J. Usón et al., "Enabling Real-Time Collaborative Cultural Experiences with Free Viewpoint Video," 2025 17th International Conference on Quality of Multimedia Experience (QoMEX), Madrid, Spain, 2025, pp. 1-4, doi: 10.1109/QoMEX65720.2025.11219989.

Enabling Real-Time Collaborative Cultural Experiences with Free Viewpoint Video

Javier Usón, Victoria Muñoz, Carlos Cortés, Isabel Rodríguez, César Díaz, Jesús Gutierrez and Julián Cabrera
Grupo de Tratamiento de Imágenes, Information Processing and Telecommunications Center,
ETSI Telecomunicación, Universidad Politécnica de Madrid. Madrid, Spain
{j.usonp, victoria.munoz.murillo, carlos.cs, im.rodriguez, cesar.diazm, jesus.gutierrez, julian.cabrera}@upm.es

***Abstract*—In this demo, we showcase a system that applies extended reality (XR) and volumetric video to enhance cultural experiences. The platform enables real-time interaction between users and a presenter within a virtual environment enriched with 3D assets. Users can visualize a volumetric representation of the presenter captured in real-time using a Free Viewpoint Video (FVV) system. Additionally, the presenter can control the elements in the scene using hand gestures, recognized by an artificial intelligence (AI) model. We propose a methodology to assess the subjective quality of this experience through user studies. The complete system will be demonstrated, including the real-time volumetric capture and the immersive application to participate in the cultural experience.**

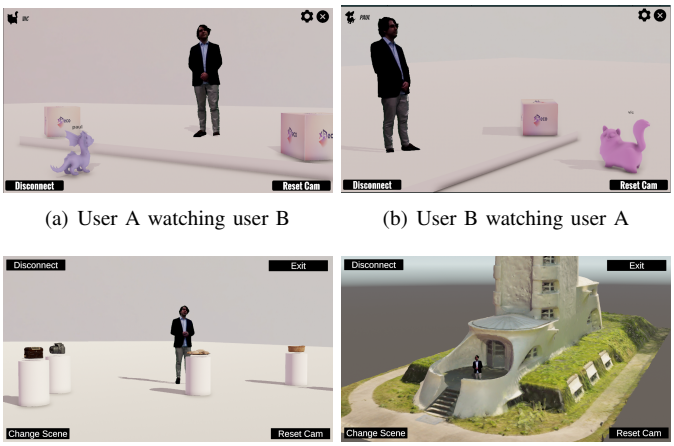
***Index Terms*—Free Viewpoint Video, FVV, Immersive Communications, Immersive Videoconferencing, Streaming media, Real-time system**

I. Introduction

Extended reality (XR) empowers cultural experiences by immersing users in virtual environments where virtual assets and 3D reconstructed objects bring them closer to real-world scenarios. In this context, volumetric video [1] enables the integration of real-life content into such experiences, the main application being volumetric representations of people, creating realistic avatars that can be viewed and interacted with by other users.

Existing XR platforms often focus on static or pre-recorded content. In contrast, we propose a system that enables immersive real-time communication and interaction between users. The core of this system is based on immersive videoconference systems [2], a technology postulated to overcome the current problems of traditional 2D videoconference, such as video conferencing fatigue, mainly caused by lack of free movement or flat representation of users [3].

In this work, we propose an experiment that involves a cultural experience built with our system and discuss the methodology that will be used to evaluate both the performance of the system and its quality of experience (QoE). This experiment aims to develop an immersive platform for cultural visits and events. The system will provide an XR experience in which users equipped with virtual reality (VR) head-mounted displays (HMDs) remotely connect to a virtual scene. This scene will showcase a presenter explaining cultural topics supported by related objects reconstructed as 3D models. The experience will include the following:



(a) User A watching user B (b) User B watching user A

(c) Topic related 3D models (d) 3D reconstructed scenario

Fig. 1. Examples of the features implemented in the cultural experience. Users are able to visualize a volumetric representation of a presenter while also having the ability to visualize and communicate to other users. Additionally, the virtual scene can be populated by 3D models that the presenter can manipulate using gesture recognition.

- A real-time volumetric representation of the presenter using Free Viewpoint Video (FVV).
- Artificial intelligence (AI)-based hand gesture recognition, used by the presenter to manipulate objects from the scene.
- Representation of other users connected to the same virtual scene as 3D avatars, as well as audio communication channel to interact with them.

