Real-Time Human Synchronization Framework for Digital Twin

Donghoon Lee, Joongho Cho and Jaeho Kim

July 10, 2023

Real-time Human Synchronization Framework for Digital Twin

Donghoon Lee
Department of Information
and Communications Engineering
Sejong University
Seoul, Republic of Korea
donghoon.sejong@gmail.com

Joongho Cho
Department of Information
and Communications Engineering
Sejong University
Seoul, Republic of Korea
whwndgh1234@gmail.com

Jaeho Kim*
Department of Information
and Communications Engineering
Sejong University
Seoul, Republic of Korea
kimjh@sejong.ac.kr

***Abstract*—The concept of digital twin, which synchronizes data from physical objects with the virtual world, has been widely used in various fields. However, previous studies have focused on synchronizing robots due to their built-in sensors, while synchronizing humans is more complex due to their high degree of freedom and the need for external sensors. Although studies have been conducted to synchronize humans using cameras, they have only considered a single space. To apply this technology in real buildings or factories, a large-scale digital twin using multiple sensors is required. To address this issue, we propose a framework that synchronizes large-scale digital twins using only the features of people in images, rather than raw data from RGB camera sensors, reducing network traffic. This framework allows for the storage and replay of only the synchronized human's features, facilitating interaction with other robots and customization.**

***Index Terms*—digital twin, synchronization, IoT**

I. Introduction

Digital twin is a technology that converts data collected from objects in the physical world and synchronizes them with the virtual world. With the development of Internet of Things (IoT) and deep learning, the digital twin is widely used in various fields in the past few years. Due to the advantages that digital twin has the potential to improve efficiencies and reduce costs in many industries, the building of the large scale digital twin attracts much attention.

While previous studies on digital twin technology have primarily focused on robot synchronization, there is a need for the synchronization of human digital twins in environments where humans and robots coexist. Unlike robots, humans require external sensors to determine their location and have a higher degree of freedom (DoF), making it more challenging to represent them in virtual space. Although some studies have been conducted on the synchronization of humans, they have only considered a single space and used cameras to detect and synchronize humans. To apply digital twin technology in real-world scenarios such as buildings or factories, it is necessary to build large-scale digital twins using multiple sensors. However, the high communication costs associated with synchronizing data collected from multiple sensors can be a significant barrier.

To address this issue, we propose a framework that synchronizes large-scale digital twins by utilizing only the features of humans in images, rather than the raw data obtained through RGB camera sensors. Our proposed framework stores and replays only the features of the synchronized human to reduce network traffic while utilizing the advantages of digital twins, which are not limited by time and space. Additionally, to facilitate interaction with robots and customization, we represent humans in the form of a humanoid.

In this paper, we present the details of our proposed framework and discuss its effectiveness in synchronizing large-scale digital twins.

II. Related Work

Several studies have focused on implementing human motion capture technology using various sensors to track and record the movements of humans.

Yi et al. (2022) [1] proposed a method that utilized a small number of inertial sensors to estimate joint torques and capture human motion in real-time. The approach demonstrated fast and accurate tracking ability and improved temporal stability and physical accuracy compared to existing human body motion tracking technology. However, since the method requires attaching separate inertial sensors to the body for human synchronization, it may be challenging to apply it to the synchronization of multiple people in real-life scenarios.

There are numerous human motion capture technologies that use RGB cameras for synchronization. Mehta et al. (2017) [2] proposed Vnect, a real-time synchronization technology that can estimate the 3D posture of the human body using only a single RGB camera. The method has rapid speed and high accuracy, and can estimate human body posture using only a single RGB camera. Mehta et al. (2019) [3] utilized computer vision technology to detect individuals in a scene and then applied the Vnect model to each person to estimate 3D postures for multiple people using a single RGB camera.

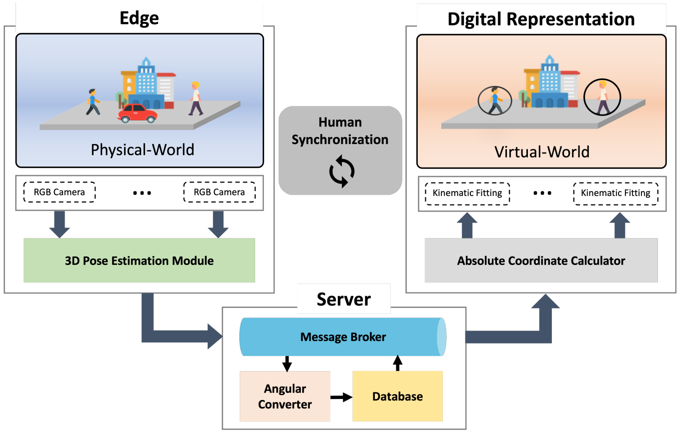


Fig. 1: Overall architecture of the framework

While these methods are capable of implementing a high-performance motion capture model of the human body using only a single camera, they only deal with synchronization in a limited area. As a result, the application of these methods to large-scale digital twin synchronization may be challenging, and alternative approaches are needed to address this issue.

