Real-Time Free Viewpoint Video for Immersive Videoconferencing

Javier Usón, Victoria Muñoz, Carlos Cortés, Daniel Berjón, Francisco Morán,
César Díaz, Jesús Gutierrez, Fernando Jaureguizar, Narciso García 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, daniel.berjon, francisco.moran, cesar.diazm,
jesus.gutierrez, fernando.jaureguizar, narciso.garcia, julian.cabrera}@upm.es

***Abstract*—In this work, we propose a demo of an immersive videoconference system using Free Viewpoint Video (FVV) technology. It makes use of the *FVV Live* system, which covers the entire FVV pipeline (capture, view rendering, and visualization) while working in real-time. The *FVV Live* system consists of nine cameras that capture an environment and a view renderer that uses the information from the cameras to generate a synthetic view at an arbitrary point.**

It is designed as a hybrid demo. While the capture and rendering processes take place at our premises, *FVV Live* can be visualized through devices connected to the Internet.

The system allows immersive navigation of a virtual scene with 6 degrees of freedom, and interaction with live-captured avatars integrated in such scene. For this purpose, it uses WebRTC connections to update the position of the virtual camera and to receive the *FVV Live* view encoded as a video.

Additionally, the user will be recorded by a simple camera and microphone setup, and the generated streams will be transmitted to our premises through the same WebRTC server. This way, people being recorded by *FVV Live* will be able to see and hear the user, enabling bidirectional communication.

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

I. Introduction

Immersive conferencing systems are set to be the next step in teleconferencing. In this sense, the possibility offered by immersive video in terms of interaction and immersion is postulated as a solution to problems such as zoom fatigue [1]. Among immersive video systems, we can find teleconferencing based on 3D avatars [2], 360 video [3], and volumetric capture [4]. Although 3D avatars are the simplest and most accessible form, they offer the lowest levels of fidelity and immersion [5]. 360 video allows immersion, but restricts movement to 3 degrees of freedom as it is a fixed-point capture. Finally, volumetric video teleconferencing offers the possibility of 6 degrees of freedom while maintaining high image fidelity. Thus, volumetric video is postulated to be the best option in terms of immersion.

However, current volumetric video systems require complex installations for multi-camera capture and high computational capacity for 3D rendering. Additionally, the number of applications capable of displaying volumetric video is limited and often requires powerful graphics processing hardware [6].

(a) Full Unity virtual Scene (b) View synthesized by *FVV Live*

Fig. 1. *FVV Live* avatars integrated inside of a Unity virtual scene. *FVV Live* renders a virtual view of the avatars (right), which is then displayed inside of the virtual scene (left).

In this work, we present an implementation of immersive volumetric videoconferencing systems based on Free Viewpoint Video (FVV). In this type of technology, a multi-camera system synthesizes a 2D view of an arbitrary position in space. Our *FVV Live* system allows the visualization of avatars in a virtual world displayed with a head-mounted display (HMD). Thanks to the head tracking system incorporated in HMDs, the demand for new views can be controlled without having to use an external controller, enabling immersive videoconferencing. Additionally, our system allows avatars to be viewed over the Internet, enabling users to connect from anywhere. In this sense, since avatar views are in 2D, we can take advantage of the existing video transmission infrastructure for both streaming and viewing, thus making immersive videoconferencing more accessible.

The synthetic view of the avatars is introduced into a virtual scene generated with the Unity engine. Fig. 1 illustrates the capture and segmentation of the avatars and their inclusion in the virtual world.

Although view capture and synthesis still require specialized hardware, our volumetric videoconferencing approach allows the visualization process to be performed from commercial devices such as the Meta Quest 3 HMD.

This experience is designed as a hybrid demo in which users can visualize and interact with a group of remote people. The system will be deployed mainly at our premises in Madrid, with the only component needed to be transported to the demo site being a terminal where the visualization application runs.

