

Multi-Users Extended Reality Remote Collaborative Operation Application based on Digital Twins 3D Projecting

Ming-Shen Jian*, Chun-Ho Cheng, Jay-Hong Cheng, Ming-Hsun Tsai, Wen-Yu Chung, Chi-Hua Chang

CCIS Lab., Dept. of CSIE, National Formosa University, Yunlin County, Taiwan 632

jianms@nfu.edu.tw*

Abstract— This research investigates image processing and artificial intelligence motion recognition for interaction between virtual 3D objects and users. Digital Twins 3D Objects and environments are presented simultaneously through internet to multiple users. Based on the Google MediaPipe AI model at edge sides, remote servers could collect the information related to the motion of used for remote interaction and virtual 3D interaction. In addition, the virtual environments with 3D object are also modified. Via virtual GPU, Web GPU, and the estimation functions, the physical environments and users can be calculated. Users needn't to have the physical controller, 3D or VR glasses for the proposed XR remote collaborative operation application. The feasibility of the proposed system can be higher with less cost.

Keywords— Image Recognition, Digital Twins, Artificial Intelligence, Extended Reality, MediaPipe

I. INTRODUCTION

Recently, the Virtual Reality (VR) and Augmented Reality (AR) are generally used and developed in education, entertainments and game [1]. By wearing so-called virtual glasses, system users could view the virtual objects. By using the auxiliary hardware devices, system users would do the interaction with the virtual objects in the virtualized world. However, in opinion, due to the glasses, the system users or operators cannot observe intuitively and intuitively. To directly have the response appropriately corresponding to the real or physical environment is not easy. In addition, limited by the operation method of the control devices or modules, users cannot be able to operate or interact intuitively with virtual objects by their hands and fingers. Furthermore, for interacting with the virtual object via control devices and modules, more cost is required. Considering the cost and quality of the user experience, Extended Reality (XR) [2] will gradually become the mainstream technology for digital virtual interaction in the foreseeable future.

In recent years, artificial intelligence image recognition has been able to identify objects, identify operators and gestures, and locate user positions [3,4]. Hence, the core technology of XR is to use computer imaging vision and artificial intelligence image recognition technology to interact with virtual environments and objects. In addition, through the Internet of Things (IoT) devices, the information and data related to the physical (real) world can be merged into the virtual environment. [9] Users can view relevant information about real and virtual environments at the same time. Some studies presented the possibility of spatial navigation [2], entertainment [5,6], and Education [7].

However, to provide the virtual environments and objects being able to interact with operators and users would require computing resources. To obtain huge computing resources for multiple users locate at different positions, Cloud Computing would be the possible solution. A cloud computing platform would do the virtualization of all the computing resources such as CPU, memory, storage spaces, Input/Output devices, etc. Even the progressing mission would require huge computing resources, based on suitable parallel processing and distributed system, the progressing mission could be done with better performance [8]. Through the Internet connection, multiple users locate at different positions could have the online XR services.

The main contributions of this research can be shown as follows:

1. Through artificial intelligence and computer vision, the body and gesture image recognition of the extended reality operator is completed. System users no longer need to hold control equipment or wearable devices,

making the operation more intuitive, intuitive and easy.

2. Utilizing the concepts of remote co-performance and Collaborative operation with appropriate projection design, users in different locations can collaboratively operate the same virtual object of the virtual space corresponding to the XR scenario or environment. The proposed system not only increases interactivity, but also provides more entertainment, educational content for users.
3. The extended reality remote collaborative operation based on Digital Twins 3D Projecting will allow users in the same or different places to conduct relevant actions and operations in advance. Entertainment, education, pre-maintenance training or teaching training, etc., can be developed and deployed with less cost.

The organization of this research is as follows: section 2 presents the research related works including motivation and purpose. Section 3 is the technical content and characteristics of the proposed system. Section 4 is the system verification. Conclusion is given in section 5.

II. RELATED WORKS

Recently, image or object recognition based on AI modules are generally used. To identify or recognize the object, labelling and segmenting the objects with huge image data would be the first step [6,10]. According to the label and segmentation data of the object, the AI training would find the features and rules for recognizing the objects. The training results would be the recognition rules or algorithms corresponding to each object. Generally speaking, to progress the AI service would require the computing resource.

Currently, to progress AI services, cloud computing platform with huge computing resource would be the popular solution [8,9]. Based on the virtualization technologies, the computing resources such as CPU, memory, storage spaces, etc., could be dynamically changed. In other words, local devices would not need to have such huge computing resource. Instead, via network accessing, the local users could obtain the AI service from the remote

cloud server. Therefore, through the suitable solution, the images at local side could be identified and recognized by the AI service on cloud.