In this paper, we present a framework that facilitates the creation of a large-scale digital twin by synchronizing multiple individuals in the physical world with their corresponding virtual representations using a human motion capture model. Our proposed framework utilizes multiple RGB cameras that are placed in different separate spaces to capture human movements in the physical world and synchronize them with their virtual counterparts in the digital twin. By applying this framework, we aim to overcome the challenges of synchronizing human movements in a complex environment where humans and robots coexist. Moreover, our proposed framework enables the creation of a customizable and interactive digital twin, which has the potential to improve efficiencies and reduce costs in various industries.

III. Human Synchronize Framework for Digital Twin

In this paper, we propose a framework for synchronizing human motion to large-scale digital twins using multiple RGB cameras installed in the physical world. The framework consists of three parts: the edge side, server side, and digital representation side. The edge side uses deep learning to track a person and estimate their 3D pose, which is then transmitted to the server. The server converts the 3D pose data into a format for synchronization with the human model in the virtual world using an angular converter and stores the data in a database. Finally, the digital representation side constitutes a virtual space expressed in a mesh format and synchronizes humans using the data from the database in the server side with a kinematic calculator.

IV. Implementation of the Framework

A. Edge Side

3D Pose Estimation Module

We utilized Yolov8 [4] to detect the location and id of a person from the images collected by the camera. The distance between the person and the camera was estimated through RootNet(2019) [5]. The estimated position and distance of the person were then input into PoseNet(2019) [5] to estimate the pose for the person's joints. The estimated 3D pose values were transmitted to the server.

B. Server Side

We utilized Kafka as a message broker for real-time data streaming, and the topic name was defined according to the camera ID. We built a PostgreSQL database and stored the pose and timestamp information together to facilitate replay later.

To extract the 3D pose of a person from the edge data and apply it to a digital twin humanoid model, we developed an angular converter. The converter converts the 3D pose values into angle values using the following equation:

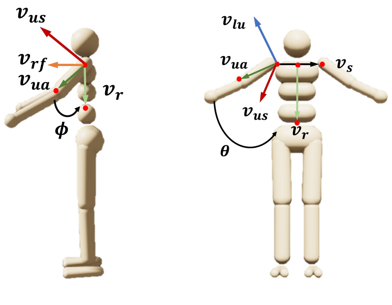
$$\phi = \begin{cases} cos^{-1}((v_{us} \cdot v_{rf})/|v_{us}||v_{rf}|) & if\ v_r \cdot v_{us} \geq 0 \\ -1 \times cos^{-1}((v_{us} \cdot v_{rf})/|v_{us}||v_{rf}|) & if\ v_r \cdot v_{us} < 0 \end{cases}, \quad (1)$$

where v_r denotes the root vector going down the center point, v_{ua} denotes a vector from upper to arm, v_s denotes a vector from the shoulder, v_{us} denotes the cross product of v_{ua} and v_s, and v_{rf} denotes the cross product of the root vector v_r and v_{us}.

Similarly, in Fig. 2a, the angle θ at which the arm moves left and right. It can be calculated by equation :

$$\theta = \begin{cases} cos^{-1}((v_{lu} \cdot v_s)/|v_{lu}||v_s|) & if\ v_r \cdot v_{lu} \geq 0 \\ -1 \times cos^{-1}((v_{lu} \cdot v_s)/|v_{lu}||v_s|) & if\ v_r \cdot v_{lu} < 0 \end{cases}, \quad (2)$$

where v_{lu} denotes the cross product of v_{us} and v_{ua}.



(a) Side view (b) Front view

Fig. 2: A simulated humanoid robot model provided by Mojoco

C. Digital Representation Side

Rendering Tool We employed Nvidia's Omniverse Isaac Sim as a rendering tool for creating the virtual world. This tool is equipped with a physics engine and can interact with robots.

Absolute Coordinate Calculator To synchronize the relative coordinate data collected from the camera with absolute coordinates, we applied an absolute coordinate calculator. This ensured that the virtual representation of the human's motion was accurate and consistent with their physical movements in the real world.

