Real-time Lower-limb Motion Embodiment in Virtual Reality from a Single Waist-wearable Camera

Chenghao Xu , Lifeng Zhu , Aiguo Song
Jiangsu Key Lab of Remote Measurement and Control
School of Instrument Science and Engineering
Southeast University

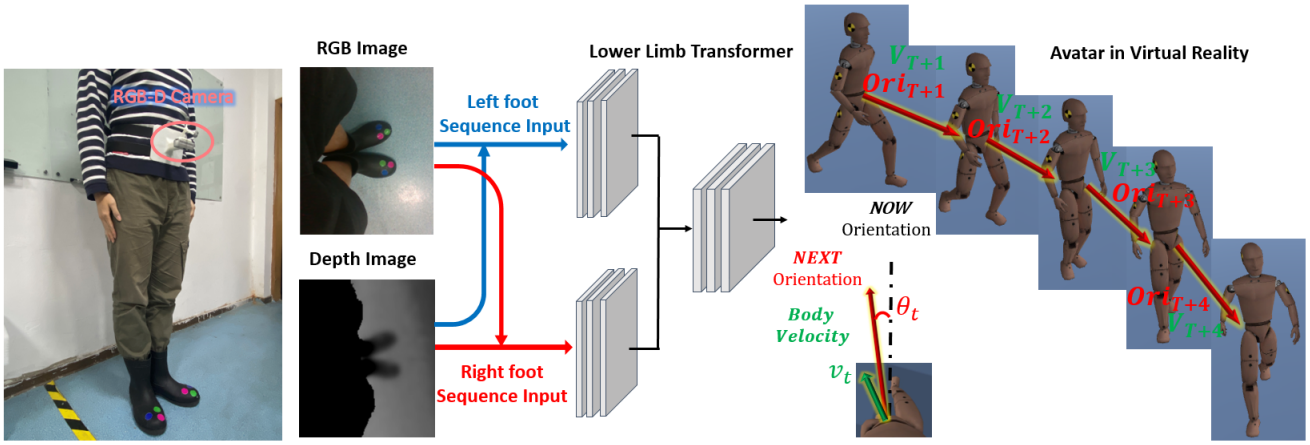

Figure 1: We propose a method for the embodiment of human lower-limb motion in VR using a single waist-wearable camera (left). We estimate the global velocity (lower right) by learning from the features of the feet observed by the camera (middle) and recover the lower-limb motions (upper right) with the body trunk moving according to the estimated global velocity. By augmenting the virtual avatar with the recovered motion, we produce lower-limb embodiment in VR.

Abstract

Background: For the interactions in virtual reality, it is essential to map the user's physical motion to the avatar in the virtual world. While reliable lower-limb motions are available under pre-installed cameras, the range of the walking motion is limited by the infrastructures. Method: We propose a new wearable solution to reproduce the lower-limb motions and map them to the virtual avatar in real time. We employ a single depth camera and design a waist-wearable layout to capture the lower-limb motions relative to the waist. By exploiting the vision data observed by the camera, we further estimate the global velocity of the user. Results: Experiments are carried out to verify our solution. We quantitatively evaluate the estimated global velocity with an optical motion capture system. We also map the recovered lower-limb motion to the avatar and utilize a standard questionnaire to measure the sense of embodiment. The experiments show that our wearable solution are feasible and effective, being applicable to different people from the perceptual perspective. Conclusions: The results verify that users are allowed to naturally explore the virtual world with the embodiment using the lightweight equipment.

***Keywords:** lower-limb , waist-wearable , VR embodiment*

1. Introduction

As one of the most important daily human activities, motion of lower limbs contains a great amount of information about human kinematics.

In today's virtual reality (VR) interactions, developers are able to reconstruct hand movements through hand controllers, while special devices for reconstructing lower-limb motions are less than hand tracking tools. However, when users are in virtual reality and cannot see their whole body in motion, it creates a strong sense of vertigo and unrealism [14]. Therefore, it is necessary to capture and map

the real lower-limb motion to the virtual avatar in real time for interactions in VR. The setup of base station limits the user's range of motion, as the user can only move and interact in a pre-defined area. Therefore, a solution that liberates movement restrictions is necessary.Wearable devices are a great solution to enlarge the interaction space.

Our ambition is to reconstruct the embodiment of the human lower limb in VR with just a single wearable camera. Due to the limited field of view (FOV) of a camera, it is not trivial to obtain sufficient visual information to recover the lower limb motions. Therefore, we need to work on the layout of the wearable camera for capturing lower-limb motions. Besides, it is challenging to effectively exploit the global information from the ego-centric view to track the lower limbs.

In this work, we design a waist-wearable layout for the real-time embodiment of human lower limb motion in VR, as shown in Figure. 1. The camera hanging from the waist shoots down to capture the movement of the feet. Since only the motion of the limbs relative to the waist can be obtained, the global motion of the body trunk cannot be directly captured. To solve this problem, we propose a learning-based method–Lower Limb Transformer, which can effectively cope with the problems of occlusion, to further estimate the linear and angular velocity of the body. We quantitatively evaluate the reconstruction quality of our method. In addition, we conduct user experiments to evaluate the perception of reconstructed lower limbs for the virtual embodiment in VR.

Our contributions in this paper are summarized as follows:

- We propose a novel waist-wearable layout for the embodiment of human lower limb motion in Virtual Reality.
- We propose Lower Limb Transformer(LLT), specifically designed for the reconstruction of lower limb motions with global movements.
- We evaluate the quality of the lower-limb motion and the perceptual acceptance of the virtual embodiment.

2. Related Work

In virtual reality, embodiment of the lower limb has been an issue worthy of study. Kilteni *et al*. [9] defined the sense of embodiment toward a proxy body when its properties are processed as if they were the properties of one's own biological body. The lower-limb embodiment in VR is usually separated into three different levels[9], the sense of self-location, agency, and body ownership.