In addition to transmit the images or streaming, the first step for objects recognizing included in the images is deploying the YOLO algorithm and computing environment [12]. According to the pre-trained formula or functions with the specific characteristics of the target object, even similar music instrument [13], underwater objects [14], or target object with less different in color [15], can be detected.

Digital twin technology [16-18] based on the Internet of Things, Cloud Computing, Big Data, Artificial Intelligence, Virtual Reality or Augmented Reality, etc., was generally used and deployed for the industry 4.0, education and remote interaction applications. Through the data exchanging and suitable expression, various responses or actions corresponding to the current status or scenario could be chosen and done.

III. PROPOSED SYSTEM

In this research, the proposed system could be divided into different partitions: User Gesture Recognition, Virtual Environment Rendering, System Deployment.

User Gesture Recognition

Through a general photography lens, hands/fingers movements and postures are detected based on YOLO image recognition similar to the proposed research [4]. Then, different recognized hand/finger postures could be used as input commands. These various commands are on demand defined corresponding to different operating functions. Therefore, in this research, it is possible to interact with the virtual environment and objects without wearing any auxiliary hardware devices.

In this research, we use a database to store information and data for artificial intelligence correspond to the user's status. The proposed system captures the data in the database to create different teaching training and operating situations. Then, it connects to the Internet to conduct online operations and communication between multiple users, while synchronizing the extended reality of multiple operators. In order to distinguish and identify the

operation methods and goals of different users, it is necessary to use artificial intelligence image recognition to recognize the user's behaviour or gestures. The recognized gestures corresponding to the specific command content preset in the database could be used as operations control methods for extended reality virtual objects.

Considering being lightweight and quickly deployed on different types of hardware devices, this research uses Google MediaPipe as the basis for artificial intelligence recognition [2]. Google MediaPipe is a development framework for processing machine learning applications. It can be applied to mobile devices (such as PC, Android, IOS) and utilize efficient management of resources (CPU and GPU) to achieve low-latency performance. In this research, the hand recognition module was used to interact with the virtual environment. The perception pipeline can use MediaPipe to construct modular graphics, including model reasoning, media processing algorithms and data conversion. MediaPipe has many recognition modules. In this study, the hand recognition module was used to interact with the simulated environment.

Based on MediaPipe face mesh, through “cv.contourArea” method, the area of the user face could be calculated according to the following equation:

$$Area = \frac{1}{2} |\sum_{i=1}^n (x_i \times y_{i+1} - y_i \times x_{i+1})| \quad (1)$$

where n indicates the total number of the feature points corresponding to the face mesh. In addition, when the value of i is n , the value of $i+1$ will be set as 1. Finally, the size or area of the user face can be evaluated and compared with the normal/average area or size of faces [11]. In other words, comparing with the area or size between the measured face and average face, the possible distance from the camera to the face could be estimated. Following figure shows the evaluation results from the MediaPipe. When an user stand about 1 meter in front of the camera, the MediaPipe outputs the measured x , y , z , and the face size values of the same face.

```
x: 0.46933871507644653
y: 0.37862429022789
z: 0.021847710013389587

Face size : 175.5 square pixels

x: 0.4696352481842041
y: 0.37876802682876587
z: 0.021173691377043724

Face size : 172.5 square pixels
```

Figure 1. The example of the face recognition evaluation result according to the MediaPipe

Virtual Environment Rendering

However, considering the XR Remote Collaborative Operation for multiple users, the presentation of virtual environment or each virtual object corresponding to different operators' perspectives or angles would be an important issue. Therefore, if the only single remote server is responsible for rendering virtual objects or environments from various perspectives, it will occupy a large amount of computing resources on the server side. In this research, considering the computing resource of the local side (user side) would be various, the rendering procedure of the 3D virtual objects could be done by two available methods: 1. rendering by cloud, and 2. rendering at local through network graphics card as the Web GPU.

Due to the complexity of the 3D virtual object rendering, the assistance of Graphics Processing Unit (GPU) is needed. However, not every device at end user side could have the GPU progressing resources. Hence, if the GPU of the users' devices is not available, the cloud server would render the virtual object in advance. According to the parallel processing and distributed system structure [8], the rendering work could be separated and distributed into individual GPU for image rendering based on the cloud computing environment. Finally, the rendered image could be delivered to the client users. In addition to the edge devices for parallel processing, the cloud computing platform could also provide the virtual machines with the assigned GPU for the image rendering. The remote client user can connect to the corresponding virtual machine for the rendered images.

