SILHOUETTE-BASED SYNTHETIC DATA GENERATION FOR 3D HUMAN POSE ESTIMATION WITH A SINGLE WRIST-MOUNTED 360° CAMERA

Ryosuke Hori[⋆] *Ryo Hachiuma*[⋆] *Hideo Saito*[⋆] *Mariko Isogawa*[†] *Dan Mikami*[†]

[⋆] Department of Information and Computer Science, Keio University, Japan
[†] NTT Media Intelligence Laboratories, Japan

ABSTRACT

In this paper, we propose a framework for 3D human pose estimation with a single 360° camera mounted on the user's wrist. Perceiving a 3D human pose with such a simple setting has remarkable potential for various applications (e.g., daily-living activity monitoring, motion analysis for sports enhancement). However, no existing work has tackled this task due to the difficulty of estimating a human pose from a single camera image in which only a part of the human body is captured and the lack of training data. Therefore, we propose an effective method for translating wrist-mounted 360° camera images into 3D human poses. We also propose silhouette-based synthetic data generation dedicated to this task, which enables us to bridge the domain gap between real-world data and synthetic data. We achieved higher estimation accuracy quantitatively and qualitatively compared with other baseline methods.

***Index Terms*—** 3D human pose estimation, 360° camera, data synthesis, silhouette, domain adaptation

1. INTRODUCTION

Vision-based 3D human pose estimation has been widely researched in recent years. Especially, 3D human pose estimation with wearable cameras is key to many important applications, such as lifelogging in terms of medical assistance, monitoring for life support, virtual reality, and sports activity analysis. Several methods that use wearable cameras mounted on the head or chest have been proposed recently [1–6]. However, thus far, there is no method with a more practical camera setting, i.e., a single wrist-mounted camera which could be introduced in smartwatches in the future.

Therefore, we propose a framework for estimating 3D poses from images taken with a single wrist-mounted camera. This task is quite challenging as some human body parts are hidden from the camera's line of sight. As shown in Fig. 1, we make use of a single 360° camera (GoPro Max) and a convolutional neural network-based framework following Yuan and Kitani's work [5] to estimate a 3D human pose with only limited visual information. The difficulty is how to prepare the training data; there is no existing dataset for 3D human pose estimation with wrist-mounted cameras. Existing works with body-mounted cameras that faced this issue tackled it by using synthetic data [1–4]. However, as widely known in these previous works, there is a domain gap issue between real-world data and synthetic data.

Fig. 1. Our 3D human pose estimation with a single wrist-mounted 360° camera.

To overcome these issues, we also describe a simplified method for generating training data. We generate silhouette-based equirectangular image sequences given only existing motion capture (MoCap) data to train the network. It reduces the data generation cost, and because the data are fully silhouette-based, it reduces the problem of domain gaps between synthetic data and real-world data.

To summarize, our contributions are as follows:
(1) We are the first to propose a 3D human pose estimation framework given a single wrist-mounted camera, which contributes to many important applications.
(2) To reduce data generation cost, and to bridge domain gaps between synthetic and real-world data, we describe a method for silhouette-based training data synthesis. This data generation method has the potential to be used for other camera settings.
(3) We provide extensive experimentation and show that our method outperforms other baseline methods.

2. RELATED WORK

Human pose estimation has long been studied in the computer vision community [7]. In particular, 3D pose estimation using a monocular camera, which is the most widely used sensor in the world, has been actively examined because of its usefulness in various situations, such as video surveillance, human–computer interaction (HCI), and self-driving [8].