Many researches are continuously exploring the feasibility and optimization of human embodiment [3]. Embodiment of lower limb is also a major topic of embodiment [20, 11]. As mentioned in[29], human motion in the virtual world does not have to be exactly identical to real-world motion for embodiment. Usually, in the virtual world, users are not able to perceive subtle deviations, such as minor changes in direction or velocity. [40] Researches[28, 1] have also shown that the human body has a certain tolerance for lower limb movement errors in VR. With the observation, the embodiment of lower limbs in the virtual world has its practical significance, such as redirected walking [25, 21].

Lower-limb motion capture is closely related to the virtual embodiment of lower-limb and other interactions in VR. In general, the highest accuracy can be achieved by using external devices and reflective markers for lower limb motion capture. For instance, marker-based optical motion capture (Mocap) systems [18, 16]. There are also monocular systems that do not rely on markers [6, 17, 19, 13]Although external devices can provide lower-limb motion with high accuracy, there are some insurmountable drawbacks due to their measurement methods. One is that such systems need to be installed in special laboratories, which requires a complex and costly installation process and takes a lot of time for preparation. Secondly, the measurement range is limited by the monitoring area of the cameras, which does not allow users to move freely.

This limitation can be lifted by a wearable device. Wearable IMU-based lower-limb motion reconstruction is one solution. It integrates the sensed acceleration and angular velocity to recover the motion trajectory of the foot. Existing gait analysis methods based on wearable inertial sensors commonly use the zero-velocity update (ZUPT) [27, 31] to eliminate integration errors accumulated in the swing phase, and the effectiveness of ZUPT relies on the accurate detection of the fully standing phase[39]. However, it is difficult to find a threshold with universal applicability among the existing methods for determining the real standing phase[32, 38]. Yi *et al*. [35] proposed a real-time motion reconstruction method that uses six IMUs to estimate the pose and global translation by fitting the body size through a neural network. However, each sensor is usually limited to measure properties at a fixed location, and it is difficult to match information between multiple sensors.

Another type of wearable solution is to use wearable cameras [7, 22, 10]. However, these systems were not able to find the global motion in the world coordinate system. Meanwhile, some wearable camera solutions use an inside-out approach to estimate human root motion by shooting outward from a camera tied to user's body [26, 36, 37]. They relied on bundle adjustment to solve for the transform of the camera. However, if no prominent visual features were inside the camera view, the SLAM-based method may lose tracking of the camera.

In contrast to these methods, our approach attempts to estimate the translation and rotation of the entire body as

well as the lower limb motion in the world coordinate system through a wearable depth camera.

In addition, multi-sensor fusion methods are also emerging. Cha *et al*. [2] proposed a real-time system for dynamic 3D capture, relying on cameras embedded in a head-mounted device and IMUs worn on the wrist and ankle. Machine learning is used to estimate the wearer's motion combining the inputs from vision and inertial sensors. Winkler *et al*. [33] proposed a reinforcement learning framework that takes in sparse signals from an HMD and two controllers to simulate full body motions. However, such multi-sensor fusion methods are not specifically designed for the motion of the lower limbs. They are all data-driven to regress the feet motion through sensors worn on the head and hands. Besides, they all use multiple sensors, while we try to work with only one wearable camera.

3. Method

Our ultimate goal is to reconstruct the motion of the human lower limb for virtual embodiment with just a single wearable camera. We will first talk about the wearable capturer settings as well as the data acquisition step. Then we will introduce our Lower Limb Transformer(LLT) to extract the global motion from the captured local signals. At last, we talk about how we drive the virtual avatar with the reconstructed motion.

3.1. Data Acquisition

Our key idea is to adopt wearable vision for lower-limb motion embodiment. Limited by the FOV of the camera, it is difficult to capture sufficiently large views of the entire lower limbs from a single camera. Inspired by the headset capturer in [2], we propose to use the downward views from the camera. Instead of installing the camera on the headset and guessing the knee joints from the captured leg motion, we propose to estimate the lower-limb motion by directly tracking the feet, which are the end-effectors of the lower limbs. In this case, we are able to reliably find the low-limb motion from data-driven inverse kinematics [23, 8, 15]. Considering that the feet are mostly visible to the waist, we design to install the camera at the waist in that the body trunk is relatively stable and slow in motion than human head, and the image quality is expected to be better than headset capturers.

We make a prototype of our waist-mounted capturer to implement the designed layout, as shown in Figure. 2. We 3D-print a platform to fix the camera and stick the platform to a belt. With the wearable design, the camera is considered to be rigidly attached to the waist without noticeable sliding. If the camera is tightly attached to the waist to shoot down, it will be obscured by clothing. We optimize the layout for a better view by suspending the camera 10cm away from the waist. In our physical prototype, we use the

Figure 2: The prototype of our waist-mounted capturer. The extended brunches fixed with reflective markers are used to collect the training data. They can be removed after the training stage.

Realsense depth camera, which can acquire both color images and binocular IR images for the estimation of the depth channel. Similar to the work [12], we set up three colored marker on each shoe to facilitate tracking the feet, as shown in Figure. 2.

There are still challenging problems to recover the virtual embodiment from the captured images. Because we track the feet in the view of the waist-mounted camera, we may only directly estimate the low-limb motion relative to the waist. For embodiment in VR, we also need to track the global motion of the user so that we are able to well align the virtual avatar to the VR scene. Because the HMD only tracks the head motion, which may move when the user observes the VR scene, we cannot directly use the tracked motion of the HMD to define the global motion of the virtual avatar. Therefore, we will work to compute for the global motion of the body trunk using the data from the waist-mounted camera. Our insight is that the relative motion of the feet hides the information of the global velocity of the body trunk. As illustrated in Figure.3, the gait from the egocentric view is different when the user straightly walks or makes a turn. The two feet are not always visible to the waist-mounted camera. We will explore how to work with the unstructured data for producing the virtual embodiment of lower-limb motion.