In opposition, if there exists the GPU at the client user side, to render the image and virtual object at

user side by transmitting a few related data would be also a good solution. Since there would be multiple users in the virtual environment at the same time, reducing the graphics computing load on the server side can also avoid network delays and congestion caused by the operating from multiple users or large amounts of 3D virtual objects or environment information transmission. Therefore, via delivering the 3D virtual objects or environment information, the user side devices could render the virtual objects or environment through the Web GPU. Therefore, only a few data would need for transmission through network.

To deploy the environment for more possible 3D virtual object designing, Unity for the implementation of proposed, which is commonly used for 3D environment construction, is used. Its functions provide the application of online (network) GPU. Unity can achieve outstanding results in color rendering through Web GPU. When performing technical services and teaching training, the color distinction of the drawn extended reality virtual objects or environments will be more realistic.

When considering various 3D models for teaching training or technical operations, there may be different situations or scripts. Therefore, the database will be connected as a storage for storing data, including object parameters, restrictions and trigger conditions, etc. In addition to using Unity to complete the 3D virtual environment and objects, the proposed system in this research also needs to be able to provide different users with collaborative operations in different locations through the Internet.

IV. VERIFICATION RESULTS

In this research, in order to be able to create a virtual environment for different needs, we create rooms in Unity. When entering the room, a script connected to python will be triggered. The camera will be turned on by python and start recognition progress. At the same time, another script will receive the data transmitted by python. Next, the unity will capture the data of the corresponding scene.

As shown in Figure 2, the gesture artificial intelligence recognition results are obtained by capturing the operator's gestures through the camera.

From the predefined observer angle, the recognized gestures, finger skeletons and other information can be rendering in the extended reality virtual space. The left side of the picture shows the perspective of the gesture operator. The left and right hands have been recognized and rendered by the MediaPipe hand recognition module as 3D virtual gestures and skeletons (purple line segments). The right side of the picture is the perspective facing the gesture operator. Hence, the virtual rendering result of the gesture appears in a shadow state (darker gray) instead of the reflection on the left (whiter and brighter). It shows that the 3D virtual gestures and skeletons drawn using the MediaPipe hand recognition module can be quickly integrated into the extended reality virtual environment. Users can be more intuitive and intuitive of manipulating virtual objects in extended reality.

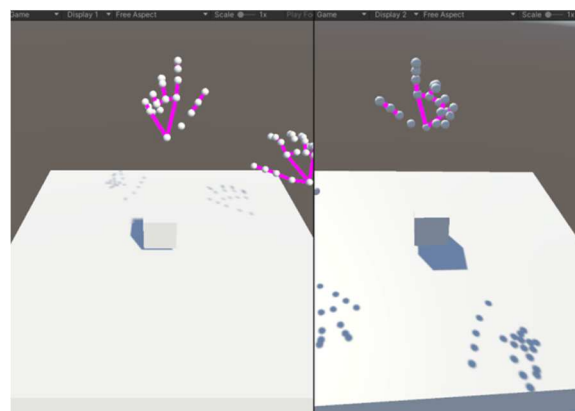


Figure 2. The example of the MediaPipe hand recognition

When the gesture recognition module completes and transmits the recognition results, the result information is processed by Unity. Corresponding to different recognition results are represented by different operation command inputs. Therefore, unity can operate virtual objects according to the operation commands, including clicking, translation and movement. The results of these extended reality computing interactions will be rendered on the unity screen.

Figure 3 presents the virtual 3D objects operation and environment for multiple users. In this scenario, the users will locate at the same side (both front side or back side) to operate and collaborate the 3D objects based on the gesture artificial intelligence recognition. Only one camera is required for the

gesture artificial intelligence recognition in this collaboration scenario. It also means that the virtual environment could be transmitted and presented in different locations for different users with the same collaboration scenario. Multiple users could collaborate with each other by operating the same virtual 3D objects or digital twins scenario.



Figure 3. The digital twin verification example of the proposed system: collaboration of multiple users in XR machine operation.

Figure 4 shows the new application of digital twins based on the proposed system. Multiple users will locate at different viewpoints corresponding to the virtual 3D objects of digital twins. Hence, the location, distance, etc., of the virtual 3D objects would be different corresponding to the users' viewpoints. Based on the gesture artificial intelligence recognition and relative position of 3D objects, the operation at one user side and the collaboration between multiple users could be still correctively presented.

