3D Human Body Reconstruction from Head-Mounted Omnidirectional Camera and Light Sources

Ritsuki Hasegawa, Fumihiko Sakaue and Jun Sato
Nagoya Institute of Technology, Nagoya 466-8555, Japan

Keywords: Human Body, Head-Mounted Camera, Shadow, SMPL, Light Source, Omnidirectional Camera.

Abstract: In this paper, we propose a method for reconstructing a whole 3D shape of the human body from a single image taken by a head-mounted omnidirectional camera. In the image of a head-mounted camera, many parts of the human body are self-occluded, and it is very difficult to reconstruct the 3D shape of the human body including the invisible parts. The proposed method focuses on the shadows of the human body generated by the light sources in the scene and uses it to perform highly accurate 3D reconstruction of the whole human body including the hidden parts.

1 INTRODUCTION

HEREHEREHEREIn recent years, 3D human pose analysis has been utilized for various applications. In sports science, human pose analysis is used to improve athletes' form, and in security, pose analysis is used to detect dangerous human behavior.

Many methods of 3D human pose estimation have been proposed in the past. The standard method is to place cameras around the human body and recover human poses from images taken by these cameras (Elhayek et al., 2015a; Mehta et al., 2017a; Mehta et al., 2017b). However, these methods are difficult to use in real-world environments because of the space required to place the cameras and the fact that the cameras cannot be moved.

3D pose estimation methods using self-attached devices such as inertial measurement units (IMUs) (von Marcard et al., 2017) and Structure-From-Motion (SFM) (Shiratori et al., 2011) that uses multiple cameras have also been proposed. These methods have problems such as the need for prior calibration and the inability to perform real-time estimation due to pose optimization after measurement.

Recently, Mo^2cap^2 (Xu et al., 2019) has been proposed for solving these problems. It uses a single fisheye camera mounted on the human head and uses deep learning to achieve a highly accurate real-time estimation of human poses. However, as shown in Fig. 1 (a), many parts of the human body are often hidden in the image obtained from the head-mounted camera, and accurate estimation is not possible in such cases. Therefore, in this paper, we propose a method for using a combination of object and shadow images to accurately recover the entire shape of the human body from a single head-mounted image.

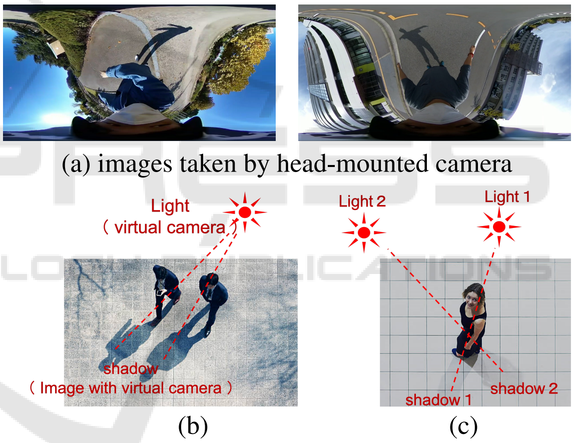
(a) images taken by head-mounted camera
(b)
(c)

Figure 1: Head-mounted camera images and shadows.

As shown in Fig. 1 (b), in many cases, the head-mounted camera image contains not only the image of the object but also a shadow of the object. The shadow is nothing but image information obtained by observing the object from the viewpoint of a light source. In other words, if the light source is considered a camera, the shadow is an image from the viewpoint of this virtual camera. Fig. 1 (c) shows an indoor image, in which we have multiple shadows. This situation often arises in indoor scenes, since indoors often have multiple light sources. In such a case, each of the shadows can be considered as a camera image with a

Hasegawa, R., Sakaue, F. and Sato, J.
3D Human Body Reconstruction from Head-Mounted Omnidirectional Camera and Light Sources.
DOI: 10.5220/0011724600003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 982-989
ISBN: 978-989-758-634-7; ISSN: 2184-4321

different viewpoint, and so we have many camera images of the human body in a single image. Thus, by using shadow images, it is possible to recover the 3D human body using camera images with a large number of different viewpoints. In this paper, we improve the accuracy of the reconstruction of hidden parts in the head-mounted image by using such shadow-based multi-viewpoint images.