3.2. Lower Limb Transformer

With the captured images from the waist-mounted camera as the input, we introduce our Lower Limb Transformer (LLT) for producing the global velocity of the body trunk, which will be later used in the virtual embodiment. The

Figure 3: The gait from the egocentric view is different when the user walks or makes a turn.

pipeline of LLT is illustrated in Figure. 4. In this work, we model the human as a rigid body rather than a mass point. The output global velocity includes a translational and a rotational component. Because the captured images are from an egocentric view, in this stage, we also model the output translational velocity $\mathbf{v}_i$ and the rotational angle θ_i at the ith frame in the egocentric coordinate system. In other words, at each time step, we estimate the body trunk to move along the vector $\mathbf{v}_i$ in the egocentric view and the marching direction turns about an angle of θ_i, as shown in Figure. 5.

3.2.1 Feature Extraction

Here we introduce the feature extraction stage from the captured images. Since the user may walk on the floor with unknown patterns or wear different clothes, it is not reliable to transfer the learned network for arbitrary users in different trials. We, therefore, propose to extract features using 3D vision based on the designed wearable solution. Specifically, we track the relative pose of each foot with respect to the camera as the features using the markers on the shoes (Figure.2).

To extract the 3D vision features, we first segment the colored markers in the image. The RGB images are converted into HSV channels, denoised and segmented using color thresholds in real time, similar to [12]. With the aligned depth channel, each segmented marker point in the image coordinate system is assigned with its depth, forming its 3D coordinate in the camera coordinate system.

Because the 3D position of each visible marker may not be accurate due to the various lighting conditions or motion blur, we further extract the 3D transform of the foot to regulate the tracked markers. Specifically, we store the 3D positions of the markers on the left foot $\bar{\mathbf{p}}_i$ and those on the right foot $\bar{\mathbf{p}}_i$ in the standing pose as the initial state. Suppose in the ith frame, the visible markers are extracted to be $\mathbf{p}_i$ and $\mathbf{q}_i$. We solve for the best rigid transform of the left foot by fitting $\mathbf{p}_i$ to $\bar{\mathbf{p}}_i$ using Procrustes methods[5]. Because the markers on one foot may not always be visible to the camera, we solve for the 3D rigid transform when all the three markers are visible. If only two markers are visible, we assume the foot does not twist and rigidly transform on a plane parallel to the ground. In this case, a 2D rigid transform is fitted. If more than two markers are lost in the 3D vision, we do not solve the rigid transform at the frame and tag it as invisible. The same process is adopted for the right foot.

After the feature extraction stage, we organize the features sent to the LLT network as $X^{(l)} = \{\mathbf{x}_i^{(l)}\}$ and $X^{(r)} = \{\mathbf{x}_i^{(r)}\}$. At the ith frame, the feature vector corresponding to the left foot $\mathbf{x}_i^{(l)} = \{\mathbf{t}_i^{(l)}, \mathbf{u}_i^{(l)}, \mathbf{v}_i^{(l)}\}$, where $\mathbf{t}_i^{(l)}$ is the tracked translation of the left foot, $\mathbf{u}_i^{(l)}$ and $\mathbf{v}_i^{(l)}$ are the first two columns of the tracked rotation matrix. The rotational component for the neural network is organized as suggest in [41]. We stack the feature vector $\mathbf{x}_i^{(l)}$ into the time series $X^{(l)}$ and feed the time series into the LLT network along with the time series of the visibility tag $S^{(l)} = \{s_i^{(l)} \mid s_i^{(l)} = 0 \ or \ 1\}$. The features for the right foot $X^{(r)}$ are organized in the same way.

3.2.2 Network Architecture

In order to learn from the time series with patterns and missing data, we design a network architecture using the self-attention mechanism proposed in [30]. To this end, we propose the Lower Limb Transformer, an attention-based two-stream network. The structure of the LLT network is shown in Figure.4. We first deal with the features extracted from the tracked left and right feet separately in two streams. For the ith time step, we design a window size N and feed the features of the left and right feet as a time sequence starting from its previous N frames. Two Transformer blocks are used to extract deep features from the feature sequence of the left foot $X^{(l)}$ and the right foot $X^{(r)}$, respectively. Facing the case of missing foot features, we use the visibility tag $S^{(l)}$ and $S^{(r)}$ to trigger the mask operation in order to keep invisible data out of the calculation of self-attention. Suppose the outputs of the single-foot masked-attention block are $Y^{(l)}$ and $Y^{(r)}$ respectively, we concatenate them to form a feature $Z = Concat(Y^{(l)}, Y^{(r)})$ as the integrated feature of two feet. Observing that there are states that both feet are visible or one of them is visible, we design the two-stream structure to dynamically trigger the features from the two feet according to the visibility tags. In this case, the network produces useful deep features in presence of unstructured visibility states.

