Digital Human Interaction Based on Mono Camera for Digital Twin

Myeongseop Kim†
Contents Convergence Research Center
Korea Electronics Technology Institute
Seoul, Korea
myeongseopkim@keti.re.kr

Taehyeon Kim
Contents Convergence Research Center
Korea Electronics Technology Institute
Seoul, Korea
taehyeon.kim@keti.re.kr

Kyung-Taek Lee
Contents Convergence Research Center
Korea Electronics Technology Institute
Seoul, Korea
ktechlee@keti.re.kr

***Abstract*—In virtual reality, the importance lies not only in digitally replicating real spaces, assets, forms, and actions but also in digitalizing the movements of individuals within these spaces from the perspective of digital twin technology. However, traditional methods, which require the use of Head Mounted Displays (HMD), VR devices, and various sensors attached to humans, impose significant constraints when reflecting the diverse actions of users within the actual digital twin. To overcome this, we introduce the "Digital Human Interaction based on Mono Camera" method. This approach simulates full-body human actions using an easily accessible mono camera. Through this method, we've identified the potential to broaden the range of user movements that can be mirrored in virtual reality while significantly reducing costs. Crucially, our approach is future-oriented and universally applicable because it does not require specialized equipment or sensors.**

***Index Terms*—Virtual Reality, Digital Twin Technology, Digital Human Interaction, 3D Pose Estimation, Avatar Construction**

I. Introduction

Digital Twin (DT) seamlessly integrates cyber and physical spaces, offering a realistic simulation and monitoring experience for product design, production, prediction, and anomaly detection in advanced manufacturing, PLM (Product Lifecycle Management), and smart healthcare sectors [1]. For a digital twin to accurately reflect the real world, it must encompass not only entities such as digital devices and spaces but also human entities that interact with other digital twins within that space [2], [3]. To digitalize these human entities, we need digital human interaction technology that mirrors real human movements onto virtual humans in real-time. Recent studies have introduced various methods to project human movements onto virtual avatars. Some of these methods use KinectV2 and Basler sensors to extract user joints, implementing virtual animation tracking through mathematical calibration [4]. Others employ real-time depth cameras and HMD, like the Oculus Quest, to execute full-body skeleton tracking provided by Microsoft Rocketbox avatars [5]. While the majority of research utilizes Kinetic sensors and HMD to capture real-time body movements [6], some studies have mirrored real human movements in the cyber world using devices attached to humans, such as the AR arm [7] and Hi5 VR gloves [8]. However, methods relying on depth cameras or wearable devices encounter limitations in their broad applicability. Our innovative digital human interaction method, as illustrated in Fig 1, employs a universally applicable mono camera to create an avatar that replicates a person's full-body movements, eliminating the need for specific devices or additional sensors.

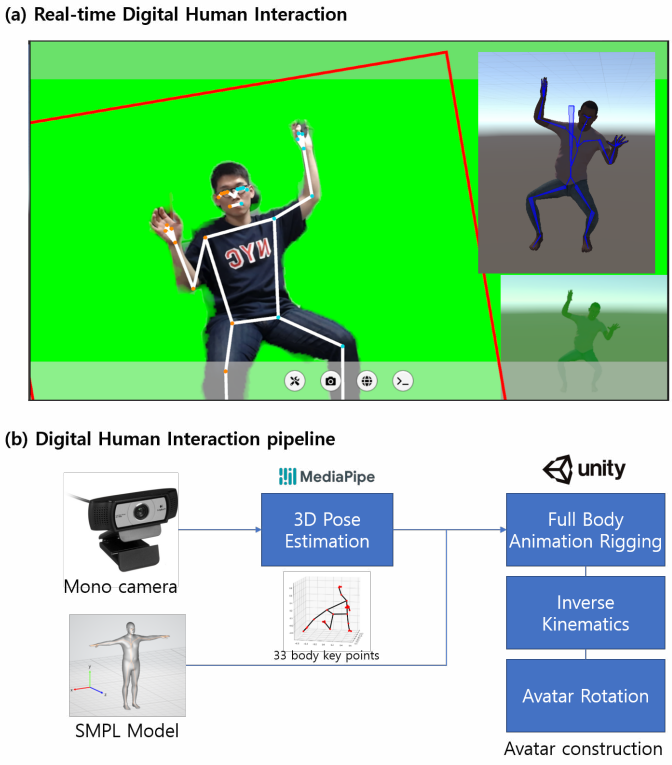

Fig. 1. (a) Real-time digital human interaction, (b) Digital human interaction pipeline