2 RELATED WORK

Human motion capture has traditionally been performed using a large number of fixed cameras. Another standard method is to use markers, which require humans to wear markers or special suits. The markerless motion capture system (Bregler and Malik, 1998; Joo et al., 2016) overcomes these limitations of markers. While most methods capture human motions in indoor scenes, a motion capture method with a small number of cameras in outdoor scene (Burenius et al., 2013; Elhayek et al., 2015b; Pavlakos et al., 2017) has also been proposed. However, there was a problem that it took time to set up the camera system and there were restrictions on the setup location.

Pose estimation using monocular camera images provides real-time pose estimation by using deep learning. Some of these methods estimate 3D pose directly from the image (Pavlakos et al., 2016; Tekin et al., 2016), while others combine 2D pose estimation with depth or heat maps to estimate 3D pose (Zhou et al., 2017; Pavlakos et al., 2018). However, these methods use images taken by standard perspective cameras placed to obtain good views of the entire human body, whereas our method uses an omnidirectional camera mounted on the human head. The camera images obtained from the head-mounted omnidirectional camera are highly distorted and have many self-occlusions, so the existing methods cannot be applied.

Recently, a head-mounted camera based 3D pose estimation method, Mo^2Cap^2 (Xu et al., 2019), was proposed. In this method, while the body parts visible from the camera can be estimated with high accuracy, the hidden parts of the body cannot be estimated well. Thus, we in this paper propose a method for estimating the entire body including such hidden parts by using shadows generated by light sources.

For 3D human body reconstruction, a parametric human model is often used. Skinned Multi-Person Linear (SMPL) model (Loper et al., 2015; Pavlakos et al., 2019; Anguelov et al., 2005) is one of the most popular parametric human body models, which is represented by a mesh of 6890 points and 23 joint points. They are determined by the pose parameter θ and the body shape parameter β. Many methods have been proposed for recovering the 3D human body by using the SMPL model (Kanazawa et al., 2018; Bogo et al., 2016; Kolotouros et al., 2019). Kanazawa et al. (Kanazawa et al., 2018) proposed an efficient method for estimating the SMPL parameters from an image by using deep learning. This method can recover SMPL parameters from a single viewpoint image, and the camera parameters are estimated simultaneously with the SMPL parameters.

While these methods estimate the human body shape from front-parallel camera images, our method uses an omnidirectional camera mounted on the human head. Therefore, the images used in our method are highly distorted and many parts of the human body are self-occluded. So, the existing methods cannot be applied to our task. In this paper, we solve these problems by learning the relationship between the distorted images and SMPL parameters and using virtual multi-viewpoint images obtained from shadows in a single image.

3 RECOVERY OF 3D HUMAN BODY USING SHADOWS

In this research, the object and shadows in a single omnidirectional image are combined for recovering highly accurate 3D human shapes. For this objective, we first explain the 3D human model and its parameters.

3.1 3D Human Body Model

For representing the 3D human body, we use the SMPL model, $M(\theta,\beta) \in \mathbb{R}^{3\times 6980}$, which has the pose parameter $\theta \in \mathbb{R}^{3K}$ and the shape parameter $\beta \in \mathbb{R}^{10}$, where K represents the number of joint points, and in this research $K = 23$.

Estimation of the SMPL model is performed using the floor coordinate system. The floor coordinate system consists of two axes parallel to the floor surface and one vertical axis. The origin of the floor coordinate system is at the intersection of a line perpendicular to the floor that passes through the camera's viewpoint and the floor. We represent the relationship between the SMPL model and the floor coordinates by the rotation $R \in \mathbb{R}^3$ and the translation $t \in \mathbb{R}$. Thus, the SMPL model represented based on the floor coordinates is described as follows:

