{T}$ and the deformed shape $\bar{T}+\Delta T$ common to all times is minimized. Here, we define the distance D_r for regularization of the two shapes T_1 and T_2 as follows:

$$D_r(T_1,T2) = ||T_1 - T_2||^2 + C(T_1,T_2) + E(T_1,T_2) \quad (6)$$

where C is the Chamfer distance of the two shapes and E is the earth mover's distance(Solomon et al., 2014). That is, the following equation is minimized for shape estimation.

$$\begin{aligned}\varepsilon = \sum_{i=1}^{n} & |I_i^b - R^b(W(\bar{T}+\Delta T+\Delta T_i + B_s(\beta),\theta_i))| \\ & + |I_i^c - R^c(W(\bar{T}+\Delta T+\Delta T_i + B_s(\beta),\theta_i),\rho,S| \\ & + L((T+\Delta T) \\ & + D_r(\bar{T},\bar{T}+\Delta T)\end{aligned} \quad (7)$$

The time deformation represented by ΔT_i should represent the shape that cannot be represented by ΔT by minimizing the deformation. Therefore, we minimize the deformation here as well.

$$\begin{aligned}\varepsilon = \sum_{i=1}^{n} & |I_i^b - R^b(W(\bar{T}+\Delta T+\Delta T_i + B_s(\beta),\theta_i))| \\ & + |I_i^c - R^c(W(\bar{T}+\Delta T+\Delta T_i + B_s(\beta),\theta_i),\rho,S| \\ & + L((T+\Delta T) \\ & + D_r(\bar{T},\bar{T}+\Delta T) \\ & + D_r(\bar{T}+\Delta T,\bar{T}+\Delta T+\Delta T_i)\end{aligned} \quad (8)$$

This allows us to estimate ΔT and ΔT_i separately.

5.4 Optimization Considering the Degrees of Freedom of the Target

Finally, we consider minimization methods for the error. The evaluation equation presented in the previous section is not convex and has many local solutions. Therefore, if all parameters are optimized at the same time, there is a high probability of falling into a local solution. Therefore, estimation is performed in order starting from the parameters with the lower degrees of freedom. After fixing the estimated parameters, the parameters with higher degrees of freedom are estimated in turn. Finally, all parameters are estimated by optimizing them simultaneously. Considering the number of parameters, the posture parameter θ, the SMPL deformation parameter β, ΔT, and ΔT_i are estimated in turn. After the shape parameter is estimated from the silhouette information, the reflectance parameter ρ and the light source parameter S are estimated simultaneously by introducing the RGB error. Finally, the final solution is obtained by simultaneously optimizing all parameters of the loss function for the silhouette image and the loss function for the RGB image. The above methods can estimate posture, shape, and texture from time-series images.

6 EXPERIMENTAL RESULTS

6.1 Environment

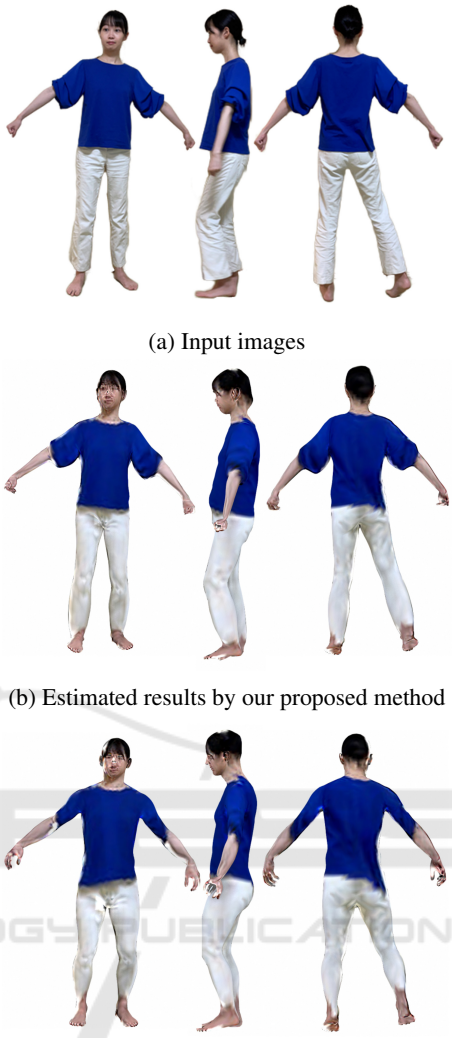
We present the results of an experiment in which we estimated the posture and shape of a person in the input images using the method described above. For the input images, we used 20 images taken from a single viewpoint of a person half-turning in place, and 25 images taken from a single viewpoint of a person moving as if swinging from side to side. Figure 2a and Fig. 4a show some of these images. These images were taken with a smartphone camera. The camera was not fixed on a tripod, but was held in the hand. Therefore, the detailed camera parameters differ from time to time. The movement of the person was made so that the distance between the camera and the person was kept constant to some extent.