The volumetric representation of the presenter is based on the *FVV Live* system [4] [5]. This technology allows its users to visualize a real-time captured scene by freely deciding on their point of view. This system follows a remote rendering approach, transmitting only 2D images corresponding to the user's perspective, rather than the full volumetric information.

Figure 1 shows examples of the volumetric representation of the presenter, interactive 3D objects, and the representation of other users as 3D avatars.

II. System Description

Figure 2 shows a schematic view of the architecture designed to carry out the experiment. The main components are detailed in the following subsections.

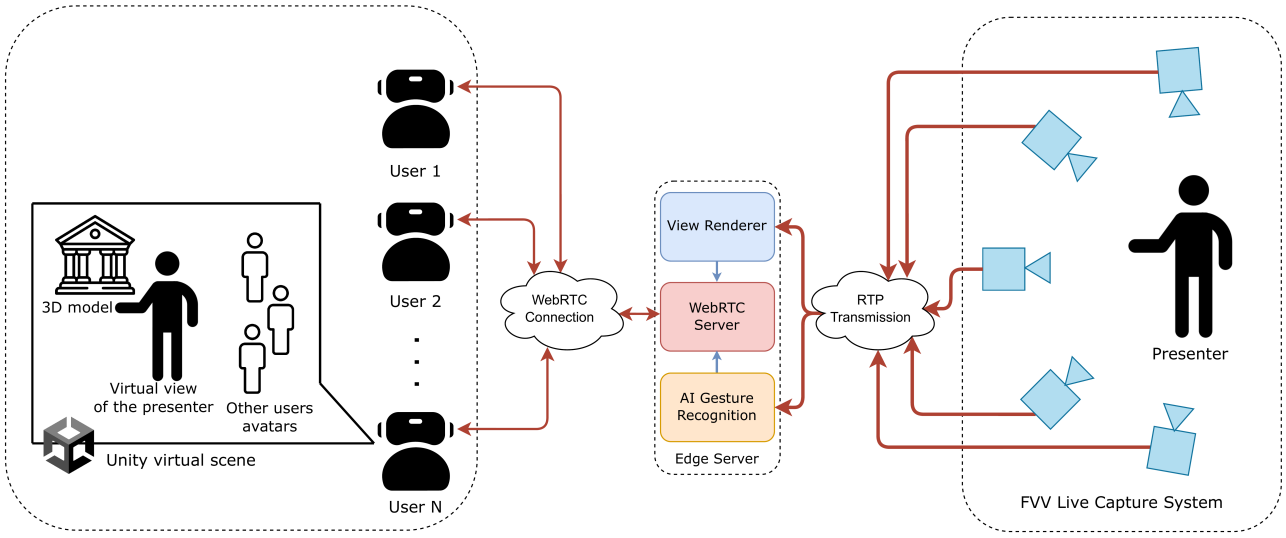


Fig. 2. Architecture of the system dedicated to the experiment. Users access the scene through a WebRTC connection managed by the WebRTC Server component. On the other side, a presenter is captured by the *FVV Live* capture module, generating one RGB+D stream per camera which is transmitted to the View Renderer instances and to the gesture recognition model using a low-latency protocol (RTP). Users receive each rendered virtual view as a video stream and the signaling related to gesture control by a WebRTC datachannel.

A. Volumetric Capture

In the experiment, the *FVV Live* capture module captures a presenter in real time using a set of stereo cameras capable of obtaining the geometric information of the scene in the shape of color and depth images (RGB+D).

Segmentation is applied to the camera feeds to separate the foreground (the presenter) from the background to enhance the integration of the presenter into the scene. Since this segmentation must be performed in real time, a green cyclorama such as the one shown by Figure 3 is used to facilitate the task.

The color and depth information is encoded as two separated H264 video streams, with regular lossy compression used for color and a custom lossless scheme for depth [4]. Finally, to ensure low-latency transmission, streams are encapsulated in RTP [6] and sent to the edge server for processing.

B. Virtual View Rendering

The *FVV Live* Virtual View Renderer is used for the volumetric representation of the presenter in the virtual scene. The system follows a remote rendering approach, where the user communicates their position to the View Renderer which synthesized a virtual view coherent to that position. The method for selecting the virtual point of view to be rendered is detailed in [7].