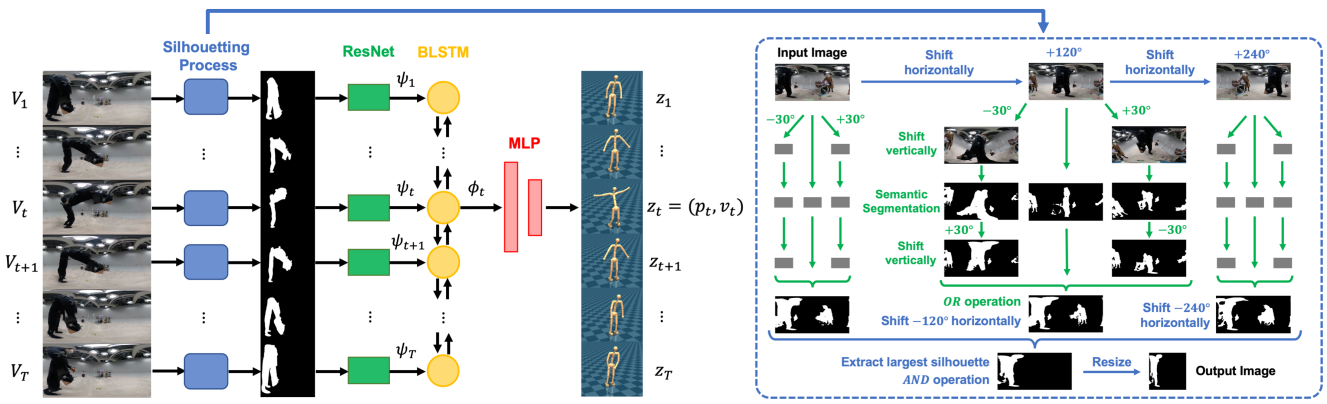

Fig. 2. Overview of our 3D pose estimation method using a single wrist-mounted 360° camera. The training process requires only synthetic silhouette images as shown in Fig. 3. For inference, the equirectangular images taken in the real environment are converted into silhouette images by the silhouetting process (right blue box) and then inputted in the network (left).

Three-dimensional human pose estimation with body-mounted cameras has also been widely investigated in recent years. Previous researchers used multiple body-mounted cameras for whole-body pose estimation [9] and used RGB-D cameras for upper-body (i.e., hands, arms, torso) motion estimation [10]. In recent years, methods for whole-body 3D pose estimation with more practical settings have been proposed, such as using a wearable camera with a wide viewing angle to capture more body parts [1–4, 11]. Moreover, several studies have achieved 3D pose estimation under the severe condition in which the human body is completely hidden from the camera's line of sight [5, 6, 12]. Thus far, these existing methods mount the camera on the user's head or chest. This paper explores the potential for another camera setting for user-mounted cameras, i.e., a single wrist-mounted camera that is considered more practical because it could be introduced in smartwatches in the future.

The majority of recent wearable camera-based 3D human pose estimation methods used fisheye cameras, expecting more body parts to be captured [1–4]. As the conventional datasets for 3D human pose estimation cannot be directly applied to these fisheye camera-based methods, these methods use synthetic datasets dedicated to each method to train their network. However, these data synthesis processes have a huge cost to bridge domain gaps between real and synthetic data. In contrast to the approach of synthesizing more realistic, high-dimensional data, Xu *et al*. proposed a low-dimensional synthetic data generation approach for a pedestrian trajectory estimation to bridge the domain gaps [13]. Inspired by this method, we use silhouette synthetic data aiming at reducing the data generation cost and bridging the domain gap.

3. PROPOSED METHOD

We propose a method for 3D pose estimation using images from a wrist-mounted 360° camera, trained only with synthetic silhouette data generated at a lower cost than existing conventional methods. During inference, we apply a silhouetting process to the actually captured images to bridge the domain gap between the synthetic data and real-world data.

3.1. Human Pose Estimation Network

The 3D human pose estimation network (Fig. 2) is inspired by the method proposed by Yuan and Kitani [5]. The network $\mathcal{F}$ takes the input of the equirectangular video frames $V_{1:T}$ in which the person is silhouetted and predicts the humanoid state $z_{1:T}$ at each frame. The humanoid state z_t consists of the pose p_t (position and orientation of the root, and joint angles) and velocity v_t (linear and angular velocities of the root, and joint velocities). The model encodes the silhouette image to ResNet-18 [14] to extract the feature vector $\psi_{1:T} \in \mathbb{R}^{128}$ and feeds it to bidirectional long-short term memory (BiLSTM) to generate the visual context $\phi_{1:T} \in \mathbb{R}^{128}$ for each frame. We then feed it to the multilayer perceptrons (MLPs) and predict the humanoid state $z_{1:T}$. The mean squared error (MSE) is used as the loss function: $L(\zeta) = \frac{1}{T}\sum_{t=1}^{T} \|\mathcal{F}(V_{1:T})_t - \hat{z}_t\|^2$, where ζ is the parameter of this network $\mathcal{F}$, and $\hat{z}_t$ is the ground-truth humanoid state. The optimal $\mathcal{F}^*$ can be obtained by an SGD-based method.