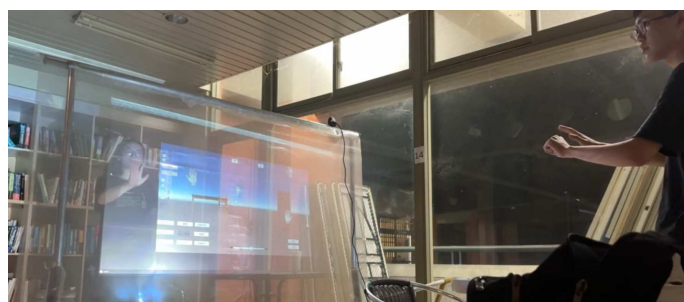


Figure 4. The digital twin verification example of the proposed system: collaboration of multiple users corresponding to various viewpoints of virtual 3D objects.

Using the 3D space network connection method built by Unity itself, multiple users can connect to the same 3D space. Therefore, using the gesture

recognition module to transmit the recognition results, the operation status of individual users can be identified. At last, by using the new identification program for unity, you can operate a single 3D virtual object with more possibilities, including operation with only one hand, operation with two hands, operation with two peoples' four hands, etc. In addition, through different users, even in different locations, via extended reality, the operations of multiple users can be completed in the same direction, opposite directions, etc., to complete multiple users' remote collaboration.

V. CONCLUSIONS

This research can overcome the time and physical space limitations of traditional physical education and training. System users could carry out collaborative work or educational training in the digital twin virtual world anytime and anywhere. The proposed system can simulate various physical scenes and objects, and promote the possibility of collaborative cooperation through instant feedback and evaluation.

ACKNOWLEDGEMENT

Thank you for the contribution and assistance of Cloud Computing and Intelligent System Lab. in National Formosa University, Taiwan.

REFERENCES

- [1] M. C. Hsieh and Y. H. Lin, "VR and AR applications in medical practice and education," *The Journal of Nursing*, 64(6), 12–18, 2017doi:10.6224/JN.000078
- [2] Hong-Peng Yin, Bo Chen, Yi Chai, Zhao-Dong Liu. "Vision-based object detection and tracking: a review." *Acta Automatica Sinica*, 42(10): 1466–1489, 2016.
- [3] H. -Y. Huang, M. -H. Tsai and M. -S. Jian, "Multiple Merged Structures Based on Image Recognition for Converting Application of Natural Language Artificial Intelligence Service," 2024 26th International Conference on Advanced Communications Technology (ICACT), Pyeong Chang, Korea, Republic of, 2024, pp. 158-162, doi: 10.23919/ICACT60172.2024.10471930.
- [4] N. Capece, G. Manfredi, V. Macellaro and P. Carratù, "An easy Hand Gesture Recognition System for XR-based collaborative purposes," 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRINE), Rome, Italy, 2022, pp. 121-126, doi: 10.1109/MetroXRINE54828.2022.9967592.
- [5] T. Akiyoshi, Y. Shimizu, Y. Takahama, K. Nagata and T. Sawabe, "Hype D-Live: XR Live Music System to Entertain Passengers for Anxiety Reduction in Autonomous Vehicles," 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Sydney, Australia, 2023, pp. 148-156, doi: 10.1109/ISMAR59233.2023.00029.
- [6] U. Žáková and D. Sedláček, "Studying the Role of Non-Player Users (Spectator and Operator) in XR Single and Multiplayer Experiences," 2023 14th IEEE International Conference on Cognitive

