Improving Accuracy of Position Information in Egocentric Pose Estimation Methods

Sho Nakashima
Gradurate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan
sharon2444@ruri.waseda.jp

Hiroshi Watanabe
Gradurate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—In Virtual Reality (VR) with Head Mounted Displays (HMDs), detecting how the wearer moves is very important. If these movements are reflected in the VR with the HMD, the reality of the user's experience will be enhanced. The HMD detects the wearer's movements with sensors mounted on it and it's controller. However, this alone has limitations in reflecting the wearer's movements. Egocentric pose estimation can be used without hardware modification and is expected to improve the accuracy of wearer motion detection. Even today, there are ways to improve the accuracy of detection of the wearer's movements, but all of them require additional devices or a large space, which also detracts from the ease of use. Therefore, the ability to use the system without hardware changes is very important in that it does not compromise ease of use. However, there is no VR that uses egocentric pose estimation, indicating that the accuracy of existing methods is not suitable for use in VR. Therefore, we propose a method to improve the accuracy of location information, by modifying the reward function of the existing method. Evaluation experiments show that the proposed method improves both quantitatively and qualitatively.**

Keywords—Pose Estimation, Egocentric Pose Estimation, VR

I. Introduction

In recent years, VR has been used in various fields such as medicine, tourism, sports, and games. Among VR, HMD have become especially popular. The reason for the use of VR with HMD is that it is easy to have a realistic experience. It is available and very easy to use, as long as it and its controller are available. The HMDs also create a realistic experience through the movement of the images in accordance with the user's head, stereoscopic viewing by showing separate images to the left and right eyes, and reflection of the user's own movements using sensors. Of course, HMDs have their disadvantages. For example, there are steaminess and discomfort caused by wearing the HMD, VR sickness, and reality that is not as realistic as reality. These shortcomings are caused by the discrepancy from reality, and more realistic VR is required. One of the reasons for the discrepancy between VR and reality is the limitation of body tracking by sensors on HMDs and controllers, which can detect head and hand movements but cannot provide information on other parts of the body. Currently, better body tracking is possible with additional devices, but these additional devices are expensive or large, and they detract from the ease of use that is an advantage of HMDs.

In order to reflect the wearer's movements more accurately without sacrificing the ease of use of HMD, we study the use of posture estimation using egocentric pose estimation. Since there is no VR that uses such a posture estimation method, existing methods are not accurate enough.

Therefore, we propose a method to improve the accuracy of position information in particular by modifying the reward function of the existing method for the use of posture estimation using Egocentric pose estimation in VR.

II. Related Work

One of the previous method on egocentric estimation is provided by Yuan et al. [1]. Ego-Pose Estimation and Forecasting as Real-Time PD control is egocentric pose estimation by Ye Yuan et al. It is possible to estimate and predict human pose from egocentric video using proportional derivative control-based policies learned via reinforcement learning. The following is an overview of the conventional methods.

- The input, egocentric video, is passed through PWC-Net [2] , ResNet-18 [3] , and LSTM in that order to obtain the state from the obtained context and humanoid pose.
- The state of the next frame is generated by executing actions determined by the policy.
- A reward determined by the reward function is given through the physical simulation environment.
- By repeating the above, the humanoid control policy is trained.

The following four reward functions are used in existing methods.

- Pose reward r_p: Function to evaluate the difference between the generated pose p_t and the ground-truth pose $\hat{p}_t$. It is represented by the following equation.

$$r_{\mathrm{p}} = \exp\left[-2\left(\sum_j \left\| q_t^j \ominus \hat{q}_t^j \right\|^2\right)\right] \tag{1}$$

where q^j denotes the quaternion in the local direction of joint j.

- End-effector reward r_e : Function to evaluate the difference between the generated end-effector vector e_t and the ground-truth end-effector vector $\hat{e}_t$. It is represented by the following equation.

$$r_{\mathrm{e}} = \exp\left[-20\left(\sum_e \left\| e_t - \hat{e}_t \right\|^2\right)\right] \tag{2}$$

where the end-effector vector e_t represents the vector from the root to each end-effector (foot, hand, head).

