Improving motion matching for VR avatars by fusing inside-out tracking with outside-in 3D pose estimation

George Fletcher
University of Bath
gf321@bath.ac.uk

Donal Egan
Trinity College Dublin

Rachel McDonnell
Trinity College Dublin

Darren Cosker
University of Bath, Microsoft

CCS CONCEPTS

• **Computing methodologies** → **Animation**.

KEYWORDS

Animation, VR, Self-Avatars, Motion Matching

1 INTRODUCTION

Achieving high quality animation for VR self-avatars on consumer devices that retain a sense of embodiment and responsiveness is an important, yet difficult task. A key factor in the difficulty is the limited sensor data available from typical 'inside-out' tracking HMDs, whereby the headset orients itself with respect to the environment using sensors on the headset (as opposed to 'outside-in', which makes use of external sensors) to generate sparse 6DoF spatial signals for the HMD and two controllers/hands. Due to this, it has been common to animate VR self-avatars by using a mixture of inverse kinematics (IK) and animation state machines, driven by various heuristics and assumptions, such as the character root/hips being directly below the user's head, and changing to different locomotion modes as the HMD moves vertically (e.g. crouching). However, such methods are difficult to design so that they produce both high animation quality and strong correspondence with a user's pose.

To improve fidelity and correspondence with the user, recent techniques leverage motion capture data to train machine learning models to estimate a user's pose [Du et al. 2023; Winkler et al. 2022] given the sparse signals from an inside-out tracking VR system. Such methods are promising, but iteration times are often too high for use in production scenarios such as game development.

Motion matching [Büttner and Clavet 2015] is a popular real-time animation method in industry due to its high quality output, lower iteration time, and relatively low inference time. Furthermore, the recent work MMVR by [Ponton et al. 2022] has demonstrated an application of motion matching to VR self-avatars, by generating the character trajectory required for motion matching, through the prediction of character root orientation (hips) from the sparse HMD and controller signals using a lightweight neural network (NN). Such an approach can provide a better approximation of user hip direction than naively assuming that the hips are always aligned with the HMD, but we found the network can struggle to generalise, given its limited capacity, likely producing an average of multiple plausible hip orientations (Figure 2: 30-40s).

Authors' addresses: George Fletcher, University of Bath, , gf321@bath.ac.uk; Donal Egan, Trinity College Dublin, ; Rachel McDonnell, Trinity College Dublin, ; Darren Cosker, University of Bath, Microsoft,

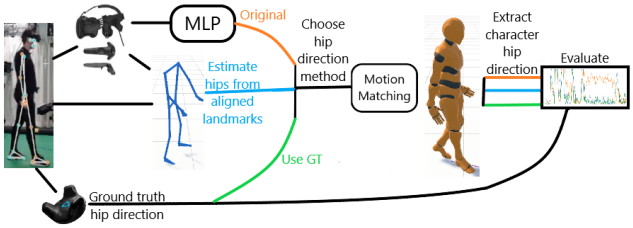

Figure 1: Hip direction drives MMVR [Ponton et al. 2022]. The original utilises an MLP to predict hip direction from HMD and controllers (orange). Our system aligns a 3D pose estimation signal (blue) to the HMD tracking signals to obtain a hip orientation estimate. We evaluate the resulting directional error (yaw) between the driven character's hip orientation and the ground truth user hip direction

In this work, we look to overcome this issue by including human pose estimation from webcam video, within the motion matching framework, in addition to the VR head and hand signals as used in MMVR. Our investigations highlight how this combined approach can reduce ambiguity when predicting the user's hip orientation, leading to more accurate body pose predictions.

2 APPROXIMATING 3D JOINT POSITIONS

We utilise an off-the-shelf 3D pose estimator: MediaPipe BlazePose GHUM 3D [Bazarevsky et al. 2020; Xu et al. 2020] to obtain camera-space 3D landmarks. We align the three 'Nose' and left/right 'Wrist' landmarks with the VR user's HMD and controller positions by:

(1) Manually correcting for the camera pitch so landmarks only need a further correction of yaw and translation
(2) Vertically aligning landmarks by either translating so that the lowest landmark is always on the ground, or aligning the 'Nose' landmark to HMD
(3) Performing a constrained least-squared minimisation (e.g. Kabsch, brute-force iterative) to find the optimal horizontal translation and yaw correction between the three 'Nose' and left/right 'Wrist' landmarks, with the positions, manually offset to correspond to the landmarks, of the HMD-provided HMD and controller positions.

The resulting aligned landmarks are a reasonable and fast, albeit crude, approximation of various user's landmark positions, notably the left and right 'Hip' landmarks, from which we can extract quantities with which to drive motion matching.

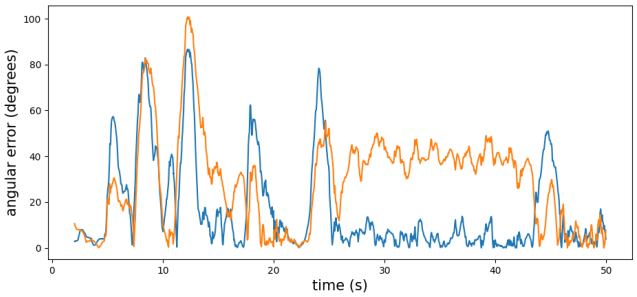

Figure 2: Plot of angular error on our test data. Blue curve shows our pose estimation driven hip direction with average error = 16.7 degrees. Orange curve shows error for MMVR's NN predicted hip direction; average error = 28.9 degrees.