$$\hat{X} = RM(\theta,\beta) + t \tag{1}$$

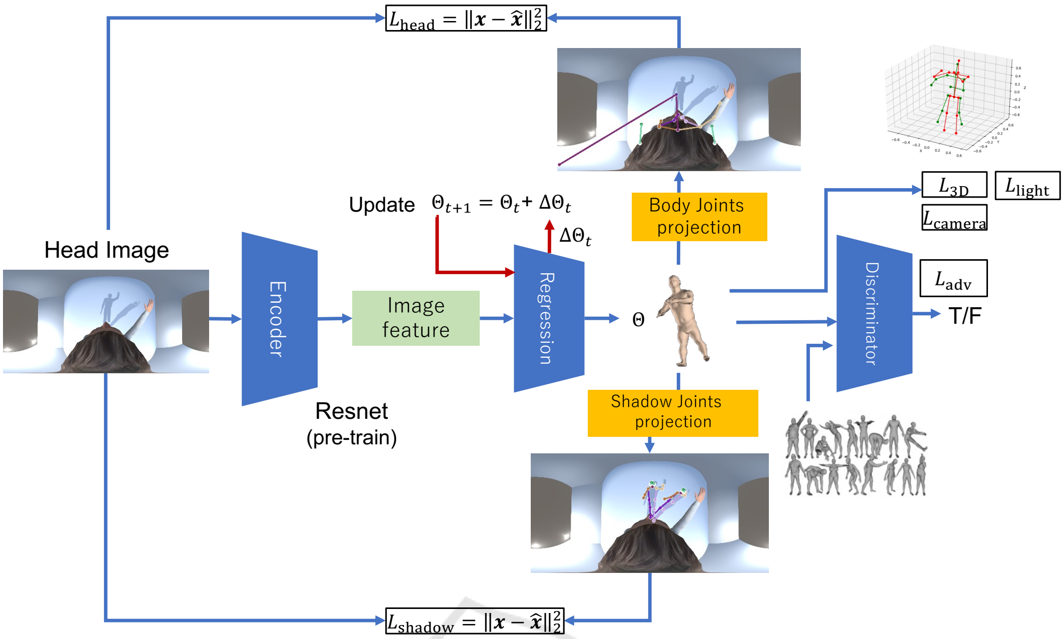

Figure 2: Network structure of Method 1.

The reason for representing the human body model in the floor coordinate system is to consider the shadow projection of the human body on the floor. We will consider this shadow projection later.

3.2 Projection to Omnidirectional Camera

The human body represented in the floor coordinate system is then converted to the camera coordinate system and projected onto the image.

In this research, the camera is fixed to the human head and its pose is equivalent to the head pose. Therefore, the camera pose can be obtained from the estimated SMPL parameters. The camera coordinate system is defined so that the direction from the neck to the top of the head is the z-axis, the direction from the right eye to the left eye is the x-axis, and the y-axis is orthogonal to these two directions. Then, the human body represented in the floor coordinate system can be converted to the camera coordinate system.

In this research, we use an omnidirectional camera. Therefore, the images are represented in equirectangular format. In equirectangular format, a 3D point $\mathbf{X}= (X,Y,Z)$ in the camera coordinate systems is projected to an image point as follows:

$$\begin{cases} \lambda = \tan^{-1}\left(\frac{Y}{X}\right) \\ \varphi = \tan^{-1}\left(\frac{Z}{\sqrt{X^2+Y^2}}\right) \end{cases} \tag{2}$$

where, λ and φ represent the longitude and latitude in the image.

Then, the human body represented by the SMPL model is projected to the omnidirectional image as follows:

$$\hat{\mathbf{x}} = \Pi(RM(\theta,\beta)+t) \tag{3}$$

where Π denotes the projection to the omnidirectional image with the coordinate transformation.

3.3 Shadow Projection

Shadow projection is performed by obtaining a shadow on the floor surface in the floor coordinate system, converting it to the camera coordinate system, and projecting it onto the omnidirectional image.

The shadow is formed at the intersection of the straight line connecting the light source and the object and the floor surface. Assuming that the light source is at infinity and its direction is $(l_x,l_y,1)$, the shadow point $\mathbf{X}' = (X',Y',Z')$ on the floor which corresponds to the 3D point $\mathbf{X} = (X,Y,Z)$ can be represented as follows:

$$\mathbf{X}' = \mathbf{L}\mathbf{X} \tag{4}$$

where, $\mathbf{L}$ denotes the following matrix:

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & -l_x \\ 0 & 1 & -l_y \\ 0 & 0 & 0 \end{pmatrix} \tag{5}$$