Next, the concatenated feature Z is fed into another attention block for capturing the correlation of the overall foot motion in the time window. The final output is passed through a $MLP + FC$ blocks [24], where the output is the linear and angular velocity in ego-centric coordinate system over a sequence of time. To provide supervision for training the lower limb transformer, we define the following loss

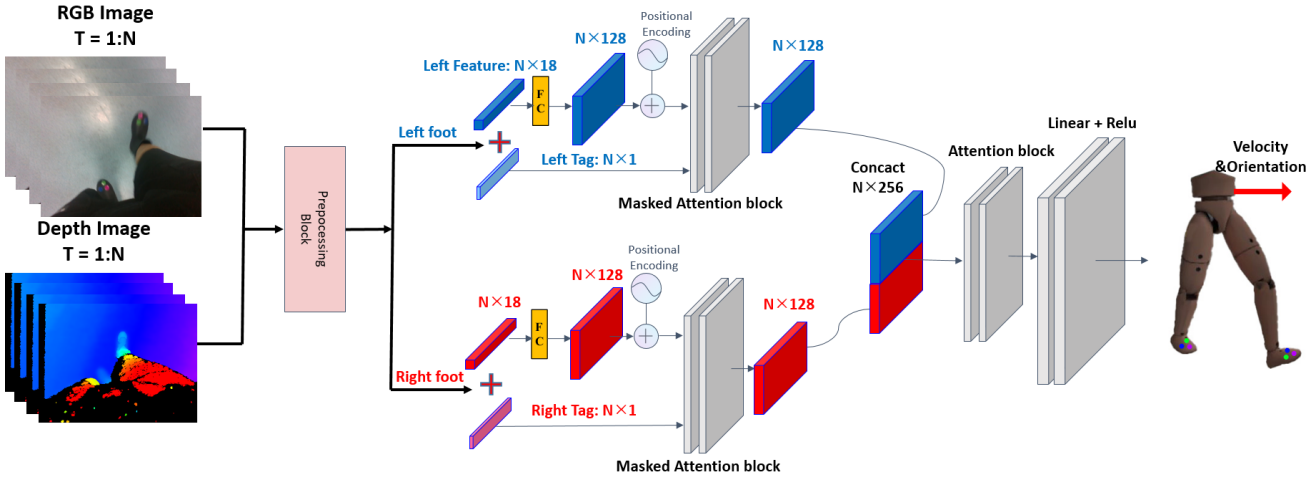

Figure 4: The pipeline of our proposed Lower Limb Transformer.

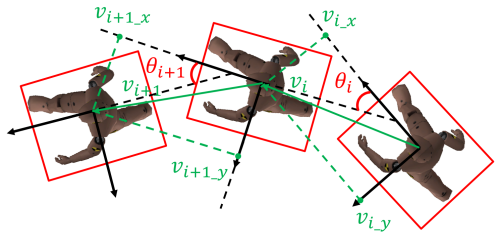

Figure 5: Linear velocity $\mathbf{v}_i$ and the rotational angle θ_i at the ith frame in the egocentric coordinate system

function:

$$L = L_v + \lambda L_\theta \tag{1}$$

where we use the mean squared error (MSE) for both the linear velocity loss L_v and angular velocity loss L_θ, and λ weights those two components.

3.2.3 Training

Dataset. Because the wearable layout of the camera is new, we construct a dataset of the captured images from the waist-mounted camera as well as the corresponding velocity of the body trunk by ourselves. As shown in Figure.2, we attach reflective markers and use an optical Mocap system, OptiTrack, to precisely track the rigid transform of the camera. Because the camera is rigidly attached to the waist, we extract the global velocity of the body trunk with the tracked motion and convert the velocity to the ego-centric coordinate system as the training data. We have recruited three normal users walking freely with the waist-mounted camera. In total, we have collected a dataset of more than 1×10^5 frames of the captured images and the tracked velocity of the body trunk as the training data.

Network Training. For the robustness of the network and sufficiency of training data, we utilize a sliding window (size $N = 50$) for the training data preparation. We use the Adam optimizer with a learning rate of 1×10^{-2} and train for nearly 20 epochs for convergence. The total number of epochs is 30 with proper early stopping and the batch size is 32. Moreover, we set the $\lambda = 1$ in loss.

Implementation details. We have high requirements for the real-time performance of the network. Since the output data of the pre-processing module is around 25FPS, if the neural network processing is too slow, it will lead to lagging of the embodiment in VR. We use the last frame of the output sequence as the estimated velocity.

Our computational setting includes a desktop with NVIDIA RTX 3080 Ti,i5-10500 CPU for training and evaluating the model. The Unity3D editor is used to run inverse kinematics and visualize the virtual avatar in VR.

3.3. Lower Limb Motion Embodiment

With the tracked feet and the regressed global velocity, we are ready to complete the virtual embodiment for the lower-limb motion in real time. First, we use inverse kinematics to trigger the local lower-limb motion with the tracked feet as the end-effector motion. In our implementation, we use a template rigged mesh of an average human body as the proxy, convert the tracked feet to the coordinate system at the root of the kinematics chain and solve for the lower-limb joint angles using inverse kinematics. [23, 8]. With the obtained joint angles, we can embody the low-limb motion into a virtual avatar by applying forward kinematics.

In order to reproduce the full 3D motion of the virtual avatar, we integrate the global velocity obtained from LLT to translate and rotate the virtual avatar. As for the camera