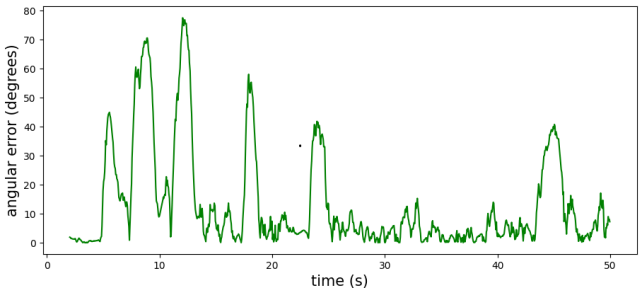

Figure 3: Plot of angular error between character hips and ground truth hip tracker when MMVR is driven directly by the ground truth hip tracker direction. Demonstrates the natural turning latency of the motion matching system. Average error = 13.3 degrees

3 DRIVING MOTION MATCHING

In traditional motion matching, one of the matching metrics computes the error in root position and root velocity between the current and candidate frames, and another, the error in future root trajectory at fixed times in the future, compared with an artificial trajectory generated by the 'gamepad' direction and magnitude. In MMVR, the root velocity and trajectory are generated by the predicted hip orientation and tracked HMD velocity in the same way a gamepad direction and magnitude generates a trajectory.

We observe that swapping out MMVR's NN predicted hip direction with a hip direction approximation from the aligned pose estimation landmarks, computed by a cross product of the left-to-right hip landmarks direction vector, with the world-space up vector, results in a lower RMSE between the resulting character's hip direction yaw angle, and the ground truth hip direction yaw angle on our preliminary test data, as shown in Figure 2. The sudden spikes in the figures occur in all regimes when the user turns their body and the motion matching system is yet to, or is in the process of, performing an appropriate transition to rotate the virtual character to follow. To validate this fact, we show the resulting error when the hip direction is driven directly by the ground truth data, as shown in Figure 3.

Our test data consists of signals produced by an HTC Vive system and Vive tracker attached to the waist, and synchronised pose estimation landmarks produced by a basic sequence of forward, backward and strafing locomotion. The data is then played-back in Unity to simulate the same VR and pose estimation inputs to quantitatively evaluate the different hip direction estimation methods (ours v.s. MMVR's NN v.s. ground truth) with regards to character and user hip alignment.

4 DISCUSSION AND FUTURE WORK

VR self-avatar locomotion is a challenging problem and we believe there are various avenues to improve the motion matching system's results (besides refining the original pose matching dataset).

The error of our system is often comparable to, or slightly worse than, the original approach during spikes in error due to turning, likely a result of the noise and latency of the pose estimation system or occlusion. However, note that such spikes also occur when driving using the ground truth hip direction (figure 3).

On the other hand, we noticed that our system tends to reduce the turning error more rapidly, such as in the time span 13-17s, and reduces bad cases of the NN staying misaligned for prolonged periods, such as 30-40s in Figure 2. We believe this may be due to issues with generalisation of the NN, or 'averaging' of the one-to-many mappings of HMD signals to hip direction, which the estimate of hip direction from pose estimation can overcome.

Also, we noticed that the motion matching system has a capability to handle the upstream noise from the pose estimation landmarks, such as due to occlusion, as the motion matching system usually either chooses not to transition due to the quality metric enforcing pose consistency, or quickly re-transitions back to a suitable animation in a way that is imperceptible due to inertialisation blending.

In future work, we seek to improve the base motion matching system via refining the pose data, but also experiment with improving the alignment, and using the pose estimation foot landmark data to guide the system to select transitions that lead with the same foot as the user during locomotion to improve lower body correspondence.

We are also interested in using this system to drive a non-humanoid avatar, such as a dog, and perform experiments of animal embodiment.

REFERENCES

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. BlazePose: On-device Real-time Body Pose tracking. *CoRR* abs/2006.10204 (2020). arXiv:2006.10204 https://arxiv.org/abs/2006.10204

Michael Büttner and Simon Clavet. 2015. *Motion Matching - The Road to Next Gen Animation.* https://www.youtube.com/watch?v=z_wpgHFSWss&t=658s

Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. 2023. Avatars Grow Legs: Generating Smooth Human Motion from Sparse Tracking Inputs with Diffusion Model. arXiv:2304.08577 [cs.CV]

J. L. Ponton, H. Yun, C. Andujar, and N. Pelechano. 2022. Combining Motion Matching and Orientation Prediction to Animate Avatars for Consumer-Grade VR Devices. *Computer Graphics Forum* 41, 8 (2022), 107–118. https://doi.org/10.1111/cgf.14628 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14628

Alexander Winkler, Jungdam Won, and Yuting Ye. 2022. QuestSim: Human Motion Tracking from Sparse Sensors with Simulated Avatars. In *SIGGRAPH Asia 2022 Conference Papers.* ACM. https://doi.org/10.1145/3550469.3555411

Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2020. GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 6183–6192. https://doi.org/10.1109/CVPR42600.2020.00622