Thus, the shadow of the human body represented by the SMPL model is projected to the omnidirectional image as follows:

$$\hat{\mathbf{x}} = \Pi(LRM(\theta,\beta)+t) \tag{6}$$

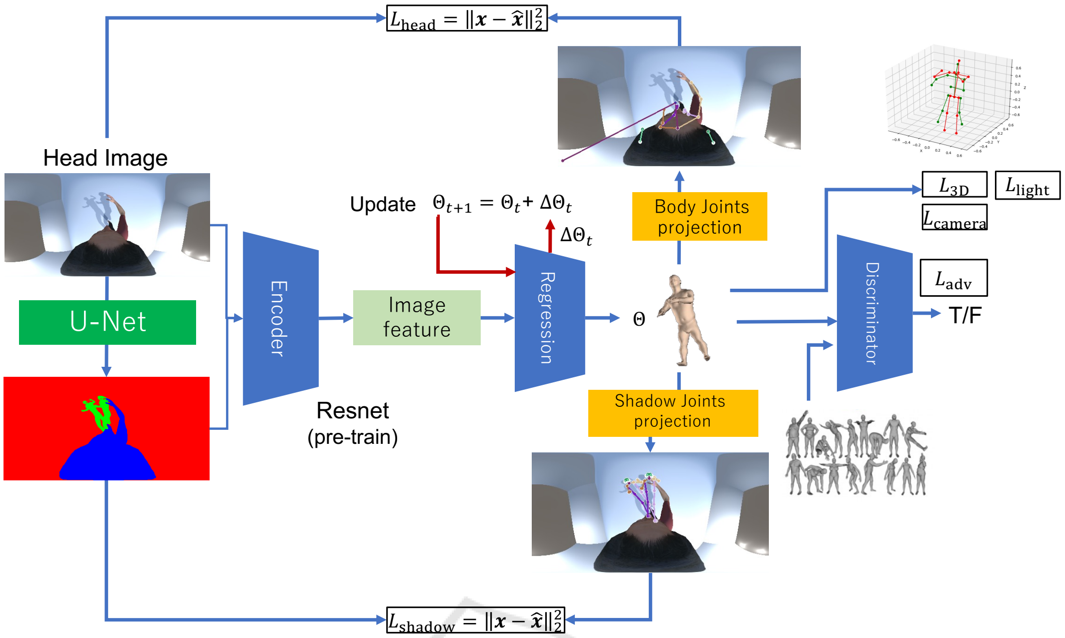

Figure 3: Network structure of Method 2.

where Π denotes the projection to the omnidirectional image with the coordinate transformation, and L denotes the shadow generation described by Eq. (4).

3.4 Network and Estimation

In this research, we estimate the SMPL parameters $\{\theta, \beta\}$, rotation R, translation t, and light source parameters l simultaneously. Thus, we estimate an $83 + 2n$ dimensional vector $\Theta = \{\theta, \beta, R, t, l\}$ by using the network, where n is the number of light sources, i.e. the number of shadows.

The network structure of the proposed method is shown in Fig 2. Our network is based on the network proposed by Kanazawa et al. (Kanazawa et al., 2018), but it is different from their network in some respects. Since our method uses shadows in the scene, we estimate not only the SMPL model but also the light sources in the scene. By using the loss obtained from the shadows, our network can estimate the SMPL model accurately, even if the head-mounted camera image has heavy self-occlusions.

The network takes a single head-mounted camera image as input and outputs the parameter $\Theta = \{\theta, \beta, R, t, l\}$. Estimation is performed in two steps. First, the image is input to Encoder, which extracts image features using a pre-trained Resnet (He et al., 2015). Next, the output of the Encoder is fed to the residual network to estimate Θ. The residual network is used here because it is difficult to estimate the rotation parameters directly in a single estimation.

The estimation is evaluated by L_2 loss of the estimated 3D joint point L_{3D}, camera parameters L_{camera}, and light parameters L_{light}. The camera parameter loss L_{camera} is used in our network since the accuracy of camera pose estimation greatly affects other parameters and is particularly important. The reprojection errors for body joints in the image L_{head} and shadow joints in the image L_{shadow} are also evaluated. The reprojection error of shadow joints L_{shadow} is computed for the number of shadows.

In addition, the proposed network has the discriminator and performs adversarial training (Goodfellow et al., 2014). This is to prevent the estimated SMPL from deviating too much from the real human body structure by discriminating between the estimated SMPL parameters and the SMPL parameter datasets for various real poses.