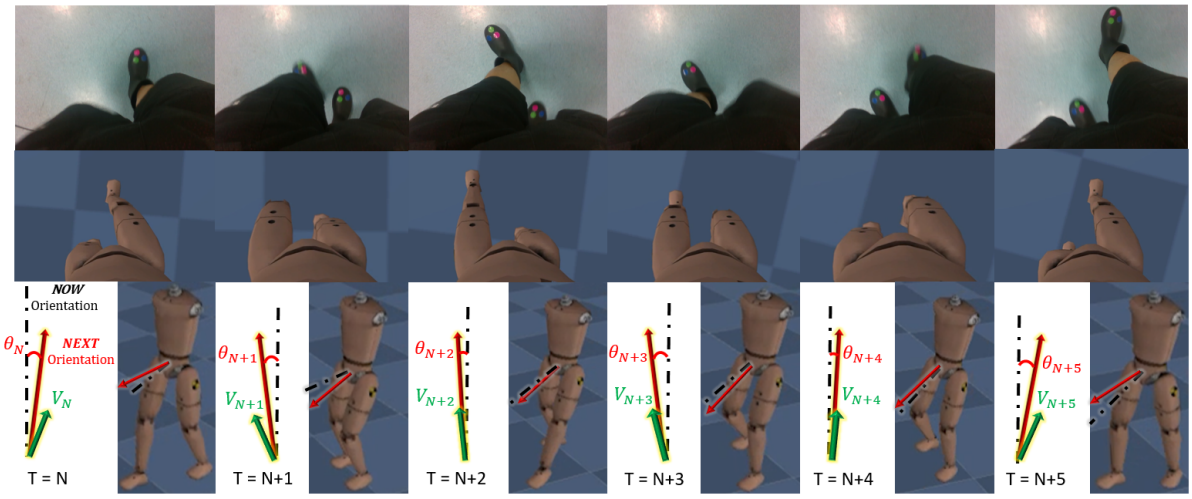

Figure 6: The lower limb embodiment (middle row) from the captured images (upper row) from the waist-mounted camera. We also plot the estimated velocity and the reconstructed motion under the third-person observation.

control in the virtual scene, in our prototype system, we use the Oculus RiftS as the HMD, which is free of base stations. We fix the viewpoint of the VR headset at the head position of the virtual avatar to produce a first-person perspective.

4. Experiment

4.1. Results

In this section, we conduct qualitative and quantitative experiments on our proposed system.

We tested our system by wearing the camera and freely moving in a room. The virtual embodiment was obtained and displayed in the HMD in real time. We show one clip of the results in Figure.6. In the top and middle row of the figure, we compared the captured raw images from the waist-mounted camera and the recovered motion of the virtual avatar. In the lower row of Figure.6, we also plotted the estimated velocity of the body trunk from the proposed LLT along with the third-person view of the reconstructed avatar motion.

The global velocity was aligned in the egocentric coordinate system with a dashboard style, where we used the black dashed line to illustrate the forward orientation of the body in the current time step. The translational velocity V_i was represented by the green arrow. The red arrow indicated the orientation of the body in the next frame and the estimated angular displacement θ_i was the angle between the orientations of the two frames. We integrated the global velocity in the egocentric coordinate system and shown the corresponding 3D motion from a third-person perspective at the side. By observing the grid pattern on the ground, the motion was perceived to match with the input.

4.1.1 Quantitative Evaluation

In our quantitative evaluation, we compared the reconstructed motion with the results from a Mocap system as the reference. The user wore our waist-mounted camera and was allowed to walk freely walked in a room with the OptiTrack as the Mocap system. We attached reflective markers on the waist-mounted camera to track its position.

After the capture, we extracted the velocity of the tracked camera using Mocap and converted it into the egocentric coordinate system as the reference data. Then we compared the reconstructed velocity from our system and the results were plotted in Figure.7. In all test datasets, the average error of the output body speed is 0.0288m/s and the average error of the angular velocity is 0.0234rad/s.

We also integrated the estimated velocity to illustrate the walking trajectory of the body trunk as shown in right column in Figure.7. The trajectory captured by OptiTrack was also drawn in the subfigures.

By comparing the integrated trajectories, we found the recovered trajectories from our wearable capturer are similar to the real data, but we could still observe deviations. Especially with the increase of time, the gap of trajectories became larger. This drift is due to the accumulation of the velocity error in the integration. It is common for wearable capturers and known as the dead reckoning in source-free navigation systems [34]. Note that it is challenging to quantitatively compare the global trajectory using inside-out cameras and the global trajectories usually were not compared in previous studies.

For the applications of virtual embodiment in VR, as the users have tolerances on perceiving the distortion of the walking view, the trajectory of the virtual avatar does not