Myeongseop Kim is the corresponding author and also first author.
This work was supported by the Technology Innovation Program (20018295, Meta-human: a virtual cooperation platform for a specialized industrial services) funded By the Ministry of Trade, Industry Energy (MOTIE, Korea).

Additionally, we've incorporated a Mediapipe Pose 3D pose estimation model into the system, enabling precise posture estimation using just the mono camera. This estimated skeleton data is harnessed to rig the movements of the digital human, leveraging inverse kinematics. We believe that this strategy is apt for future-focused digital human interaction and can seamlessly integrate into digital twin scenarios.

II. Digital Human Interaction Based on Mono Camera Method

As illustrated in Fig 1, we capture real-time video images using a mono camera and extract key points of the human body using 3D pose estimation technology. Utilizing this extracted data, we construct avatars in Unity, a leading virtual reality environment, to facilitate digital human interaction [9]. For avatar construction, we employ the Skinned Multi-Person Linear (SMPL) [10], a realistic 3D human body model that is parameterized based on the body's shape and pose. To implement the digital human interaction system, we rig the animation of the digital human avatar using the extracted 3D pose estimation data, leveraging inverse kinematics [11] from computer graphics and applying forward rotation correction.

A. Apparatus and Environments

The mono camera we used for digital human interaction is highly versatile. It encompasses models such as the Logitech C930, Logitech 4K Pro Magnetic, and even built-in webcams on laptops, all without specific constraints. We employ a single camera to capture the input video for the 3D pose estimation model. Our system operates within an environment powered by the Unity 3D engine, which is used to realize the digital human. This system was implemented using Unity version 2021.2.3f1.

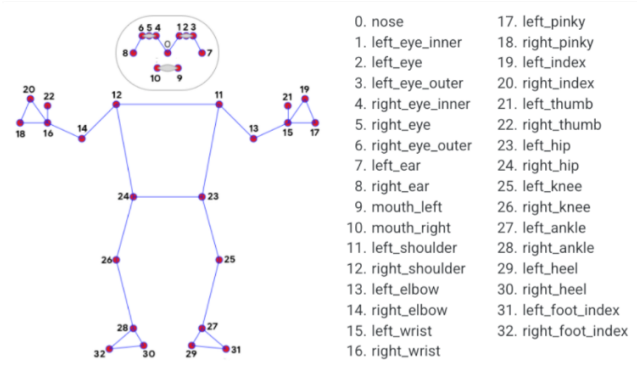

Fig. 2. 33 Pose Landmarks [13]

B. 3D Pose Estimation

With advancements in computer vision technology, markerless gait analysis based on video footage has become feasible using pose estimation models such as OpenPose [12], BlazePose [13], and YOLO-Pose [14]. These pose estimation techniques employ computer vision and machine learning algorithms to extract human poses from videos and track the movement of the body's joints and limbs in 2D or 3D spaces. Notably, Mediapipe Pose, an open-source tool developed by Google and built upon the BlazePose model, offers faster inference speeds and more reliable results than other models [15]. As a result, we developed a Pose detector integrated with Mediapipe Pose to execute 3D pose estimation from real-time video captured by a Mono Camera. Leveraging this technology, we can real-time estimate the pose of a person's body, face, and hands using 33 key points (as shown in Fig 2). We utilize specific data from these estimated key points to rig the avatar's animation.