The virtual views rendered are transmitted to the user as a video stream. In the virtual scene, this video is played on a billboard, a plane that rotates to always face the user.

Since users can navigate the scene with 6 DoF independently, each one of them requires a specific viewpoint of the presenter. To serve simultaneous users, several instances of the View Renderer component are deployed, each one synthesizing views for one user.

C. AI-based Gesture Recognition

Gesture recognition is used to allow the presenter to control the virtual scene in a convenient way. The gesture recognition module receives the color video streams from the capture and processes them with a deep learning model [8].

Gestures can be used to control several aspects of the virtual scenario, such as rotating or scaling 3D models, or changing the scenario around users completely.

D. WebRTC Server and Client Application

Users connect to the experience through an application built in Unity, which is available in a desktop version (for regular 2D screens) and a version developed for HMDs (VR). Both versions use WebRTC connections [9] managed by the WebRTC Server component to access the following services:

- Access their specific instance of View Renderer.
- Receiving the events generated by the presenter's gestures
- Receiving the position of each of the other users to place their 3D avatars on the virtual scene in real-time.
- Bi-directional audio conference to communicate with the presenter and the rest of the users.

WebRTC enables low-latency communications by using RTP media transmission and DTLS [10] with retransmissions for data channels. In this experiment, RTP is used by the video generated in the View Renderer instances and by the audio conference streams. On the other hand, DTLS is used by signaling messages with the position of the users and to notify the events generated by gestures.

III. Quality of Experience Analysis

This methodology aims to capture how the immersive system compares with existing formats and how robust the

Fig. 3. *FVV Live* capture module setup. 9 stereo cameras (Stereolabs ZED) capture both the color and depth information from the scene in a green cyclorama setup to ease the process of real-time segmentation.

experience is under real-world transmission conditions. To evaluate the effectiveness of the proposed cultural experience based on XR, we will focus on a subjective assessment of QoE.

Methodological recommendations exist in the literature on how to conduct QoE studies according to the multimedia use case. Traditional video experiences have been evaluated according to the following standards: ITU-T BT 500 for televised video [11] , ITU-T 910 [12] for video in multimedia applications, P.913 [13] for visualizing multimedia content on devices such as tablets and laptops. These recommendations set out the subjective evaluation methodology to measure QoE in such video scenarios. However, our system includes not only video, but also interactivity. There are also recommendations for scenarios where interactivity and video are combined, such as P.920 [14]. ITU-T P.920 sets out questionnaires and tasks for the evaluation of QoE in interactive audiovisual communications. Finally, P.919 [15], Subjective test methodologies for 360º video on head-mounted displays sets out a framework for evaluating experiences while wearing an HMD, which includes recommendations for evaluating cybersickness.

Based on these recommendations, we propose the following methodology:

- Interactive audiovisual quality using the standardized questionnaires contained in P.920.
- Cybersickness, particularly in comparison to traditional 2D videoconferencing platforms, contained in P.919.
- Usability, focusing on the naturalness of communication through avatars and gesture-based controls using the System Usability Scale.
- Sensitivity to video quality and latency, evaluating how the subjective experience of users changes under varying levels of encoding quality and network delay conditions.

Participants will experience all three formats in randomized order. In addition, participants will experience the XR session under different network and compression configurations to assess their tolerance to reduced quality and increased latency.

IV. Demo Setup

In our demo, three simultaneous users will be able to test the experience. The cameras will capture one of them, while the other two will visualize the virtual scene through the desktop or HMD Unity-based application.

The equipment needed to be transported to the demo site consists of the following:

- Nine Stereolabs ZED cameras and a microphone, managed by three Capture Servers.
- A portable green screen setup.
- One "Edge Server" to run the View Renderer instances, the gesture recognition model and the WebRTC Server.
- One WiFi router connected to the Edge Server.
- One laptop to run the desktop Unity application.
- One Meta Quest 3 to run the HMD Unity application.

With respect to networking, the purpose of the router is to perform the demo inside a local network. The capture servers will have a direct wired connection to the Edge Server to avoid any bandwidth problems when delivering RGB+D information. Additionally, the router will ensure a strong connection from the clients to the edge server.

V. Conclusions

In this work, we propose a demo setup to test a collaborative cultural experience built with our immersive videoconference system.