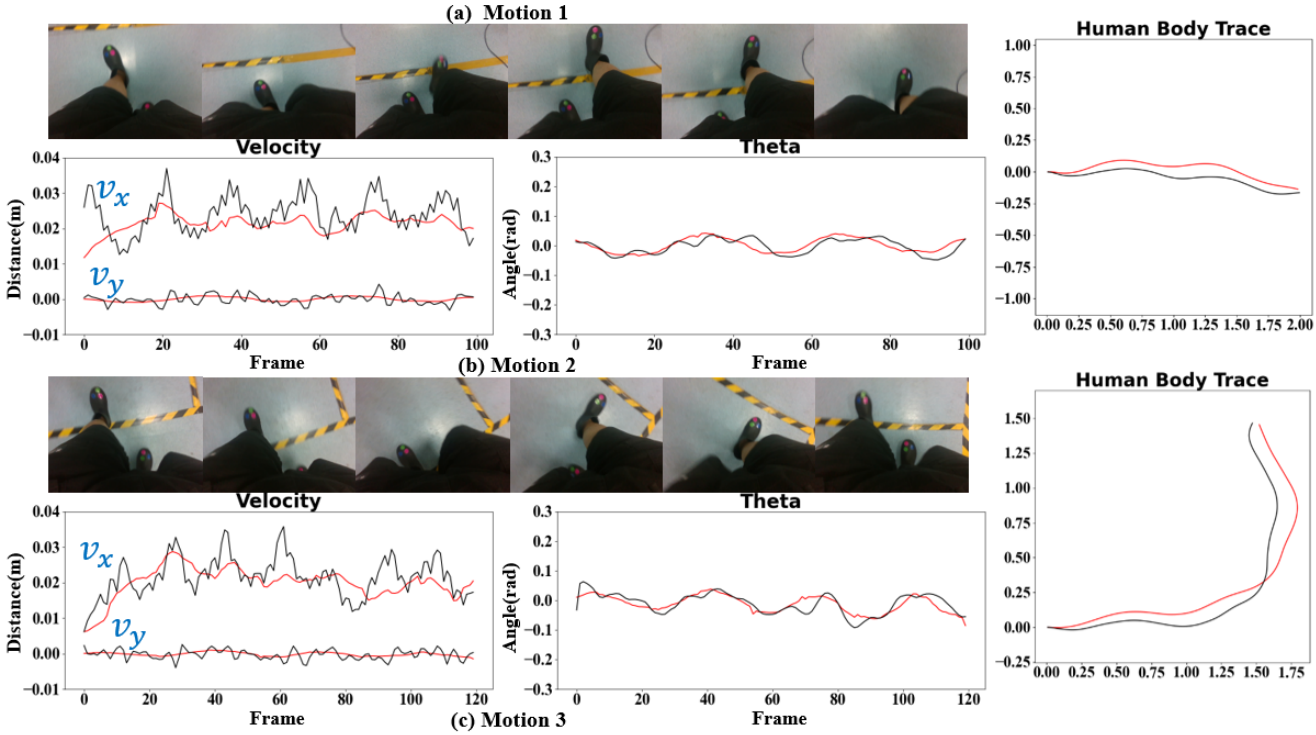

Figure 7: The estimated translational velocity (lower left) and angular velocity (lower right) of straight walking compared with the optical Mocap results. We plot our recovered velocity in red lines and the reference data in black lines. The comparison between recovered trace (red lines) and the reference trace (black lines) obtained from Mocap systems.

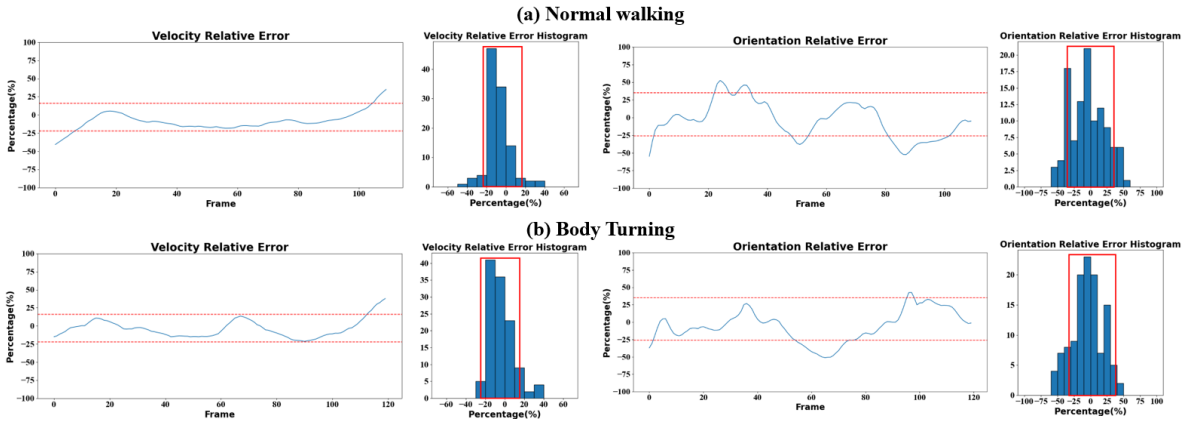

Figure 8: The relative error of velocity and angle velocity while users making different kinds of motions

have to be exactly the same as the real trajectory [21, 29]. As long as the estimated egocentric velocity from our framework is within a certain range, the user will not perceive such deviation of the embodied motion. According to [29], the deviation in speed should be within downscaling by 14% and upscaling by 26%. Users can be turned physically about 49% more or 20% less than the perceived virtual rotation. Note that these scales are measured for mapping from the virtual walking to the real world. For the virtual embodiment, we essentially develop an inverse map from the real-world motion to the virtual avatar. Therefore, the tolerance on the scale of the linear speed is in $[-20.635\%, 16.276\%]$, and the relative error of the angular speed should be in $[-32.888\%, 25\%]$ according to [29]. We collected the relative errors of the estimated velocity and plotted them by checking with the perceptual range, as

shown in Figure.8.

From the results, we found that the relative error of the speed located inside the range of the reported perceptual tolerance at most of the time. We also checked the histogram of the relative error. In the test set of straight walking, the percentage of relative error of linear speed falling in the ideal interval is 95% and the percentage of relative error of angular speed falling in the ideal interval is 71%, while 94.17% and 77.5% during turning right and 82.73% and 71.81% during turning left. As the majority of the data were within the perceptual tolerance and we did not observe extremely large relative errors in a continuous time slot, we expected the virtual embodiment worked well in terms of perception by inspecting the objective data.

4.1.2 User Study

We further conducted a user study and collected subjective data to check whether the final embodiment was accepted by common users. We recruited 10 participants in this study. They were university students (9 male) aged from 21 to 27 (mean 24.1). They all had either normal or corrected- to-normal vision. 5 of them had experience in exploring VR environments or playing VR games.

In the user study, subjects wore an Oculus headset and the designed belt mounted with the camera. Subjects were asked to walk freely around the field, completing straight and turning movements, and observe the virtual lower-limb embodiment through the HMD. We compared our method as the experimental mode to embodiment using a template motion as the reference mode. In the reference mode, we generated virtual embodiment using the same virtual avatar while its motion was simply created by rigidly transforming a pre-programmed template walking animation using the integrated path from the HMD. In the experiment, each participant tested the embodiment with the reference mode and experimental mode in a random order. Then, they were asked to score the embodiment of the lower limb on a questionnaire and interviewed for more subjective feedback.

We used a standard questionnaire to measure the sense of embodiment based on [4]. We adapted the questionnaire for the lower limb reconstruction, which consisted of nine items and three subsets of questions, including body ownership (Ownership), agency and motor control (Agency), and body position (Location), as listed in Table. 1. The users are allowed to actively observe their lower limbs in the virtual world. In addition, the avatar walks with the captured motion is compared to a pre-programmed template motion. Then the subscales of ownership, agency, and location scores as shown in Figure.9.