Thus, the training loss of our network is described as follows:

$$L = w_1 L_{3D} + w_2 L_{light} + w_3 L_{camera} + w_4 L_{head} + w_5 L_{shadow} + w_6 L_{adv} \quad (7)$$

We train the network by minimizing the loss. In our experiments, we chose $w_1 = 1.0$, $w_2 = 6.0$, $w_3 = 7.5$, $w_4 = 750$, $w_5 = 1.0$, and $w_6 = 2.0$.

3.5 Estimation with Shadow Extraction

Up to now, a camera image was input to Encoder without any processing. However, it may be diffi-

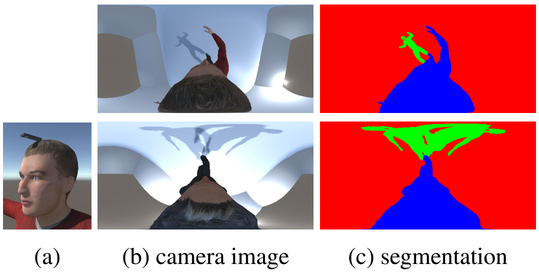

Figure 4: Dataset images. (a) shows a head-mounted omnidirectional camera, (b) shows camera images, and (c) shows segmented images.

cult for the network to identify shadows in the image. Therefore, we also consider an alternate method in which shadow regions are extracted in advance and input to Encoder with the original camera image. We call the method without shadow extraction Method 1 and call the method with shadow extraction Method 2 respectively.

The network structure of Method 2 is shown in Fig. 3. Segmentation images are generated from the head-mounted camera images and input to different Encoders to extract image features. These are then combined and input to a residual network for parameter estimation. The loss function for training is the same as in Method 1.

Shadow regions are extracted by generating segmentation images. Although instance segmentation can distinguish between objects and can handle multiple shadows separately, it has a problem in the representation of overlapping shadows. Therefore, we use semantic segmentation in this research. Segmentation is performed by U-Net (Ronneberger et al., 2015). The image is segmented into three classes: human body, shadow, and others. Shadows are represented as a single class and are not separated from each other.

4 DATASETS

In this research, two types of datasets are used for training: a dataset of head-mounted camera images and an SMPL dataset. The Human3.6M dataset (Ionescu et al., 2014) is annotated with 3D joint points. However, the camera used in our method is an omnidirectional camera and the image format is different. The MonoEye dataset (Hwang et al., 2020) is a dataset taken with a fisheye lens, but the camera is positioned at the thorax, which is different from the viewpoint of the head-mounted camera used in our method. Therefore, we need to create a new dataset for training our network. In this research, annotation of shadow joint points is required for the head-mounted camera image, but it is difficult to create such a dataset by using a real head-mounted omnidirectional camera. Therefore, we created a dataset with synthetic images using Unity. Human models created by Autodesk Character Generator (Autodesk, 2022) and 210 animations downloaded from Mixamo (Mixamo, 2022) were used for generating synthetic images. The position of the light source and its intensity was selected randomly,

Table 1: Number of shadows and accuracy of human body recovery.

	MPJPE (mm)
0 shadow	259.39
1 shadow	200.24
2 shadows	166.46

Table 2: Accuracy of proposed method with/without shadow segmentation (MPJPE (mm)).

	Method 1	Method 2
1 shadow	200.24	182.58
2 shadows	166.46	143.97

The dataset consists of a single head-mounted camera image with 3D joint points, 2D joint points, 2D shadow joint points, light source parameters, and true-value annotations for camera pose. The ground truth of the segmentation image was generated by creating images with and without shadows and taking the difference between them. A part of the dataset is shown in Fig. 4. Fig. 4 (a) shows a head-mounted omnidirectional camera used for creating the dataset, Fig. 4 (b) shows camera images, and Fig. 4 (c) shows segmented images.

We also used Mosh dataset (Loper et al., 2014) as the ground truth SMPL dataset for adversarial training. The generator is trained so that the discriminator cannot distinguish between the generated SMPL and the ground truth SMPL.