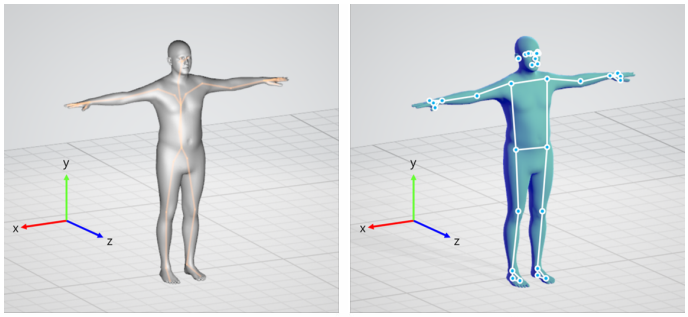

Fig. 3. (a) SMPL body key points, (b) Our body key points

C. Avatar Construction

For avatar construction, we utilize the SMPL model, which is widely accepted for representing realistic human body poses and is extensively used for 3D pose estimation of humans in images and videos [16]. We synchronized the SMPL model with real-time user movements using inverse kinematics (IK), a crucial technique in animation, by leveraging specific joint coordinates. In Unity 3D, we first repositioned the results of the 3D Pose Estimation onto the avatar object for animation rigging. We then executed avatar animation rigging using the IK tool and a Unity Asset [17] designed for human IK control. However, as the pre-defined body keypoints of the SMPL model (as shown in Fig 3 (a)) and the results from our 3D pose extraction (Fig 3 (b)) are not identically structured, we couldn't match them one-to-one. Thus, we individually implemented IK rigging for the arms and legs, which could be matched. Since IK dynamically implements joint movements based on the endpoint of a fixed object, animation issues arise when rigging the body from non-frontal views. Therefore, we projected the 3D Pose estimation results onto the XZ plane and calculated a vector orthogonal to the extension line connecting both hip coordinates, which serve as the body's center. By rotating the avatar object based on this, we meticulously adjusted the digital human avatar.

III. Results and Discussion

In the realm of virtual reality, ensuring that digital representations of human beings are accurate and dynamic is crucial. Through our study, we have taken significant strides

toward this objective by leveraging widely accessible mono cameras. Here, we discuss the implications of our findings, their advantages, and potential areas for future refinement.

Utility of Mono Cameras: Traditional digital human interaction methods, reliant on depth cameras and wearable devices, often face limitations, especially in broader applicability. Our innovative approach using a universally applicable mono camera represents a potential game-changer. By eliminating the need for specialized equipment or sensors, our method becomes financially feasible and universally applicable.

Performance of the 3D Pose Estimation: Integrating the Mediapipe Pose model for 3D pose estimation has proven to be a strong decision. This model, developed by Google, has provided faster and more consistent results than other existing models. Given that our approach solely depends on the accuracy of pose estimation for avatar construction, this choice reinforces the reliability of our methodology.

Efficiency in Avatar Construction: Utilizing the SMPL model coupled with inverse kinematics for avatar construction within Unity 3D was pivotal. This combination allowed for an accurate and dynamic replication of real human movements. However, some challenges arose due to the misalignment between the predefined keypoints of the SMPL model and our 3D pose extraction results. In the future, refining the process to bridge this gap more efficiently could further enhance the realism of the digital avatars.

Implications for Digital Twin Technology: Our method, rooted in the "Digital Human Interaction based on Mono Camera" approach, presents significant implications for Digital Twin Technology. Given that the core of Digital Twin is to integrate cyber and physical spaces seamlessly, our research effectively addresses the human element, which has often been a complex component to digitalize accurately. Now, with a simple mono camera setup, the range of human movements that can be mirrored in virtual reality has widened substantially.

IV. Future works

While our approach is groundbreaking, it is not without limitations. For instance, as the full accuracy of avatar movements in non-frontal views remains a challenge, future research can explore methodologies to overcome such constraints. Moreover, further optimization of the system for even more diverse camera models can ensure the ubiquitous application of this technology.

