MediaPipe-based Real-time Interactive Avatar Generation for Metaverse

Esmot Ara Tuli[1], Ahmad Zainudin[2], Md Javed Ahmed Shanto[1], Jae Min Lee[1], and Dong-Seong Kim[1]
[1]Department of IT convergence Engineering,Kumoh National Institute of Technology, Gumi, South Korea, 3917
[2]Department of Electronic Engineering, Kumoh National Institute of Technology, Gumi, South Korea, 3917
{esmot, zai, shantoa729, ljmpaul, dskim}@kumoh.ac.kr

***Abstract*—The Metaverse constitutes a collective virtual world platform that integrates a range of emerging technologies, including artificial intelligence, web 3.0, blockchain, advanced hardware, and other cutting-edge innovations. In the metaverse, the metahuman is considered a vital part. The demand for concurrent interaction, motion, and emotion generation in metaverse avatars, mirroring real-world human counterparts, is witnessing a notable upsurge within academic and industrial domains. In order to address this concern, the present study proposes a real-time motion and expression generation approach integrated with MediaPipe for the Metaverse. Furthermore, we deploy the Mediapipe integrated metahuman creation process in the Unreal engine, which shows the performance is better than the other existing methods.**

***Index Terms*—Human-centered computing, human-avatar interaction, motion capture, metaverse.**

I. Introduction

The Metaverse represents a virtual universe that emulates real-world activities, encompassing a multitude of diverse engagements and interactions. The meta-human constitutes a vital component within the Metaverse, and its significance has been particularly amplified during the COVID-19 pandemic. There has been a sudden surge in demand for a human-centric Metaverse experience, driven by the unique circumstances and challenges posed by the pandemic [1]. In the metaverse, meta-humans can participate in a variety of diverse activities, such as classroom, training, conferences, concerts, entertainment, travel, shopping, business, social gatherings, etc [2]. In order to facilitate these functionalities, users must be empowered to transmit real-time motion information within the metaverse environment. Consequently, it is imperative that meta-human consistently align with the motion of the respective users they represent.

In [3], the authors propose a Wi-Fi-based smart home system for human pose estimation in metaverse avatar simulation. They utilize WiFi channel state information (CSI) to classify human pose and activity, as well as generate avatars based on the data obtained from WiFi sensing data. The collection of CSI sensing data suffers from phase shift problems, which negatively impact the extraction of information such as behavior and heartbeat, consequently leading to a decline in the performance of avatar construction. Therefore, paper [4] correct CSI phase shifts using phase compensation and sliding window. Furthermore, they propose a wireless sensing dataset for the metaverse. Instead of depending single device, wang et al.[5] propose semantic sensing information transmission for the metaverse. Combine multiple sensing data for example mobile phones, wifi, and others encode first using semantic encoding before sending to the metaverse to reduce information loss. However, the sensor-based metahuman creation process has some major limitations. For instance, facial details expression is difficult to create in the metaverse, also it is difficult to identify gender using sensor data. In paper [6], proposes real-time virtual humans in the metaverse as well as generating six facial expressions by setting parameters. This process is not dynamic, avatars only use predefined expressions. In [7], adopt Octree-based real-time metahuman simulation in the metaverse. However, the performance of synchronization and delay needs to improve.

Our work uses an open-source deep learning-based platform, Mediapipe which is developed by Google, to generate real-time human poses, movements, and facial expressions in the metaverse.The remainder of the research is organized as follows: Section II illustrates the proposed system. Implementation description described in section III. Finally, section IV concludes the paper with the future direction.

II. Proposed System

The proposed system model is illustrated in Figure 1. As we can see, the user's real-time video is captured using a webcam or pre-recorded videos can be used. The captured video then passes through the Mediapipe plugin, which estimates landmarks. After that, in the metaverse, the metahuman avatar will be connected with the Mediapipe plugin using an animation blueprint. Where avatar movement and facial expression will reflect the real-time input provided by the Media-pipe plugin.

We employ the MediaPipe Holistic model which is a combination of two Mediapipe models: pose detection and face detection. Although the pose detection model also includes face detection, it does not capture facial expressions in detail.By using the Holistic model, we can generate an avatar that exactly mimics real-human expressions in real-time. The Holostick landmark utilizes a machine model to facilitate the continuous generation of gestures, poses, and actions. It relies on 468 face landmarks, 33 pose landmarks, and 21 hand landmarks for each hand.