5 EXPERIMENT

Experiments were conducted using the synthetic image dataset described in section 4. Three types of light sources (0, 1, and 2) were used for training and testing, where 0 light means we do not use shadow information. The training data consists of 3150 images without shadows (15 models, 210 poses), 15750 images with one shadow (15 models, 210 poses, 5 light source positions), and 15750 images with two shadows (15 models, 210 poses, 5 light source positions). The test data consists of 100 images each.

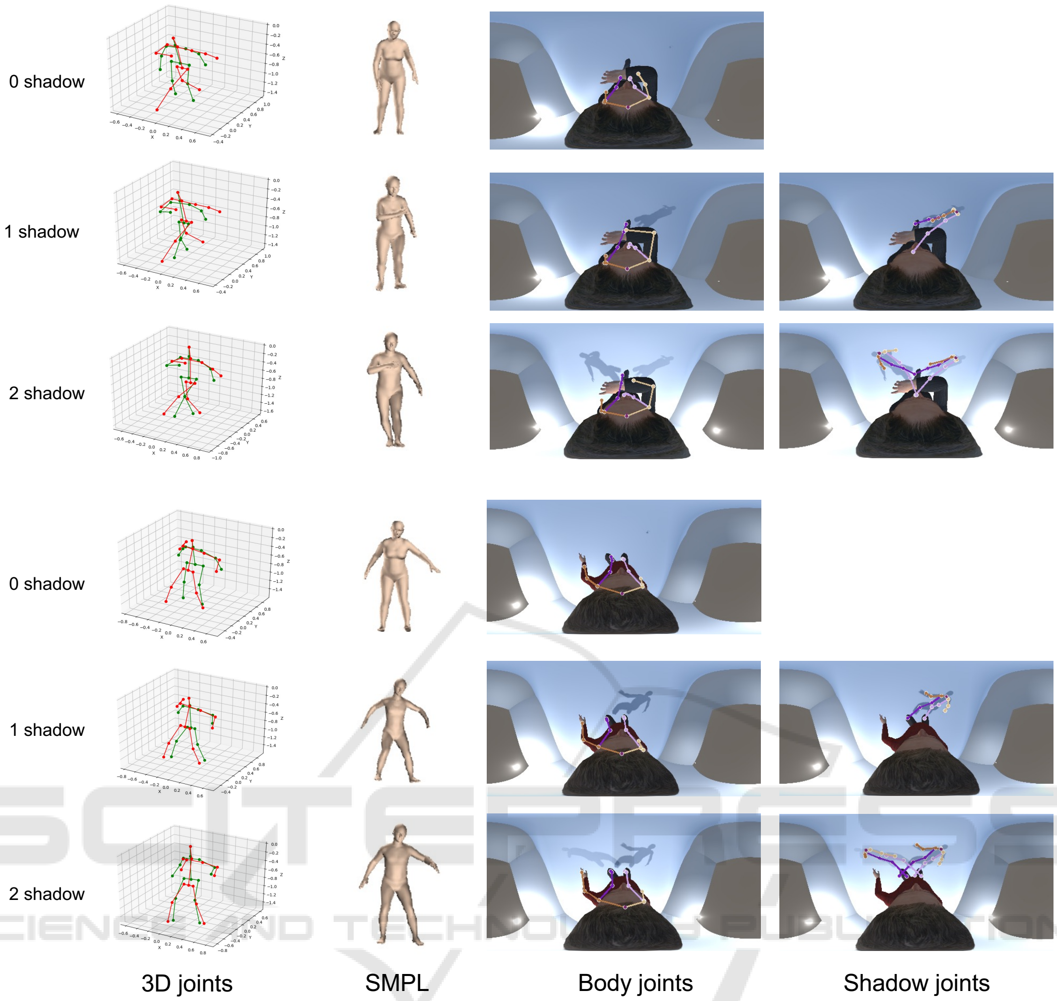

Figure 5: Result of the proposed method in the case of 0 shadow, 1 shadow, and 2 shadows. Red lines and green lines in 3D joints represent true poses and estimated poses respectively. Body joints and Shadow joints show reprojection of body and shadow joints on the input omnidirectional image.

We first show the results of Method 1. Fig. 5 shows the 3D joints and SMPL human body estimated from our method in the case of 0 shadow, 1 shadow, and 2 shadows respectively. We also show the reprojection of the estimated 3D joints of the body and shadows. As shown in this figure, the accuracy of the estimated results improves as the number of shadows increases. In particular, the accuracy of estimation in *Z* axis, which cannot be determined from the human body image alone, improves as the number of shadows, i.e. the number of viewpoints, increases. We can also find that the accuracy of invisible feet also improves.