V. Conclusion

In conclusion, the "Digital Human Interaction based on Mono Camera" method showcases significant promise in revolutionizing the way we approach digital human representation in virtual spaces. By simplifying the apparatus while maintaining or even enhancing accuracy, our research paves the way for the more accessible and widespread adoption of virtual reality and digital twin applications.

Acknowledgment

This work was supported by the Technology Innovation Program (20018295, Meta-human: a virtual cooperation platform for a specialized industrial services) funded By the Ministry of Trade, Industry Energy (MOTIE, Korea).

References

[1] F. Tao, H. Zhang, A. Liu, and A. Y. Nee, "Digital twin in industry: State-of-the-art," IEEE Transactions on industrial informatics, 15(4), 2018, pp. 2405-2415.

[2] W. Shengli, "Is human digital twin possible?," Computer methods and programs in biomedicine update, 1, 100014, 2021.

[3] M. Dallel, V. Havard, Y. Dupuis, and D. Baudry, "Digital twin of an industrial workstation: A novel method of an auto-labeled data generator using virtual reality for human action recognition in the context of human–robot collaboration," Engineering Applications of Artificial Intelligence, 118, 105655, 2023.

[4] P. Fechteler, W. Paier, A. Hilsmann, and P. Eisert, "Real-time avatar animation with dynamic face texturing," In 2016 IEEE International Conference on Image Processing (ICIP), pp. 355-359, September 2016.

[5] M. Gonzalez-Franco, Z. Egan, M. Peachey, A. Antley, T. Randhavane, P. Panda, Y. Zhang, C. Y. Wang, D. F. Reilly, T. C. Peck, A. S. Won, A. Steed and E. Ofek, "Movebox: Democratizing mocap for the microsoft rocketbox avatar library," In 2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), pp. 91-98, December 2020.

[6] J. Zhao, Z. Wang, Y. Peng, and Y. Wang, "Real-time generation of leg animation for walking-in-place techniques," In Proceedings of the 18th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry, pp. 1-8, December 2022.

[7] A. Boschmann, D. Neuhaus, S. Vogt, C. Kaltschmidt, M. Platzner, and S. Dosen, "Immersive augmented reality system for the training of pattern classification control with a myoelectric prosthesis," Journal of neuroengineering and rehabilitation, 18(1), 1-15, 2021.

[8] S. Kang, G. Kim, K. T. Lee, and S. Kim, "Giant finger: Visuo-proprioceptive congruent virtual legs for flying actions in virtual reality," In 2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), pp. 933-934, March 2023.

[9] V. Voleti, B. Oreshkin, F. Bocquelet, F. Harvey, L. S. Ménard, and C. Pal, "Smpl-ik: Learned morphology-aware inverse kinematics for ai driven artistic workflows," In SIGGRAPH Asia 2022 Technical Communications, pp. 1-7.

[10] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," ACM transactions on graphics (TOG), 34(6), 1-16, 2015.

[11] A. Aristidou, J. Lasenby, Y. Chrysanthou, and A. Shamir, "Inverse kinematics techniques in computer graphics: A survey," In Computer graphics forum, Vol. 37, No. 6, pp. 35-58, September 2018.

[12] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7291-7299, 2017.

[13] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking", arXiv preprint arXiv:2006.10204, 2020.

[14] D. Maji, S. Nagori, M. Mathew, and D. Poddar, "Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2637-2646, 2022.

[15] C. S. T. Hii, K. B. Gan, N. Zainal, N. M. Ibrahim, S. A. M. Rani, and N. Abd Shattar, "Marker free gait analysis using pose estimation model," In 2022 IEEE 20th Student Conference on Research and Development (SCOReD), pp. 109-113, IEEE, November 2022.

[16] Y. Sun, Q. Bao, W. Liu, Y. Fu, M. J. Black, and T. Mei, "Monocular, one-stage, regression of multiple 3d people," In Proceedings of the IEEE/CVF international conference on computer vision, pp. 11179-11188, 2021.

[17] RootMotion, Final IK. Unity Technologies. https://assetstore.unity.com/packages/tools/animation/final-ik-14290 (Accessed: 2023-07-13), 2022.