ZStudio: Portable and Real-time Motion Capture Studio for Creators in the Metaverse

Jung-Seok Cho
jungseok.cho@naverz-corp.com
Avatar AI, NAVER Z
South Korea

Seongchan Jeong
seongchan_jeong@naverz-corp.com
Avatar AI, NAVER Z
South Korea

Geonwon Lee
geonwon.lee@naverz-corp.com
Avatar AI, NAVER Z
South Korea

Jaehyun Han
jaehyun.han@naverz-corp.com
Avatar AI, NAVER Z
South Korea

HyeRin Yoo
hyerin.yoo@naverz-corp.com
Avatar AI, NAVER Z
South Korea

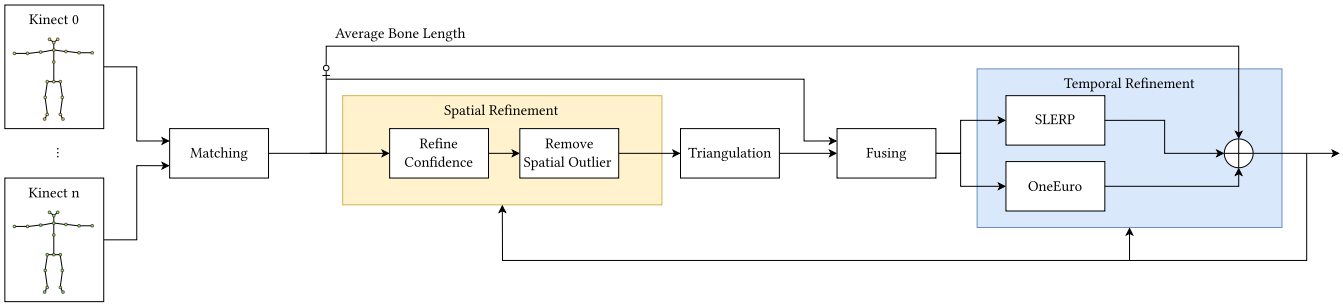


Figure 1: Overall pipeline of motion capture. Skeletons are assigned to the tracing IDs. Spatial outliers are corrected in the first stage. The triangulated and original skeleton are fused based on confidence. Lastly, the final stage removes temporal outliers.

ABSTRACT

In the metaverse, the avatar motion plays a pivotal role in users' communication who aspire to express themselves. Creators are responsible for motion generation using motion capture technology. However, most motion capture studios have clear economic and technical limitations for creators working from home. To overcome these obstacles, this research proposes a portable and real-time motion capture system(ZStudio) with commercial cameras. Furthermore, it also puts forth an algorithmic pipeline that dramatically enhances accuracy with a limited number of cameras.

CCS CONCEPTS

• **Computing methodologies** → **Motion capture**; *Motion processing*.

KEYWORDS

motion capture, mobile capture studio, avatar, computer vision

ACM Reference Format:
Jung-Seok Cho, Seongchan Jeong, Geonwon Lee, Jaehyun Han, and HyeRin Yoo. 2023. ZStudio: Portable and Real-time Motion Capture Studio for Creators in the Metaverse. In *ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG '23), Nov 15–17, 2023, Rennes, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

MIG '23, Nov 15–17, 2023, Rennes, France

ACM ISBN 978-1-4503-XXXX-X/18/06.
https://doi.org/XXXXXXX.XXXXXXX

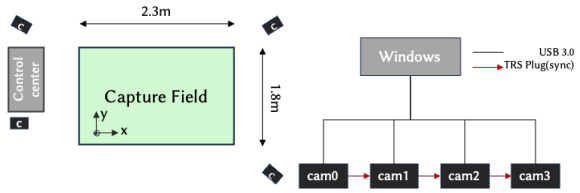
1 INTRODUCTION

Motion is a key form of non-verbal communication in both real-life and virtual world [2], where creators serve as crucial content providers in enabling users to express individuality through avatars [5]. To expedite creators, service providers have researched motion capture from two perspectives: mobilization and advancement, meaning mobile-device-level usage and micro-level accuracy respectively.

However, these two perspectives have distinct challenges. Firstly, mobilizable pose estimation methods are susceptible to occlusion and lead to depth ambiguity[6]; it impedes utilization of service. Secondly, high-performance solutions are not real-time and above all, pose physical and economic limitations for creators. In this paper proposes a motion capture system design with hardware design suitable for installation at home and motion estimation algorithms.

In summary, the contributions of this paper include:

- **Robust motion capture algorithmic pipeline:** We propose a temporal-spatial robust motion capture pipeline; it is resistant to common challenges of motion capture: left-right reversed, partially framed out, occluded joints.
- **Portable, yet scalable and realtime system:** ZStudio requires a minimum of two Azure Kinect units and an easy-installable interface with 30FPS. The system can be expanded according to creators' needs for more performance.



(a) Layout (b) Interface

Figure 2: System configuration diagram of ZStudio.