The system involves a virtual scenario that includes features such as visualizing a volumetric representation of a presenter using real-time FVV, interacting with other users using audio communication, and observing them inside of the scene as virtual avatars. Additionally, the presenter can control the virtual scene through hand gestures thanks to an AI-based detection model.

Both the client application and the real-time volumetric capture system, including the multi-camera setup, will be demonstrated. Participants will be able to experience the demo using immersive displays and control the virtual environment with their own hands.

Regarding QoE assessment, we proposed a methodology to evaluate the impact of immersive communication, AI-based controls, and transmission quality and latency on the overall experience.

Acknowledgment

This work was supported by these projects: PID2020-115132RB (SARAOS) and PID2023-148922OA-I00 (EEVOCATIONS) funded by MCIN/AEI/10.13039/501100011033 of the Spanish Government, TED2021-131690B-C31 (Revolution) funded by MCIN/AEI /10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, and TSI-063000-2021-80 (DISRADIO –Pilotos) funded by the Ministry of Digital Transformation of the Spanish Government and NextGenerationEU/PRTR.

References

[1] Yili Jin, Kaiyuan Hu, Junhua Liu, Fangxin Wang, and Xue Liu, "From capture to display: A survey on volumetric video," 2023.

[2] Pablo Pérez, Ester Gonzalez-Sosa, Jesús Gutiérrez, and Narciso García, "Emerging immersive communication systems: Overview, taxonomy, and good practices for qoe assessment," *Frontiers in Signal Processing*, vol. 2, 2022.

[3] Andrew A Bennett, Emily D Campion, Kathleen R Keeler, and Sheila K Keener, "Videoconference fatigue? exploring changes in fatigue after videoconference meetings during covid-19.," *Journal of Applied Psychology*, vol. 106, no. 3, pp. 330, 2021.

[4] Pablo Carballeira, Carlos Carmona, César Díaz, Daniel Berjón, Daniel Corregidor, Julián Cabrera, Francisco Morán, Carmen Doblado, Sergio Arnaldo, María del Mar Martín, and Narciso García, "FVV Live: A Real-Time Free-Viewpoint Video System With Consumer Electronics Hardware," *IEEE Transactions on Multimedia*, vol. 24, pp. 2378–2391, 2022.

[5] Pablo Pérez, Daniel Corregidor, Emilio Garrido, Ignacio Benito, Ester González-Sosa, Julián Cabrera, Daniel Berjón, César Díaz, Francisco Morán, Narciso García, Josué Igual, and Jaime Ruiz, "Live Free-Viewpoint Video in Immersive Media Production Over 5G Networks," *IEEE Transactions on Broadcasting*, vol. 68, no. 2, pp. 439–450, 2022.

[6] Henning Schulzrinne, Stephen L. Casner, Ron Frederick, and Van Jacobson, "RTP: A Transport Protocol for Real-Time Applications," RFC 3550, July 2003.

[7] Javier Usón, Victoria Muñoz, Carlos Cortés, Daniel Berjón, Francisco Morán, César Díaz, Jesús Gutierrez, Fernando Jaureguizar, Narciso García, and Julián Cabrera, "Real-time free viewpoint video for immersive videoconferencing," in *2024 16th International Conference on Quality of Multimedia Experience (QoMEX)*, 2024, pp. 171–174.

[8] Zihan Ni, Jia Chen, Nong Sang, Changxin Gao, and Leyuan Liu, "Light yolo for high-speed gesture recognition," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 3099–3103.

[9] WebRTC Working Group, "Web real-time communication (webrtc)," https://www.w3.org/TR/webrtc/, 2021.

[10] Eric Rescorla, Hannes Tschofenig, and Nagendra Modadugu, "The Datagram Transport Layer Security (DTLS) Protocol Version 1.3," RFC 9147, Apr. 2022.

[11] Rec. ITU-T BT.500, "Methodologies for the subjective assessment of the quality of television images, document recommendation," 2023.

[12] Rec. ITU-T P.910, "Subjective video quality assessment methods for multimedia applications," 2023.

[13] Rec. ITU-T P.913, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," 2023.

[14] Rec. ITU-T P.920, "Interactive test methods for audiovisual communications," 2000.

[15] Rec. ITU-T P.919, "Subjective test methodologies for 360º video on head-mounted displays," 2020.