- Infocommunications (CogInfoCom), Budapest, Hungary, 2023, pp. 000129-000134, doi: 10.1109/CogInfoCom59411.2023.10397540.
- [7] B. Aresh et al., "Integration of Extended Reality (XR) in Non-Native Undergraduate Programmes," 2023 15th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), Kuala Lumpur, Malaysia, 2023, pp. 285-292, doi: 10.1109/SKIMA59232.2023.10387333.
- [8] M. -S. Jian, J. -T. Gui, X. -M. Wu, "Uncentralized Artificial Intelligence Computing Agent with the Distributed Training and Computing Tasks Based on Open Source Cloud Proxy," CSCE 2023 - ICAI, Las Vegas, USA, p.93. (ISBN:1-60132-518-5)
- [9] M. -S. Jian, Y. -C. Wang, B. -H. Wu and Y. -E. Cheng, "Hybrid cloud computing for user location-aware augmented reality construction," 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon, Korea (South), 2018, pp. 1-1, doi: 10.23919/ICACT.2018.8323691.
- [10] A. Torralba, B. C. Russell and J. Yuen, "LabelMe: Online Image Annotation and Applications," in Proceedings of the IEEE, vol. 98, no. 8, pp. 1467-1484, Aug. 2010, doi: 10.1109/JPROC.2010.2050290.
- [11] I. Badri and M. Sayyouri, "Face Recognition: A Mini-Review." In: S. Motahhir, and B. Bossoufi, (eds) Digital Technologies and Applications. ICTDA 2023. Lecture Notes in Networks and Systems, vol 669. Springer, Cham, 2023. https://doi.org/10.1007/978-3-031-29860-8_47.
- [12] M. Mahendru and S. K. Dubey, "Real Time Object Detection with Audio Feedback using Yolo vs. Yolo_v3," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 734-740, doi: 10.1109/Confluence51648.2021.9377064.
- [13] C. Dewi, R. -C. Chen, Hendry and Y. -T. Liu, "Similar Music Instrument Detection via Deep Convolution YOLO-Generative Adversarial Network," 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), Morioka, Japan, 2019, pp. 1-6, doi: 10.1109/ICAwST.2019.8923404.
- [14] X. Wang, X. Jiang, Z. Xia and X. Feng, "Underwater Object Detection Based on Enhanced YOLO," 2022 International Conference on Image Processing and Media Computing (ICIPMC), Xi'an, China, 2022, pp. 17-21, doi: 10.1109/ICIPMC55686.2022.00012.
- [15] W. Yijing, Y. Yi, W. Xue-fen, C. Jian and L. Xinyun, "Fig Fruit Recognition Method Based on YOLO v4 Deep Learning," 2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Chiang Mai, Thailand, 2021, pp. 303-306, doi: 10.1109/ECTI-CON51831.2021.9454904.
- [16] Y. Ye, "A Higher Education Information Management System Based on Digital Twin Technology and Artificial Intelligence Algorithms," 2024 Asia-Pacific Conference on Software Engineering, Social Network Analysis and Intelligent Computing (SSAIC), New Delhi, India, 2024, pp. 172-176, doi: 10.1109/SSAIC61213.2024.00039.
- [17] Z. Bosić, V. Čačković, T. Marušić, R. Radović, P. Gusić and T. Žitnik, "Augmented Reality Modeling Support Using a Digital Twin," 2024 47th MIPRO ICT and Electronics Convention (MIPRO), Opatija, Croatia, 2024, pp. 771-778, doi: 10.1109/MIPRO60963.2024.10569151.
- [18] Zexuan Zhu, Chao Liu and Xun Xu, "Visualisation of the Digital Twin data in manufacturing by using Augmented Reality", Procedia CIRP, vol. 81, pp. 898-903, 2019, <https://doi.org/10.1016/j.procir.2019.03.223>



Ming-Shen Jian was born in Kaohsiung City, Taiwan in 1978. He received the B.S. from the National Chiao Tung University, HsinChu, and Ph.D degrees in Computer Science and Engineering from the National Sun Yat-sen University, Kaohsiung, Taiwan in 2007. From 2018, he was an Associate Professor and director with the National Formosa University Cloud Computing and Intelligent System Laboratory.

Currently he is also an IEEE Senior Member. Since 2009, he has been an Assistant Professor with the Computer Science and Information Engineering Department, National Formosa University. He is the author of four books, more than 50 articles, and at least 15 invention patents. His research interests include IOT development and application, Big Data, Optimal Solution, Intelligent System, and Cloud Computing. He was a Secretary of the Taiwan Association

of Cloud Computing. Dr. Jian was a recipient of the IEEE sponsored international conference Paper Award in 2016, 2017, and 2018.



Chun-Ho Cheng currently is a bachelor degree student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to Cloud Computing and Docker container services, security, and intelligent system. He joins the Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2022.



Jay-Hong Cheng currently is a bachelor degree student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to Cloud Computing and Docker container services, security, and intelligent system. He joins the Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2022.



Ming-Hsun Tsai currently is a master degree student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to object recognition and segmentation, and image processing. He joins the Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2023.



Wen-Yu Chung currently is a bachelor degree student of Dept. Computer Science and Information Engineering at National Formosa University. Her current research interests are in the area related to object recognition and segmentation, and image progressing. She joins the Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2022.



Chi-Hua Chang currently is a bachelor degree student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to object recognition and segmentation, and image progressing. He joins the Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2023.