We utilized Mitsuba3 as the differentiable renderer and adopted the Adam algorithm as the optimization method for minimizing the cost functions.

6.2 Results

The results of the posture and shape reconstruction using the proposed method are shown in Fig. 2. Figure 2c shows the result of the restoration using only SMPL, and Fig. 2b shows the result of the simultaneous optimization of posture, shape and texture parameters using the proposed method. The proposed method accurately reconstructs the shape of the subject's physical features, clothing, and hair style, while SMPL alone results in shape differences from the input. Figure 3 shows the result of viewing the generated 3D model from a different aspect. It can be seen that the natural 3D shape has been restored from different angles.

Figure 4 and Fig. 5 shows the result of restoring another object. In this experiment, a subject wearing clothing with large time variability was restored. The recovered shape shown in Fig.4 clearly represents the movement of the clothes represented by ΔT_i. Figure 5a and 5b show the results of deformation using only the common shape ΔT at all times, while Fig.4b and 4c show the results when the shape ΔT_i that changes at each time is also estimated. These results show that the introduction of ΔT_i can represent shape changes such as skirt swaying, which could not be represented by ΔT alone. However, the detailed wrinkles of the clothes are not sufficiently recovered. This may be due to the fact that the image used as the input for this study did not have much variation in shading, and the shading information was expressed as a variation in reflectance. In order to restore the image using shadows, it is necessary to reduce the degree of freedom by introducing regularization for reflectance, which will be discussed in the future.



(a) Input images

(b) Estimated results by our proposed method

(c) Estimated results by SMPL

Figure 2: Input images and estimated 3D shapes.

6.3 Evaluation

In order to evaluate the proposed method in more detail, we present the results of evaluation using synthetic images. We used a 3D character model as an input image to generate images from various viewpoints and postures, from which the shape was restored. Chamfer distance was used to compare the two shapes. For comparison, the Chamfer distance was calculated for the results of shape estimation using only the deformation parameter of SMPL. In order to eliminate the effect of the posture, a basic posture was applied to the recovered shapes for comparison. In order to compare the accuracy of the posture estimation, the distances were calculated for each of the 3D shapes deformed to the posture at each time, and the average of the distances was calculated.

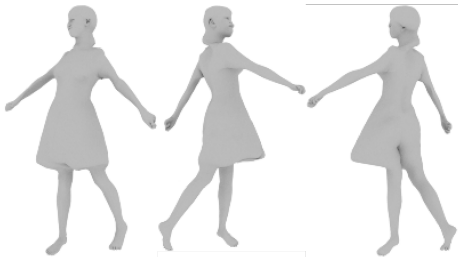
Figures 6a and 6b show the differences in the shapes at the basic posture and the differences in the shapes for the restoration results at a certain time. The blue shape is the correct shape and the orange shape is the estimated shape. The left image is the result of the estimation using only SMPL, and the right image is the result of the estimation using the proposed method. The results show that the proposed method is able to represent changes in hair style, etc., and that it is able to estimate a shape that is close to the input shape.

The results of the comparison are shown in Table 1 and 2. These values are the mean of the distances from the vertices in the input model to the nearest neighbors in the estimated model. The height of the input model is approximately 175 cm. The results show that the proposed method is able to estimate a shape closer to the input in the case of the basic pos-

Figure 3: Observation results from a different perspective.