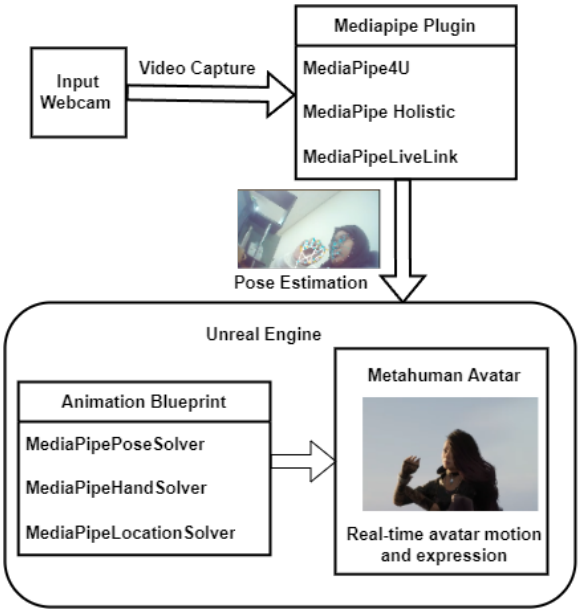

Fig. 1: Instantaneous synchronization of avatar motion with real-world physical movements.

Fig. 2: Full body motion generation and facial expression mimicking.

III. Implementation Description

In the creation of the metaverse, we utilize the Unreal Engine platform. However, it is worth noting that Mediapipe can also be seamlessly connected with other platforms, such as Unity 3D. We utilize Unreal Engine 5.1 on a Windows 10 operating system with 64-bit architecture, equipped with 16 GB RAM, and a Core i5-8500 processor. Additionally, we employ Visual Studio 2022 and Visual C++ as our development tools. In order to use Mediapipe in the Unreal Engine, it is necessary to enable specific plugins, including MediaPipe4ULiveLink, MediaPipe4UGStreamer, GStreamer, and MediaPipe4U. In our study, we employ the Unreal Metahuman character. In the case of Maximo 3D character, it is necessary to connect the skeleton bones according to the Mediapipe bone hierocracy.

IV. Conclusion

We apply Mediapipe for real-time avatar creation in the metaverse. We can utilize webcam or recorded video in this framework. This vision-based model will be suitable for meetings, conferences, and classrooms inside the metaverse. In future work, there is an opportunity to utilize additional MediaPipe features such as audio classification, text embedding, and language detection for the metaverse.

Acknowledgment

This work was supported by Priority Research Centers Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(MEST)(2018R1A6A1A03024003) and MSIT(Ministry of Science and ICT), under the Innovative Human Resource Development for Local Intellectualization support program (IITP-2023-2020-0-01612) supervised by the IITP(Institute for Information communications Technology Planning Evaluation).

References

[1] U. Zaman, I. Koo, S. Abbasi, S. H. Raza, and M. G. Qureshi, "Meet your digital twin in space? profiling international expatrsquo;s readiness for metaverse space travel, tech-savviness, covid-19 travel anxiety, and travel fear of missing out," *Sustainability*, vol. 14, no. 11, 2022. [Online]. Available: https://www.mdpi.com/2071-1050/14/11/6441

[2] T. Huynh-The, Q.-V. Pham, X.-Q. Pham, T. T. Nguyen, Z. Han, and D.-S. Kim, "Artificial intelligence for the metaverse: A survey," *Engineering Applications of Artificial Intelligence*, vol. 117, p. 105581, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197622005711

[3] J. Yang, Y. Zhou, H. Huang, H. Zou, and L. Xie, "Metafi: Device-free pose estimation via commodity wifi for metaverse avatar simulation," *arXiv preprint arXiv:2208.10414*, 2022.

[4] J. Wang, H. Du, X. Yang, D. Niyato, J. Kang, and S. Mao, "Wireless sensing data collection and processing for metaverse avatar construction," *arXiv preprint arXiv:2211.12720*, 2022.

[5] J. Wang, H. Du, Z. Tian, D. Niyato, J. Kang, and X. Shen, "Semantic-aware sensing information transmission for metaverse: A contest theoretic approach," *IEEE Transactions on Wireless Communications*, 2023.

[6] M. Zhang, Y. Wang, J. Zhou, and Z. Pan, "Simuman: A simultaneous real-time method for representing motions and emotions of virtual human in metaverse," in *Internet of Things–ICIOT 2021: 6th International Conference, Held as Part of the Services Conference Federation, SCF 2021, Virtual Event, December 10–14, 2021, Proceedings*. Springer, 2022, pp. 77–89.

[7] K. Y. Lam, L. Yang, A. Alhilal, L.-H. Lee, G. Tyson, and P. Hui, "Human-avatar interaction in metaverse: Framework for full-body interaction," in *Proceedings of the 4th ACM International Conference on Multimedia in Asia*, 2022, pp. 1–7.