3.2. Training Data Synthesis

To synthetically train the network, we generate pairs of input silhouetted equirectangular videos and the corresponding 3D pose of the camera wearer. In a virtual environment, such as Unity, the 360° camera is fixed at the virtual avatar's wrist position. By making the virtual avatar move with the MoCap data, the corresponding input equirectangular image and output 3D human pose sets can be generated. Fig. 3 depicts how the equirectangular image is generated synthetically.

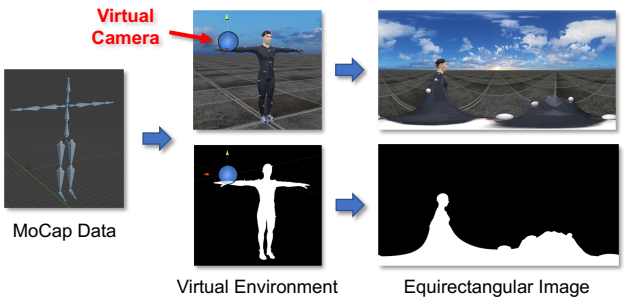

Fig. 3. Overview of training data synthesis. We generate binary images as shown in the lower right.

3.3. Silhouetting process for inference

For inference, we apply semantic segmentation to the equirectangular images captured with a wrist-mounted 360° camera and generate a human silhouetted binary image by extracting the parts labeled as human (Fig. 2). We employ HRNet [15] as the semantic segmentation network which is trained with the ADE20K dataset [16]. As the equirectangular images are heavily distorted at the top and bottom of the image, unlike the images from the general perspective projection model in the ADE20K dataset, we apply semantic segmentation not only to the original equirectangular images but also to the equirectangular images shifted vertically. Then we merge the results of the semantic segmentation to obtain the network input.

Specifically, we first generate three equirectangular images by shifting them horizontally (yaw axis) at 120° intervals. Second, we generate two images by shifting each one by ±30° vertically (pitch axis). Third, we apply semantic segmentation to each image and shift them vertically back to the equirectangular images of the original vertical angle. Fourth, in each of the three horizontally shifted images, we combine the three generated silhouette images via the *OR* operation and shift them horizontally back to the same position as the input image. Fifth, we extract the largest silhouette of each of the three images and merge it via the *AND* operation. The image is resized to get the image for input in the network.

4. EXPERIMENT

4.1. Dataset

- **MoCap Training and Test Data**: We used OptiTrack to capture the motion data to construct the dataset. Two subjects wore the 360° camera on their wrist and were asked to perform a variety of actions, including walking, jumping, crouching, and raising their hands. Each take lasted about 5 min, and each subject performed two takes. We used three of the four takes as the training data and the fourth take as the test data.
- **In-the-Wild Data**: We also collected in-the-wild data to verify the effectiveness of our method in a real-world environment. As in the previous MoCap training and test data collection, the subject wore the 360° camera on their wrist and was asked to perform a variety of actions. This dataset consisted of 11 videos each lasting about 5 sec. As it is hard to obtain ground-truth 3D poses in a real-world environment, following Yuan and Kitani [5], we captured side-view poses of the subject, which were used for quantitative evaluation based on 2D keypoints.

The dataset is available at this link.

4.2. Network Training

The weights of ResNet-18 [14] were pretrained with ImageNet [17], and the Adam [18] optimizer was employed at the learning rate of $1e-4$. The input equirectangular image was resized to 224×224. When training this network, for each time step we sampled data fragments in turn for 120 frames (4 sec) and padded 10 frames of visual features ψ_t on both sides to reduce the estimation error on the boundary frames when computing ϕ_t. We used Unity as the virtual environment to generate the training data and MuJoCo [19] to visualize estimated human poses that consisted of 52 degrees of freedom (DoFs) and 19 rigid bodies.

4.3. Evaluation Metric

Following Isogawa *et al*. [20], we used the following metrics to evaluate the accuracy of 3D human pose estimation. For the estimated and ground-truth keypoints, we set the hip keypoint as the origin and scaled the coordinate to make the height between the shoulder and hip equal to 0.5 [m]. The errors in Table 1 were measured in meters.