(a) Input images

(b) 3D shape of $\bar{T} + \Delta T + \Delta T_i$ for a frame

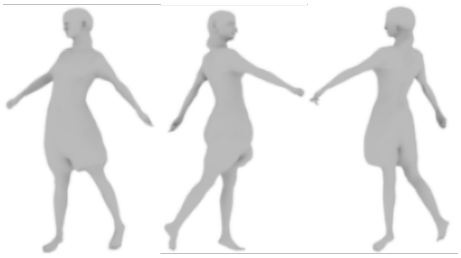


(c) 3D shape of $\bar{T} + \Delta T + \Delta T_i$ without texture

Figure 4: Estimated shape with specific deformation.

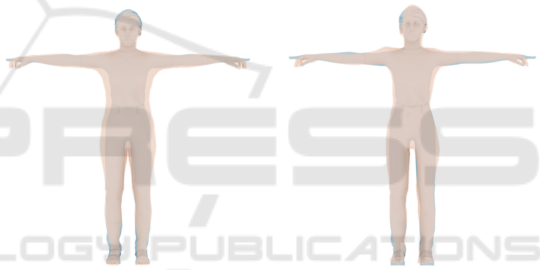
ture. The error in the evaluation at different times is also smaller. The reason why the distance is larger in this result than in the case of the basic posture is that the error in the estimated posture is reflected in the error in the shape. Therefore, it can be said that this result is also reflected in the posture estimation result. Considering this, it can be expected that the proposed method can not only estimate the shape of

(a) 3D shape of $\bar{T} + \Delta T$



(b) 3D shape of $\bar{T} + \Delta T$ without texture

Figure 5: Estimated shape without specific deformation.



(a) Comparison in basic posture

(b) Comparison in specific posture in a frame

Figure 6: Comparison between the estimated shape and the correct shape. Orange shape shows the estimated shape and blue shape shows the correct shape.

the object, but also improve the accuracy of the posture estimation. These results confirm that the proposed method can be used to estimate detailed shape and posture from time-series images that include motion.

Table 1: Comparison in basic posture using only ΔT [cm].

SMPL	Proposed model
4.00	2.95

Table 2: Average of errors at each time when using ΔT_i [cm].

SMPL	
8.12	8.02

7 CONCLUSIONS

In this study, we proposed a method for estimating the detailed shape and posture of a person from time-series video images in which the subject's posture changes. In order to further improve the accuracy of estimation, we plan to study more effective methods for representing and estimating shading information and for estimating posture.

REFERENCES

Bogo, F., Kanazawa, A., Lassner, Christoph andGehler, P., Romero, J., and Black, M. J. (2016). Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing.

Jiang, Y., Ji, D., Han, Z., and Zwicker, M. (2020). Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kato, H., Beker, D., Morariu, M., Ando, T., Matsuoka, T., Kehl, W., and Gaidon, A. (2020). Differentiable rendering: A survey. *CoRR*, abs/2006.12057.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16.

Nicolet, B., Jacobson, A., and Jakob, W. (2021). Large steps in inverse rendering of geometry. *ACM Trans. on Graphics(TOC)*, 40(6):1–13.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D., and Black, M. J. (2019). Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Robinette, K., Blackwell, S., Daanen, H., Boehmer, M., and Fleming, S. (2002). Civilian american and european surface anthropometry resource (caesar), final report. volume 1. summary. page 74.

Solomon, J., Rustamov, R., Guibas, L., and Butsche, A. (2014). Earth mover's distances on discrete surfaces. *ACM Trans. on Graphics(TOG)*, 33(4):1–12.

Wang, H.-K., Huang, M., Zhang, Y., and Song, K. (2023). Multi-view 3d human pose and shape estimation with epipolar geometry and mix-graphormer. In *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pages 28–32.

Zengin, A., Pye, S. R., Cook, M. J., Adams, J. E., Wu, F. C. W., O'Neill, T. W., and Ward, K. A. (2016). Ethnic differences in bone geometry between white, black and south asian men in the uk. *Bone*, 91:180 – 185.