A quantitative evaluation of the test results was also conducted. Table 1 shows Mean Per Joints Position Error (MPJPE) of the estimated human body in the case of 0 shadow, 1 shadow, and 2 shadows. We find that the accuracy improves as the number of shadows increases. This confirms the effectiveness of using shadows in 3D human body reconstruction from head-mounted camera images.

We next evaluate Method 2, which combines Method 1 with shadow extraction. The 3D human body estimated by using Method 2 is shown in Fig. 6. In both cases of one and two shadows, we find that Method 2 using shadow segmentation improves the accuracy of the reprojection of shadows and the overall accuracy of the recovered body compared to Method 1. In particular, in the case of the two shadows, the right arm of the image is successfully recovered since the shadow information, which is not fully used in Method 1, is captured and used for the estimation in Method 2.

Method 2 was also quantitatively evaluated by MPJPE . The results are shown in Table 2. We find that MPJPE of Method 2 is lower than that of Method

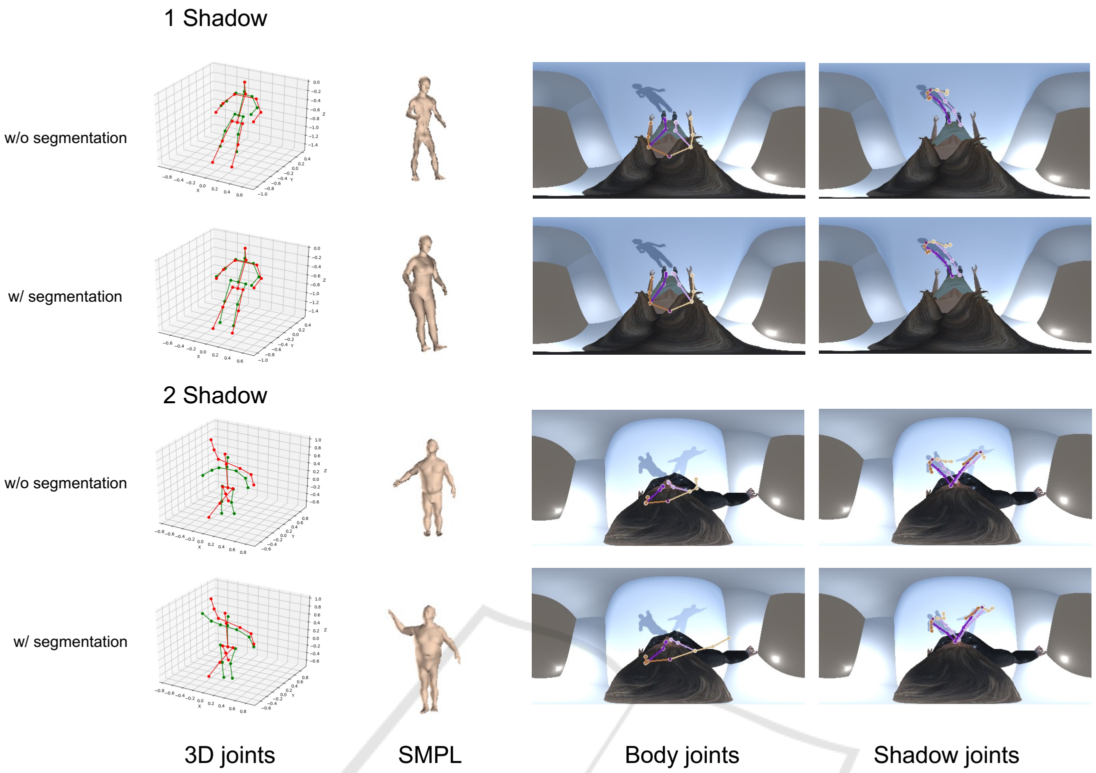

Figure 6: Results of the proposed method with and without shadow segmentation in the case of 1 shadow and 2 shadows.

1 in both cases where the number of shadows is 1 and 2. These results confirm the effectiveness of using shadow extraction in the proposed method.