- Root pose reward r_{rp}: Function that prompts the root joints (Hips) of the humanoid to have a ground-truth height $\hat{h}_t$ and a quaternion $\hat{q}_t^{\mathrm{r}}$ of joint r. It is represented by the following equation.

$$r_{\mathrm{rp}} = \exp\left[-300\left(\left(h_t - \hat{h}_t\right)^2 + \|q_t^{\mathrm{r}} \ominus \hat{q}_t^{\mathrm{r}}\|^2\right)\right] \quad (3)$$

- Root velocity reward r_{rv}: Function to evaluates the linear velocity l_t and angular velocity ω_t of the root and the deviation from the linear velocity $\hat{l}_t$ and angular velocity $\hat{\omega}_t$ of the ground truth. It is represented by the following equation.

$$r_{\mathrm{rv}} = \exp\left[-\left\|l_t - \hat{l}_t\right\|^2 - 0.1\|\omega_t^{\mathrm{r}} - \hat{\omega}_t^{\mathrm{r}}\|^2\right] \quad (4)$$

Each of the above four rewards is multiplied by its weight and divided by the sum of the weights to arrive at the final reward function r_t. It is represented by the following equation.

$$r_t = \frac{\mathrm{w_p} r_{\mathrm{p}} + \mathrm{w_e} r_{\mathrm{e}} + \mathrm{w_{rp}} r_{\mathrm{rp}} + \mathrm{w_{rv}} r_{\mathrm{rv}}}{\mathrm{w_p} + \mathrm{w_e} + \mathrm{w_{rp}} + \mathrm{w_{rv}}} \quad (5)$$

where $\mathrm{w_p}$, $\mathrm{w_e}$, $\mathrm{w_{rp}}$, and $\mathrm{w_{rv}}$ are the weight coefficients for each reward. The reward weights $(\mathrm{w_p}, \mathrm{w_e}, \mathrm{w_{rp}}, \mathrm{w_{rv}})$ are set to $(0.5, 0.3, 0.1, 0.1)$

We propose a method to improve the accuracy of location information by modifying the reward function of the existing method. Specifically, we add the following world coordinate reward r_{c} to the existing reward function.

$$r_{\mathrm{c}} = \exp[-\mathrm{k_c}((x_t - \hat{x}_t)^2 + (y_t - \hat{y}_t)^2)] \quad (6)$$

where (x_t, y_t) are the xy -coordinates of the attitude estimation result at time t, $(\hat{x}_t, \hat{y}_t)$ are the xy-coordinates of the ground truth at time t, and $\mathrm{k_c}$ is the decreasing rate of the function. In other words, the final reward function r_t of the proposed method is expressed by the following equation.

$$r_t = \frac{\mathrm{w_p} r_{\mathrm{p}} + \mathrm{w_e} r_{\mathrm{e}} + \mathrm{w_{rp}} r_{\mathrm{rp}} + \mathrm{w_{rv}} r_{\mathrm{rv}} + w_{\mathrm{c}} r_{\mathrm{c}}}{\mathrm{w_p} + \mathrm{w_e} + \mathrm{w_{rp}} + \mathrm{w_{rv}} + \mathrm{w_c}} \quad (7)$$

where $\mathrm{w_c}$ represents the weight of the world coordinate reward r_{c}. The experiment is performed with $\mathrm{w_c} = 0.1$.

III. Experiments

A. Environment

In this experiment, "mujoco" [4] is used as a physical control-based virtual environment to run the humanoid model, as in the previous method. The humanoid model used in the experiment consists of 58 degrees of freedom and 21 rigid bodies, with hips as the root joint. Most non-root joints have three degrees of freedom, but knees and ankles have only one degree of freedom. Figure 1 shows a diagram of the humanoid model used.

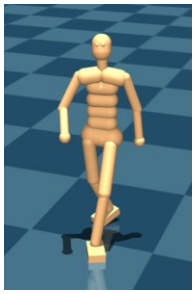

Fig. 1. Humanoid model used (mujoco)

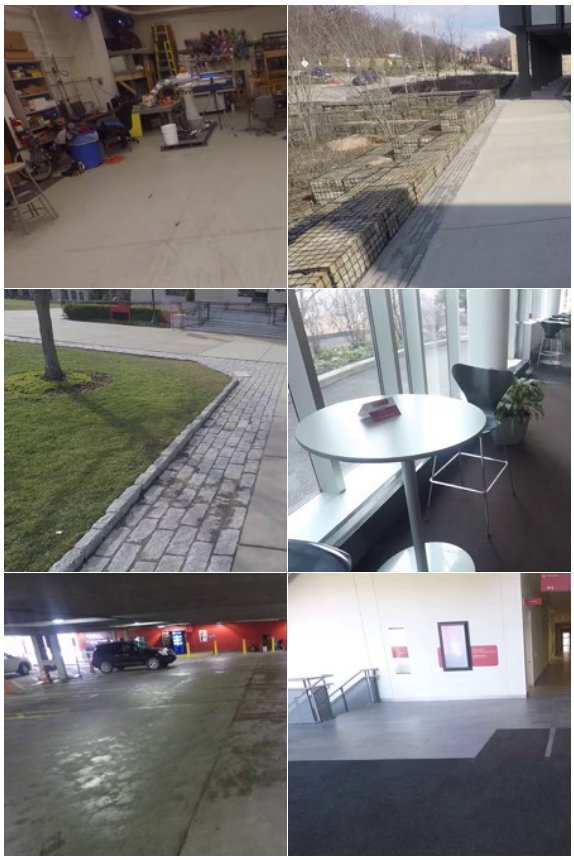

Fig. 2. Example of egocentric view image of dataset [1]