- **Mean Per-Joint Position Error (MPJPE)**: For evaluation on the MoCap test data, we employed MPJPE that measures the Euclidean distance between the estimated pose and the ground-truth pose. This metric is defined as $\frac{1}{TJ}\sum_{t=1}^{T}\sum_{j=1}^{J}\|(x_t^j - x_t^{root}) - (\hat{x}_t^j - \hat{x}_t^{root})\|_2$, where x_t^j is the j^{th} joint position of the estimated pose, and $\hat{x}_t^j$ is the ground truth. x_t^{root} and $\hat{x}_t^{root}$ represent the root joint position of the estimated and ground-truth poses, respectively.
- **2D Keypoint Error ($\mathbf{E}_{\text{key}}$)**: For evaluation on the in-the-wild data, we employed the pose-based metric $\mathbf{E}_{\text{key}}$ calculated as $\frac{1}{TJ}\sum_{t=1}^{T}\sum_{j=1}^{J}\|(y_t^j - \hat{y}_t^j)\|_2$, where y_t^j is the j^{th} 2D keypoint of the estimated pose obtained by projecting the 3D joints to an image plane with a side-view camera, and $\hat{y}_t^j$ is the ground truth extracted with OpenPose [21].

4.4. Baseline Methods

We compared our method against the following baseline methods:

Method	MoCap Test Data (MPJPE)					In-the-Wild Data
	Walk	Jump	Crouch	Raise hand	All Frames	$\mathbf{E}_{\text{key}}$
RGB	0.346	0.311	0.284	0.407	0.339 ± 0.068	0.330 ± 0.074
Optical Flow	0.118	0.192	0.145	0.128	0.132 ± 0.070	0.352 ± 0.091
SS Silhouette	0.227	0.256	0.229	0.173	0.227 ± 0.057	0.275 ± 0.073
Ours	**0.106**	**0.147**	**0.138**	**0.106**	$\mathbf{0.115 \pm 0.053}$	$\mathbf{0.198 \pm 0.083}$

Table 1. Quantitative results of pose estimation accuracy on the MoCap test data and the in-the-wild data.

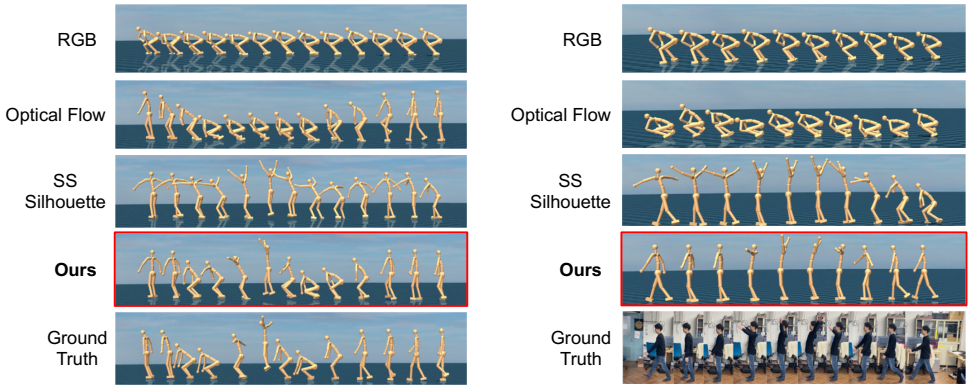

Fig. 4. Qualitative results for each method on the MoCap test data (left) and the in-the-wild data (right).

- **RGB**: A method that trains the network on synthetic RGB data as shown in the upper part of Fig. 3 and tests it on real RGB data taken with a 360° camera.

- **Optical Flow**: The PoseReg method proposed by Yuan and Kitani [5] that uses optical flows of the input RGB data obtained by PWC-Net [22] as the network input.

- **Semantic Segmentation (SS) Silhouette**: A method that trains the network on synthetic silhouette data as shown in the lower part of Fig. 3 and tests it on silhouette images created by applying semantic segmentation to real equirectangular images and extracting the parts labeled as human.

4.5. Results and Discussion

The quantitative results of the pose estimation on the MoCap test data and the in-the-wild data are shown in Table 1, and the qualitative results are shown in Fig. 4. The results show that our method outperforms the baseline methods.

The method using RGB data failed to estimate poses through the sequences. In addition, although the method using optical flow data estimated simple motions such as walking and raising hands relatively well on the MoCap test data, it failed to estimate poses on the in-the-wild data because the optical flow of the objects around the subject was estimated. These methods seem to have failed because of a large domain gap between the synthetic training data and the test data collected in the real-world environment.

The method using SS silhouette data also had lower estimation accuracy than ours. Due to the distortion that occurred in the equirectangular images, the human region segmentation failed, resulting in very noisy data with the presence of human regions other than the user who wore the camera.