From the results of the t-test analysis, we found that there was a significant difference between the two groups of data (t=-0.42). Compared to embodiment using the template motion, the gait from our method was more consistent and realistic, yielding a satisfactory embodiment in our experiments. We found that the realistic motion played an important role in terms of body ownership and agency. While we assumed our score in terms of location might not outperform the reference mode, the results still showed significance because the template gait did not match with the real

Table 1: Questionnaire to measure the sense of embodiment

Subscale	Question
Ownership	1)I felt as if the virtual representation of the foot moved just like I wanted it to, as if it was obeying my will. 2)I felt as if the virtual representation of the lower limb was someone else's. 3)It seemed like the lower limb belonged to me.
Agency	1)It felt like I could control the virtual lower limb as if it was my own body. 2)The task was easy to perform. 3)I felt as if the virtual avatar was moving by itself.
Location	1)I felt as if my body was located where I saw the virtual body. 2)I felt out of my body. 3)I felt dizzy during the experiment.

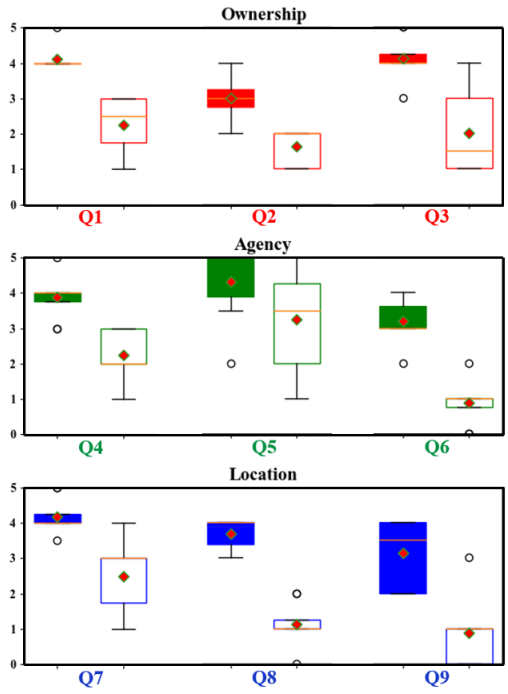

Figure 9: User-rated scores of the questionnaire. The solid bar represents the scores using our method and the hollow bar represents the scores using template motion.

walking velocity, making it hard to correctly estimate the real location.

After the experiment, we interviewed each participant and asked them about their subjective feelings about the lower-limb embodiment. Most participants believed that the virtual embodiment corresponded to their real lower limbs most of the time.

4.2. Discussion

In this work, we work for virtual embodiment of the lower-limb motion from a single wearable camera. It can be applied to various tasks in VR such as roaming in the virtual scene. The recovered lower-limb motions are also expected to be used in assessment tools for rehabilitation or as signatures for gait recognition. We test our method with various walking motions. Although walking is the most typical lower limb motions in VR exploration, we also expect to recover other lower-limb motions by incorporating more training data about different motions.

In addition, we adopt the first-person vision and innovatively locate the viewpoint around the waist. To the best of our knowledge, this is the first attempt to experiment with the lower-limb motion reconstruction with this view, whereas many wearable solutions are generally based on IMUs. In this work, we exploit modern machine learning algorithms with the self-attention mechanism and report the reconstructed lower-limb motions with only vision data from the new view. Results with better quality are expected if the inertia measurements are well fused into our pipeline.

In this work, the virtual embodiment is well perceived when the user looks downward in the virtual scene. We adopt the tracked transform of the feet to estimate the global velocity of the body trunk in presence of missing data due to occlusion. However, we do not learn for the unknown feet transforms due to occlusion. In order to produce the animation in a third-person view, in our current implementation, we extrapolate for the transforms of the invisible feet with a linear regression. In our application of virtual embodiment, we find the users in the experiment do not see their heels of the virtual avatar in casual VR interactions. Therefore, the simple treatment with the missing data does not affect the virtual embodiment in practice.

5. Conclusion

We propose a wearable solution for the virtual embodiment of lower-limb motions based on a single camera. We design a novel first-person view of the camera by mounting it on the waist. By exploiting the vision information from the new perspective, we learn for the global velocity of the body trunk and successfully reconstruct the lower-limb motions in real time. We show that the recovered lower-limb motions work well for virtual embodiment based on the results from quantitative evaluation and subject studies. With the single wearable camera, we build our system with a low cost while the embodiment is recovered in real time, which shows its potential for VR systems in open environments.

In the future, we will add more lower limb motions with more diverse user datasets to improve the generalization and accuracy of our method. Moreover, the error of our wearable solution can also be reduced by incorporating with IMU or other sensors.

Acknowledgement

This work has been supported by the NSFC under Grants No.92148205, the Natural Science Foundation of Jiangsu Province under Grants No. BK20211159, and the Fundamental Research Funds for the Central Universities.

References

[1] G. Bruder, V. Interrante, L. Phillips, and F. Steinicke. Redirecting walking and driving for natural navigation in immersive virtual environments. *IEEE transactions on visualization and computer graphics*, 18(4):538–545, 2012. 2

[2] Y.-W. Cha, H. Shaik, Q. Zhang, F. Feng, A. State, A. Ilie, and H. Fuchs. Mobile. egocentric human body motion reconstruction using only eyeglasses-mounted cameras and a few body-worn inertial sensors. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 616–625. IEEE, 2021. 3

[3] A. C. S. Genay, A. Lécuyer, and M. Hachet. Being an avatar" for real": a survey on virtual embodiment in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 2021. 2

[4] M. Gonzalez-Franco and T. C. Peck. Avatar embodiment. towards a standardized questionnaire. *Frontiers in Robotics and AI*, 5:74, 2018. 8

[5] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 4

[6] K. Guo, F. Xu, T. Yu, X. Liu, Q. Dai, and Y. Liu. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017. 2