B. Dataset

As with the virtual environment, we use the same dataset as in the previous method. The dataset consists of a pair of egocentric video and posture information of the camera wearer. The dataset is constructed from the video and the poses of a GoPro camera wearer performing various actions (e.g., walking, jogging, bending, crouching, leaning, etc.). They are taken at various locations that should be indoors and outdoors. Each take is about one minute long, and they do not segment or label the motions. Each take consists of approximately 1,800 frames. The image size is 224×224. This is a new dataset created by the authors of the previous method. Figure 2 shows an example image of the data set used.

C. Result

The results of posture estimation for the existing and proposed methods are shown in Figure 3 when the first-person viewpoint image is used as input. The left images are the results of the existing method, and the right images are the results of the proposed method. In Figure 3, the humanoid in red is the ground truth, and one in yellow is the estimated result. A part of the image of the obtained posture estimation result is shown in the figure, separated by every fifty frames. In this case, $\mathrm{k_c} = 0.3$ is used.

Figure 3 shows that the difference between the ground-truth humanoid model and the estimated humanoid model position is smaller for the proposed method, indicating that the proposed method is more accurate in estimating the position. In both cases, the difference between the ground-truth and estimated positions becomes larger in the latter half of the frame. Figure 3 shows only one take, but the same can be said for the other takes.

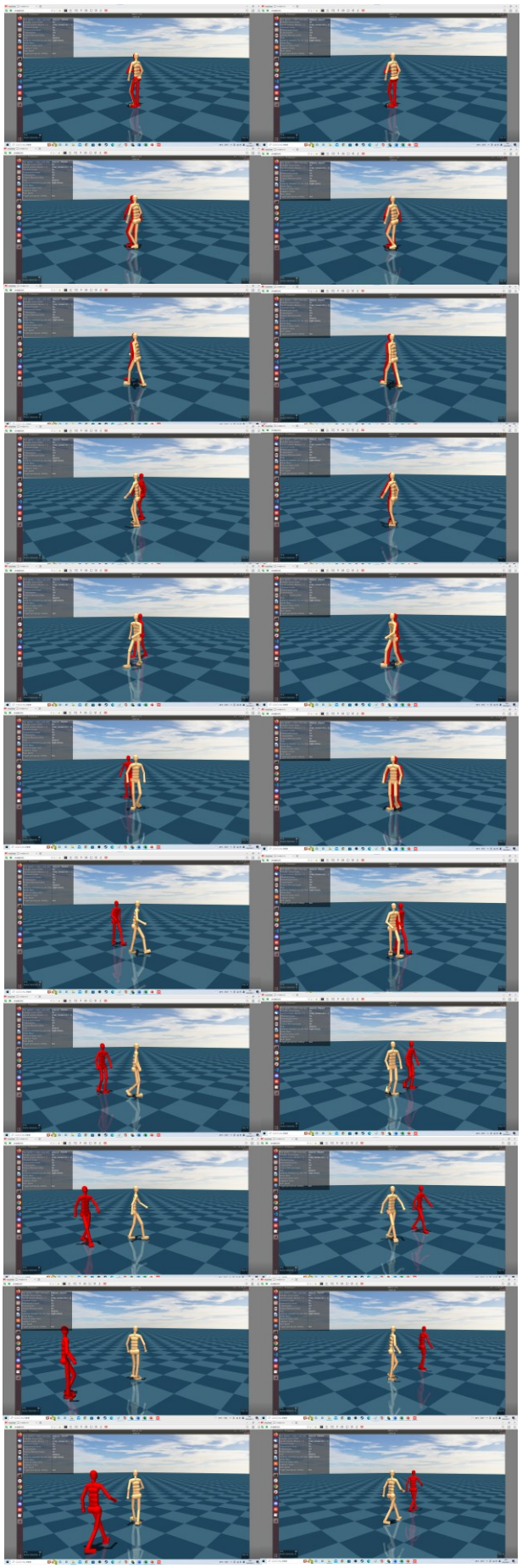

Fig. 3. **Left**: Results of conventional method. **Right**: Results of proposed method.