In contrast, our method produces 3D human poses closer to the ground truth than any other baseline. The results indicate that our method works effectively, thanks to the synthetic silhouette training data and the silhouetting process for the inference, which bridges the domain gap between synthetic data and real data.

5. CONCLUSION

We presented a framework for estimating 3D human poses with a single wrist-mounted 360° camera. Our pose estimation network is trained only on synthetic silhouette image data generated in the virtual environment. For inference, our method uses binary silhouette images generated via the silhouetting process that takes actually captured images as input. Our synthetically trained method could reduce the data generation cost and bridges the domain gaps between synthetic and real data, which has been an issue in previous researches. The experimental results showed that our method outperforms other baseline methods qualitatively and quantitatively. We believe that as our synthetically trained network for 3D human pose estimation with a single wrist-mounted camera can easily be extended to other camera settings, our method contributes to the further development of this research field.

6. REFERENCES

[1] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, P. Fua, H.-P. Seidel, and C. Theobalt, "Mo2Cap2: Real-time Mobile 3D Motion Capture with a Cap-mounted Fisheye Camera," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 5, pp. 2093–2101, 2019.

[2] D. Tome, P. Peluse, L. Agapito, and H. Badino, "xR-EgoPose: Egocentric 3D Human Pose From an HMD Camera," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7727–7737.

[3] D. Tome, T. Alldieck, P. Peluse, G. Pons-Moll, L. Agapito, H. Badino, and F. De la Torre, "SelfPose: 3D Egocentric Pose Estimation from a Headset Mounted Camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, p. 1–1, 2020.

[4] D.-H. Hwang, K. Aso, Y. Yuan, K. Kitani, and H. Koike, "MonoEye: Multimodal Human Motion Capture System Using A Single Ultra-Wide Fisheye Camera," in *Annual ACM Symposium on User Interface Software and Technology*, 2020, pp. 98–111.

[5] Y. Yuan and K. Kitani, "Ego-Pose Estimation and Forecasting as Real-Time PD Control," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10081–10091.

[6] E. Ng, D. Xiang, H. Joo, and K. Grauman, "You2Me: Inferring Body Pose in Egocentric Video Via First and Second Person Interactions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9887–9897.

[7] Z. Liu, J. Zhu, J. Bu, and C. Chen, "A Survey of Human Pose Estimation: The Body Parts Parsing Based Methods," *J. Vis. Commun. Image Represent.*, vol. 32, pp. 10–19, 2015.

[8] Y. Chen, Y. Tian, and M. He, "Monocular Human Pose Estimation: A Survey of Deep Learning-based Methods," *Comput. Vis. Image Underst.*, vol. 192, pp. 102897, 2020.

[9] T. Shiratori, H. S. Park, L. Sigal, Y. Sheikh, and J. K. Hodgins, "Motion Capture from Body-Mounted Cameras," *ACM Trans. Graph.*, vol. 30, no. 4, 2011.

[10] G. Rogez, J. S. Supančič, and D. Ramanan, "First-person Pose Recognition Using Egocentric Workspaces," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4325–4333.

[11] H. Rhodin, C. Richardt, D. Casas, E. Insafutdinov, M. Shafiei, H.-P. Seidel, B. Schiele, and C. Theobalt, "EgoCap: Egocentric Marker-Less Motion Capture with Two Fisheye Cameras," *ACM Trans. Graph.*, vol. 35, no. 6, 2016.

[12] H. Jiang and K. Grauman, "Seeing Invisible Poses: Estimating 3D Body Pose from Egocentric Video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3501–3509.

[13] Y. Xu, V. Roy, and K. Kitani, "Estimating 3D Camera Pose from 2D Pedestrian Trajectories," in *IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 2568–2577.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[15] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5686–5696.

[16] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic Understanding of Scenes Through the ADE20K Dataset," *Int. J. Comput. Vis.*, vol. 127, pp. 302–321, 2018.

[17] J. Deng, W. Dong, R. Socher, L.-J Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-scale Hierarchical Image Database," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[18] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *International Conference on Learning Representations*, 2015.

[19] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A Physics Engine for Model-based Control," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 5026–5033.

[20] M. Isogawa, Y. Yuan, M. O'Toole, and K. Kitani, "Optical Non-Line-of-Sight Physics-Based 3D Human Pose Estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7011–7020.

[21] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[22] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, p. 8934–8943.