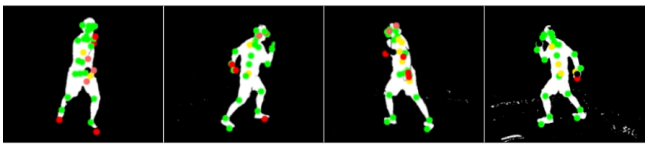


Figure 3: Joints and estimated confidence values.

As related studies, Panoptic[4] proposed a "social Interaction dataset" using 480 VGA, 31 HD, and 10 Kinect cameras. Recently, simple capture studio[1], fewer than 10 cameras, contributes to the improved accuracy of 2D estimation.

2 PROPOSED METHOD

Figure 1 illustrates the proposed system that resolves spatial and temporal outliers, which are caused due to invisible joints and self-occlusions. The visual information generated by each perspective interacts with themselves to alleviate outliers.

2.1 Hardware Configuration

Figure 2. depicts hardware configuration. The capture field is defined as the union of the field of view (FoV) of the four cameras. The capture field can be freely modified within the distance limit of Kinect in any place, and we set the camera positions empirically. Kinects connect to the desktop via USB 3.0, and a TRG cable is required for temporal synchronization. Cameras are calibrated with pre-marked patterns[3].

2.2 Pose Refinements

Pose estimation causes outliers for a variety of reasons such as motion blur. The primary idea is the confidence of each joint represents view reliability in triangulation[3]. In the first step, we regenerate spatial joint confidence. In detail, we create integrated depth maps from four Kinects and compute the joint confidence based on the obtained 3D pose. Furthermore, the similarity is measured based on the angle between the bone vector in the current and the previous frame to remove ill-pose such as a flipped and mislay estimated skeleton.

Temporal interference occurs due to physical reasons such as the rolling shutter, sensor noise, illumination, etc. We separately mitigate local and global temporal outliers with well-known processing algorithms. We refine the bone vectors locally with the SLERP algorithm. OneEuro filter purifies root position(pelvis) from temporal outliers for global coordinates.

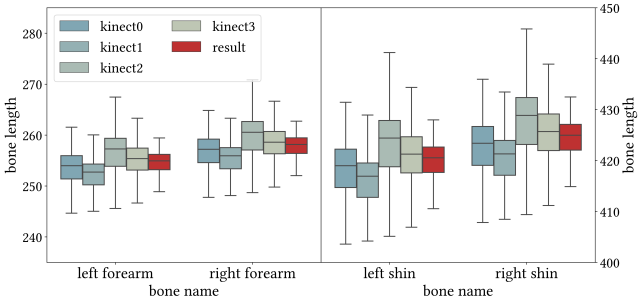


Figure 4: Box-whisker plot of bone vectors.

3 EXPERIMENT

Figure 3 illustrates that projected point clouds and apparently visible and occluded joints have high confidence (green) and low confidence (red) respectively—confidence alleviates the frequency of ill-pose occurrences in triangulation. To validate the plausibility of the result poses, we measured the variance of each length of the most dynamic: forearm and shin in Figure 4. The variance of the bone lengths refined by the pipeline was 64% less than the average of the variances of the other views and 35% less than the lowest variance view.

Additional experiments confirm that the number of views has a significant impact on achieving higher plausibility. Assuming the variance of the results obtained with four cameras was 1, variances of two and three camera systems were 1.28 and 1.15 respectively in the identical motion sequence. We confirmed our pipeline runs on 30FPS+, maximum FPS of Kinect camera, with Unity Engine on i5-13600KF, RTX 4080, and 32GB RAM.

4 CONCLUSION

In this work, we proposed the portable and real-time motion studio for creators with multiple cameras for accurate 3D pose and presented an algorithmic pipeline to solve various problems resulting from inaccurate pose estimation. In the future, we plan to present how to quantify and measure the plausibility of the motion capture system and re-targeting to avatars.

REFERENCES

[1] Renat Bashirov, Anastasia Ianina, Karim Iskakov, Yevgeniy Kononenko, Valeriya Strizhkova, Victor Lempitsky, and Alexander Vakhitov. 2021. Real-time rgbd-based extended body pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2807–2816.

[2] Ray L. Birdwhistell. 1971. *Kinesics and Context*. University of Pennsylvania Press, Philadelphia. https://doi.org/doi:10.9783/9780812201284

[3] Richard Hartley and Andrew Zisserman. 2004. *Multiple View Geometry in Computer Vision* (2 ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511811685

[4] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2015. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*. 3334–3342.

[5] Marc Erich Latoschik, Daniel Roth, Dominik Gall, Jascha Achenbach, Thomas Waltemate, and Mario Botsch. 2017. The Effect of Avatar Realism in Immersive Social Virtual Realities. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology* (Gothenburg, Sweden) *(VRST '17)*. ACM, New York, NY, USA, Article 39, 10 pages. https://doi.org/10.1145/3139131.3139156

[6] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. 2020. Deep learning-based human pose estimation: A survey. *Comput. Surveys* (2020).