6 CONCLUSIONS

In this paper, we proposed a method for recovering the 3D shape of the human body by using shadows in a single head-mounted camera image. We also proposed a method for extracting shadow regions by using semantic segmentation and combining the results to further improve the accuracy of the proposed method.

We created a synthetic image dataset using Unity and conducted experiments using the dataset. We showed that by using the shadows in the image, the estimation of the 3D human body can use multiple views, and as a result, we can drastically improve the accuracy of human body estimation, even if the head-mounted images suffer from self-occlusion.

In our future work, we will conduct the evaluation of our method using images obtained from a real head-mounted omnidirectional camera under various environments.

REFERENCES

Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., and Davis, J. (2005). Scape: shape completion and animation of people. *ACM Trans. Graph.*, 24:408–416.

Autodesk (2022). Autodesk character generator. *https://charactergenerator.autodesk.com/*.

Bogo, F., Kanazawa, A., Lassner, C., Gehler, 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.

Bregler, C. and Malik, J. (1998). Tracking people with twists and exponential maps. In *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231)*, pages 8–15.

Burenius, M., Sullivan, J., and Carlsson, S. (2013). 3d pictorial structures for multiple view articulated pose estimation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3618–3625.

Elhayek, A., Aguiar, E., Jain, A., Tompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., and Theobalt, C. (2015a). Efficient convnet-based markerless motion capture in general scenes with a low number of cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Elhayek, A., de Aguiar, E., Jain, A., Tompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., and Theobalt, C. (2015b). Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3810–3818.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Hwang, D.-H., Aso, K., Yuan, Y., Kitani, K., and Koike, H. (2020). Monoeye: Multimodal human motion capture system using a single ultra-wide fisheye camera. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, pages 98–111.

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:1325–1339.

Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T., Nabbe, B. C., Matthews, I. A., Kanade, T., Nobuhara, S., and Sheikh, Y. (2016). Panoptic studio: A massively multiview system for social interaction capture. *CoRR*, abs/1612.03153.

Kanazawa, A., Black, M. J., Jacobs, D. W., and Malik, J. (2018). End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*.

Kolotouros, N., Pavlakos, G., Black, M. J., and Daniilidis, K. (2019). Learning to reconstruct 3d human pose and shape via model-fitting in the loop. *CoRR*, abs/1909.12828.

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.

Loper, M. M., Mahmood, N., and Black, M. J. (2014). MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13.

Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., and Theobalt, C. (2017a). Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE.

Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H., Xu, W., Casas, D., and Theobalt, C. (2017b). Vnect: Real-time 3d human pose estimation with a single RGB camera. *CoRR*, abs/1705.01583.

Mixamo (2022). Get animated. *https://www.mixamo.com/*.

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)*, pages 10975–10985.

Pavlakos, G., Zhou, X., and Daniilidis, K. (2018). Ordinal depth supervision for 3d human pose estimation. *CoRR*, abs/1805.04095.

Pavlakos, G., Zhou, X., Derpanis, K. G., and Daniilidis, K. (2016). Coarse-to-fine volumetric prediction for single-image 3d human pose. *CoRR*, abs/1611.07828.

Pavlakos, G., Zhou, X., Derpanis, K. G., and Daniilidis, K. (2017). Harvesting multiple views for marker-less 3d human pose annotations. *CoRR*, abs/1704.04793.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Shiratori, T., Park, H. S., Sigal, L., Sheikh, Y., and Hodgins, J. K. (2011). Motion capture from body-mounted cameras. *ACM Trans. Graph.*, 30(4).

Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., and Fua, P. (2016). Structured prediction of 3d human pose with deep neural networks. *CoRR*, abs/1605.05180.

von Marcard, T., Rosenhahn, B., Black, M., and Pons-Moll, G. (2017). Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. *Computer Graphics Forum 36(2), Proceedings of the 38th Annual Conference of the European Association for Computer Graphics (Eurographics)*.

Xu, W., Chatterjee, A., Zollhoefer, M., Rhodin, H., Fua, P., Seidel, H.-P., and Theobalt, C. (2019). Mo^{2}Cap^2 : Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1.

Zhou, X., Huang, Q., Sun, X., Xue, X., and Wei, Y. (2017). Weakly-supervised transfer for 3d human pose estimation in the wild. *CoRR*, abs/1704.02447.