D. Evaluation Experiment

In order to quantitatively evaluate the proposed method, we conducted evaluation experiments. In addition to the evaluation items used in existing methods such as Pose Error (Epose), Velocity Error (Evel), and Average Acceleration (Aaccl), we added an additional item, Coordinated Error (Ecoor), to evaluate the improvement in accuracy of position information. The following is a description of each evaluation item.

- Epose: It measures the Euclidean distance between the generated pose sequence $p_{1:T}$ and the ground-truth pose sequence $\hat{p}_{1:T}$. It is represented by the following equation.

$$\frac{1}{T}\sum_{t=1}^{T}\|p_t - \hat{p}_t\|_2 \tag{8}$$

- Evel: It measures the Euclidean distance between the generated velocity sequence $v_{1:T}$ and the ground-truth velocity sequence $\hat{v}_{1:T}$. It is represented by the following equation. v_t and $\hat{v}_t$ are computed by the finite difference method.

$$\frac{1}{T}\sum_{t=1}^{T}\|v_t - \hat{v}_t\|_2 \tag{9}$$

- Aaccl: It uses the average magnitude of joint accelerations to measure the smoothness of the generated pose sequence. It is represented by the following equation. Where $\dot{v}_t$ is joint accelerations and G is the number of actuated DoFs.

$$\frac{1}{TG}\sum_{t=1}^{T}\|\dot{v}_t\|_1 \tag{10}$$

- Ecoor: It measures the difference between the xy coordinates of the generated pose sequence $(x.y)$ and the xy coordinates of the ground truth $(\hat{x},\hat{y})$. It is represented by the following equation. Instead of simply calculating the error at each time, it subtracts the error from the previous frame to simply measure how much the frame alone has deviated. Added because the existing evaluation item could not evaluate the improvement in the accuracy of location information.

$$\frac{1}{T}\sum_{t-1}^{T}\{\|x_t - \hat{x}_t\|_1 + \|y_t - \hat{y}_t\|_1 - (\|x_{t-1} - \hat{x}_{t-1}\|_1 + \|y_{t-1} - \hat{y}_{t-1}\|_1)\} \tag{11}$$

The values of each evaluation item for the existing and proposed methods when the value of k_c is varied from 0.3, 3, 30, 300, and 3000 are shown in Table I. The figures in parentheses in Table I show the percentage increase of each evaluation item compared to the existing method. For any evaluation item, the smaller the value, the better the result.

Table I shows that all of the proposed methods improve on Ecoor. However, other evaluation items such as Epose, Evel, and Aaccl have increased by several percent with some exceptions. This is not desirable, but Figure 3 does not show the impact of the increase in the other evaluation items. Table I also shows that, with the exception of $k_c = 30$, the smaller the value of k_c, the greater the improvement in Ecoor. The smaller the value of k_c, the smaller the rate of decrease of the

TABLE I. VALUES OF EVALUATION ITEMS AND PERCENTAGE INCREASE FOR EXISTING AND PROPOSED METHODS

method	**Evaluation Item**			
	Epose	***Evel***	***Aaccl***	***Ecoor***
walking	0.42 (+0.4%)	3.22 (+1.2)	3.93 (-0.4)	0.007 (-72.3)
jogging	0.59 (+4.9)	5.30 (+1.2)	7.19 (+1.6)	0.040 (-26.6)
crouching	0.71 (+10.0)	3.60 (+5.2)	3.29 (+5.1)	0.020 (+5.6)
leaning	0.55 (+4.8)	3.16 (+2.3)	3.46 (+5.4)	0.016 (-34.4)

TABLE II. VALUES OF EVALUATION ITEMS OF THE PROPOSED METHOD FOR EACH MOVEMENT AND THE RATE OF INCREASE

method	**Evaluation Item**			
	Epose	***Evel***	***Aaccl***	***Ecoor***
conventional	0.55	3.79	4.39	0.029
$k_c = 0.3$	0.56 (+3.3%)	3.84 (+1.3%)	4.51 (+2.7%)	0.018 (-37.1%)
$k_c = 3$	0.55 (+1.5%)	3.83 (+1.1%)	4.46 (+1.7%)	0.019 (-33.7%)
$k_c = 30$	0.54 (-1.2%)	3.77 (-0.3%)	4.60 (+4.9%)	0.025 (-13.2%)
$k_c = 300$	0.55 (+0.7%)	3.78 (-0.2%)	4.35 (-0.9%)	0.020 (-32.1%)
$k_c = 3000$	0.56 (+1.8%)	3.80 (+0.4%)	4.42 (+0.6%)	0.022 (-26.3%)

function and the smoother the decrease of the world coordinate reward r_c. The smaller the value of k_c, the greater the improvement in Ecoor, since this provides some reward for larger errors in xy-coordinates. The values of Epose, Evel, and Aaccl in Table I are found to increase in many cases compared to the existing method. The addition of the world coordinate reward has a negative effect on the generated poses as a result of the stronger forcing force on the position.