[7] J. Healey and R. W. Picard. Startlecam: A cybernetic wearable camera. In *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No. 98EX215)*, pages 42–49. IEEE, 1998. 2

[8] Y. Jiang and C. K. Liu. Data-driven approach to simulating realistic human joint constraints. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1098–1103. IEEE, 2018. 3, 5

[9] K. Kilteni, R. Groten, and M. Slater. The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments*, 21(4):373–387, 2012. 2

[10] A. Kim, J. Kim, S. Rietdyk, and B. Ziaie. A wearable smartphone-enabled camera-based system for gait assessment. *Gait & posture*, 42(2):138–144, 2015. 2

[11] L. Kruse, E. Langbehn, and F. Steinicke. I can see on my feet while walking: Sensitivity to translation gains with visible

feet. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 305–312. IEEE, 2018. 2

[12] Y. Lee, W. Do, H. Yoon, J. Heo, W. Lee, and D. Lee. Visual-inertial hand motion tracking with robustness against occlusion, interference, and contact. *Science Robotics*, 6(58):eabe1315, 2021. 3, 4

[13] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 2

[14] C. Lopez, P. Halje, and O. Blanke. Body ownership and embodiment: vestibular and multisensory mechanisms. *Neurophysiologie Clinique/Clinical Neurophysiology*, 38(3):149–161, 2008. 1

[15] X. Lv, J. Chai, and S. Xia. Data-driven inverse dynamics for human motion. *ACM Transactions on Graphics (TOG)*, 35(6):1–12, 2016. 3

[16] M. Menolotto, D.-S. Komaris, S. Tedesco, B. O'Flynn, and M. Walsh. Motion capture technology in industrial applications: A systematic review. *Sensors*, 20(19):5687, 2020. 2

[17] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer vision and image understanding*, 81(3):231–268, 2001. 2

[18] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2-3):90–126, 2006. 2

[19] L. Mündermann, S. Corazza, and T. P. Andriacchi. The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications. *Journal of neuroengineering and rehabilitation*, 3(1):1–11, 2006. 2

[20] C. D. Murray. An interpretative phenomenological analysis of the embodiment of artificial limbs. *Disability and rehabilitation*, 26(16):963–973, 2004. 2

[21] N. C. Nilsson, T. Peck, G. Bruder, E. Hodgson, S. Serafin, M. Whitton, F. Steinicke, and E. S. Rosenberg. 15 years of research on redirected walking in immersive virtual environments. *IEEE computer graphics and applications*, 38(2):44–56, 2018. 2, 7

[22] G. O'Loughlin, S. J. Cullen, A. McGoldrick, S. O'Connor, R. Blain, S. O'Malley, and G. D. Warrington. Using a wearable camera to increase the accuracy of dietary analysis. *American journal of preventive medicine*, 44(3):297–301, 2013. 2

[23] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 3, 5

[24] A. Pinkus. Approximation theory of the mlp model in neural networks. *Acta numerica*, 8:143–195, 1999. 4

[25] S. Razzaque. *Redirected walking*. The University of North Carolina at Chapel Hill, 2005. 2

[26] T. Shiratori, H. S. Park, L. Sigal, Y. Sheikh, and J. K. Hodgins. Motion capture from body-mounted cameras. In *ACM SIGGRAPH 2011 papers*, pages 1–10. 2011. 2

[27] I. Skog, P. Handel, J.-O. Nilsson, and J. Rantakokko. Zero-velocity detection—an algorithm evaluation. *IEEE transactions on biomedical engineering*, 57(11):2657–2666, 2010. 2

[28] F. Steinicke, G. Bruder, J. Jerald, H. Frenz, and M. Lappe. Analyses of human sensitivity to redirected walking. In *Proceedings of the 2008 ACM symposium on Virtual reality software and technology*, pages 149–156, 2008. 2

[29] F. Steinicke, G. Bruder, J. Jerald, H. Frenz, and M. Lappe. Estimation of detection thresholds for redirected walking techniques. *IEEE transactions on visualization and computer graphics*, 16(1):17–27, 2009. 2, 7

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[31] J. Wahlström and I. Skog. Fifteen years of progress at zero velocity: A review. *IEEE Sensors Journal*, 21(2):1139–1151, 2020. 2

[32] J. Wahlström, I. Skog, F. Gustafsson, A. Markham, and N. Trigoni. Zero-velocity detection—a bayesian approach to adaptive thresholding. *IEEE Sensors Letters*, 3(6):1–4, 2019. 2

[33] A. Winkler, J. Won, and Y. Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. *arXiv preprint arXiv:2209.09391*, 2022. 3

[34] Y. Wu, H.-B. Zhu, Q.-X. Du, and S.-M. Tang. A survey of the research status of pedestrian dead reckoning systems based on inertial sensors. *International Journal of Automation and Computing*, 16(1):65–83, 2019. 6

[35] X. Yi, Y. Zhou, and F. Xu. Transpose: real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 2

[36] Y. Yuan and K. Kitani. 3d ego-pose estimation via imitation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 735–750, 2018. 2

[37] Y. Yuan and K. Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10082–10092, 2019. 2

[38] R. Zhang, M. Loschonsky, and L. M. Reindl. Study of zero velocity update for both low-and high-speed human activities. *International Journal of E-Health and Medical Communications (IJEHMC)*, 2(2):46–67, 2011. 2

[39] R. Zhang, H. Yang, F. Höflinger, and L. M. Reindl. Adaptive zero velocity update based on velocity classification for pedestrian tracking. *IEEE Sensors journal*, 17(7):2137–2145, 2017. 2

[40] Y. Zhang and J. Hong. Direction change of redirected walking via a single shoe height change. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 375–376. IEEE, 2021. 2

[41] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 4