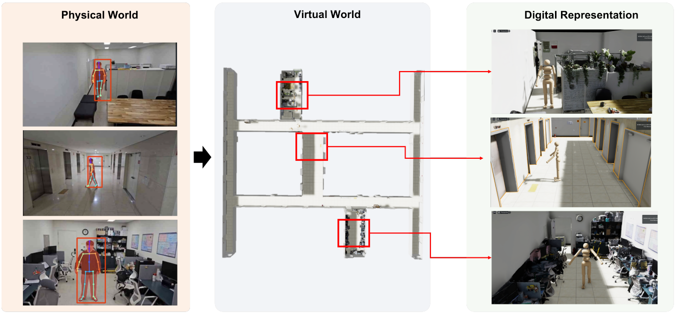


Fig. 3: Result of real-time synchronization of human motion in the scene from the physical world to the virtual world. We fixed cameras in each room and corridor, synchronizing the motion of one person in each room.

V. Experiment and Result

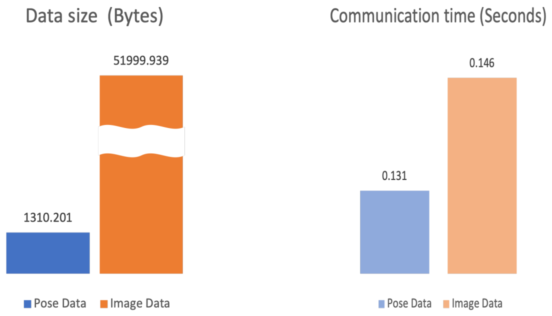
A. Experiment Setup

To evaluate the effectiveness of our proposed framework, we set up an experimental environment consisting of three Edge computers connected to three 2D RGB cameras, one Kafka server, one DB server, and one workstation to run Omniverse Isaac Sim for virtual environment rendering. To construct the virtual environment, we scanned and meshed the 5th floor of the Deayang AI center building in Sejong University, Seoul, South Korea using 3D LiDAR. In the physical world, we installed one camera in each of the two rooms on the 5th floor of the Deayang AI center building and one camera in the hallway to perform synchronization.

B. Results

In this study, we evaluated the proposed framework by synchronizing people in two rooms and a corridor, as depicted in Fig.4. We also compared the capacity and transmission speed when data was extracted from an image and when an image was transmitted to a server, assuming the same network situation. To compare the network cost of sending each data to the server, we assumed that the processing time of pose estimation was the same since the GPU used in the edge and the server was the same. The image was compressed at a rate of 50% in JPEG format and transmitted, while the extracted features were transmitted as json without compression. The average of 1000 transmissions was calculated, and the results are presented in Fig.3.

The results in Fig.4a show that when only human features are extracted from the image and transmitted to the server, the data size is reduced by about 97.5% compared to the size of the image data. This is assuming that there is only one person, and if the number of people increases, the size of the extracted data will increase accordingly. However, since there is a large difference, it is efficient to estimate multiple people. Moreover, the transmission of only human pose to the server showed about 10.2% faster results, as shown in Fig.4b, which is not proportional to the size of the data. However, considering the size of the data, the difference in speed will be wider in a congested network.



(a) Data Size

(b) Communication Time

Fig. 4: Comparison of communication time according to data differences

VI. Conclusion

In this paper, we presented a novel framework for human synchronization that can be used to construct large-scale digital twins. The proposed framework includes an angular converter and an absolute coordinate calculator to synchronize 3D pose data with virtual world humans. Our experimental results showed that our framework requires less network cost than traditional digital twin frameworks that directly receive raw data. Moreover, we demonstrated that the proposed framework can be used to synchronize a self-driving robot with a digital twin to perceive and interact with people in blind spots of sensors and simulate data that is closer to physical-world human data using stored data.

However, our current research does not synchronize the size and shape of humans, so further research is needed to achieve more precise human synchronization.

Acknowledgment

This research was partly supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program

(IITP-2023-2021-0-01816) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00545, Development of Intelligent digital twin object federation and data processing technology).

References

[1] X. Yi, Y. Zhou, M. Habermann, S. Shimada, V. Golyanik, C. Theobalt, and F. Xu, "Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 167–13 178.

[2] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *Acm transactions on graphics (tog)*, vol. 36, no. 4, pp. 1–14, 2017.

[3] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt, "Xnect: Real-time multi-person 3d motion capture with a single rgb camera," *Acm Transactions On Graphics (TOG)*, vol. 39, no. 4, pp. 82–1, 2020.

[4] "Revolutionizing the world of vision ai." [Online]. Available: https://ultralytics.com/

[5] G. Moon, J. Y. Chang, and K. M. Lee, "Camera distance-aware top-down approach for 3d multi-person pose estimation from a single RGB image," *CoRR*, vol. abs/1907.11346, 2019. [Online]. Available: http://arxiv.org/abs/1907.11346