To show the effectiveness of the proposed method for each movement (walking, jogging, crouching, and leaning), Table II shows the values of each evaluation item of the proposed method for each movement and the amount of increase over the conventional method. For the proposed method, $k_c = 0.3$. As in Table I, evaluation items other than Ecoor are also increased, but the amount of increase is only a few percent, indicating that Ecoor is greatly improved except for crouching. The large difference in the rate of increase in Ecoor for each operation is thought to be due to the smaller sample size by dividing the data set into separate data sets for each operation. In addition, the reason why there was no improvement in Ecoor in crouching is that the crouching movement by itself is only a height movement in the first place. Figure 4 also shows a comparison of the estimated results and ground-truth during crouching. It can be seen that the left side, the estimation result of the conventional method, is less accurate in location information than the right side, the estimation result of the proposed method. This seems to be a different result from Table II, but it is due to the fact that the number of samples was reduced for each operation, which caused some bad estimation results to have a larger impact on the whole. It is assumed that increasing the number of samples would also improve Ecoor during crouching. The estimation results for walking, which showed the most improvement in Ecoor in Table II, already show that it is possible to improve the accuracy of location information visually, as shown in Figure 3.

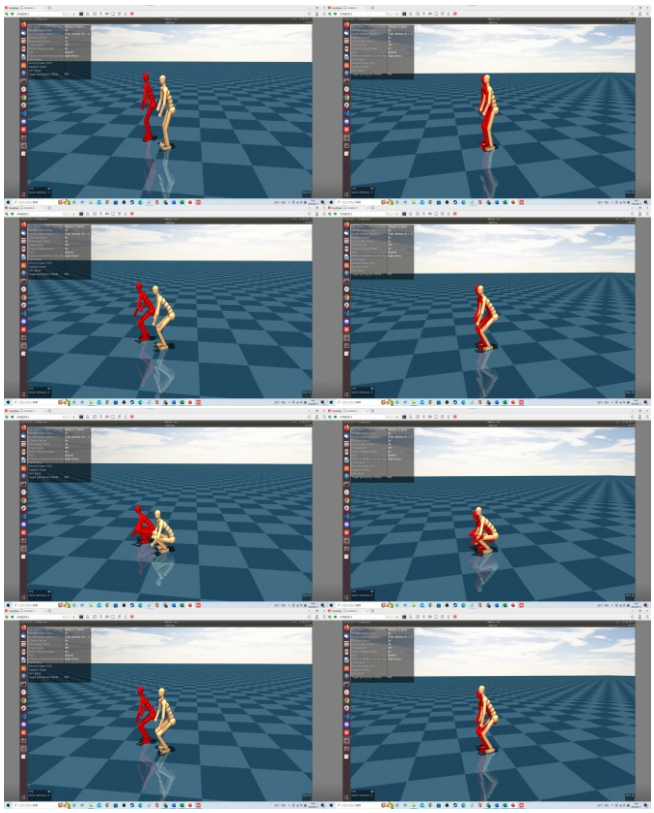

Fig. 4. Comparison of Estimation results and ground-truth during Crouching. **Left**: Results of conventional method. **Right**: Results of proposed method.

IV. CONCLUSION

We propose a new method to improve the accuracy of location information by adding a new world coordinate reward r_c to the conventional method. Comparison with the conventional method shows that the proposed method can estimate the location more accurately. Evaluation experiments show that the proposed method improves the accuracy of location information. when $k_c = 0.3$, the method improves by 37.1%. We also confirmed that the proposed method can improve the accuracy of location information for most movements.

Since this experiment was conducted using egocentric view images of the current world, new problems will emerge by using egocentric view images of the virtual world for evaluation against its use in VR games.

REFERENCES

[1] Y. Yuan and K. Kitani, "Ego-pose estimation and forecasting as real-time pd control," In IEEE International Conference on Computer Vision, pp.10082–10092, Oct, 2019.

[2] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.8934–8943, Jun, 2018.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.770–778, Jun, 2016.

[4] Mujoco, 2022, https://mujoco.org/, [access: 16/06/2022]