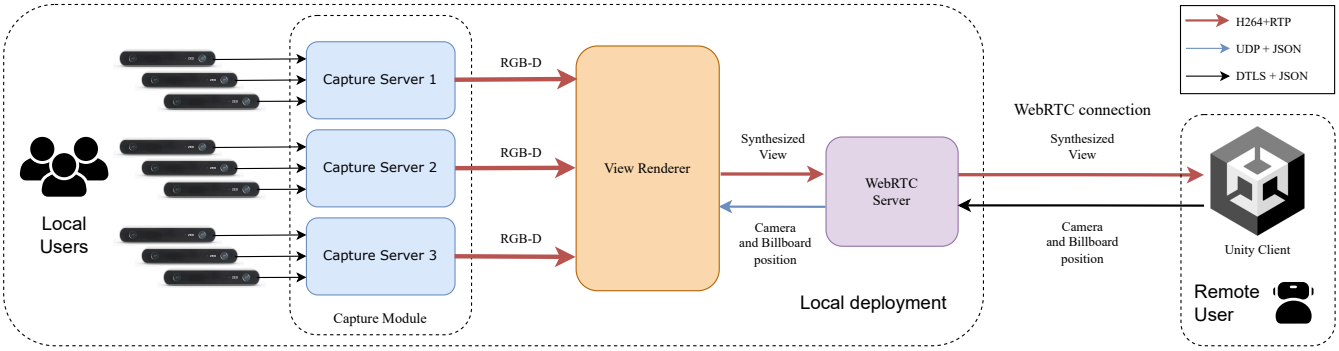

Fig. 2. Architecture of the deployment used in the demo. The Capture Module, View Renderer, and WebRTC Server are deployed locally. This way, the only remote element is the Unity application, which is able to connect to the local deployment through the WebRTC Server. H264 encoding and RTP transmission are used for video, with a custom lossless scheme for depth streams. Position data is encoded in a simple JSON and transmitted using UDP. The connection between the WebRTC Server and the Unity client makes use of the WebRTC data channel, which adds encryption to the UDP messages (DTLS).

II. System Architecture

The FVV application involved in this demo is the *FVV Live* system. Proposed in [7] and extended in [8], it is capable of covering the entire FVV pipeline while working in real time: volumetric capture of a scene, encoding and transmission of the geometrical information, and virtual view synthesis.

The proposed architecture for the demo is presented in Fig. 2. It is based on the first iteration of the integration of *FVV Live* with immersive applications built on Unity presented in [9]. Its main focus is to deploy the heavy-processing components locally, so the remote user only needs a terminal capable of running the Unity application, such as a laptop or an HMD.

The implementation is divided into two main modules: the Capture Module and the View Renderer, with an additional third module that enables communication with a Unity client through the WebRTC protocol [10].

A. Capture Module

The Capture Module handles the volumetric capture of the scene, encoding, and transmission to the View Renderer. It makes use of nine Stereolabs ZED cameras, which are capable of capturing both texture and depth information from the scene using Stereo Matching algorithms.

Captured depth information is represented by depth maps. For transmission, these maps are encoded using a lossless video format and a custom bit interleaving scheme that takes advantage of the 4 chrominance bits to achieve a 12 bits per pixel video. This encoding results in a huge output bitrate. Consequently, segmentation is applied to the depth stream to reduce the bandwidth requirement by only transmitting the depth from the captured avatars to the View Renderer.

B. View Renderer

The View Renderer is the module in charge of interpolating the captured information to synthesize the requested virtual view.

First, it receives and decodes the information transmitted by the Capture Module to retrieve the texture and depth information from the scene. Then, the virtual view is synthesized using backward depth image-based rendering (DIBR). The new view is synthesized using the 3 cameras closest to the updated viewpoint [11].

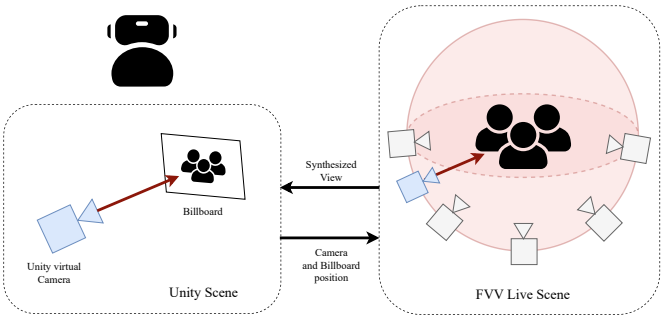

Fig. 3. The Unity application communicates the position of the virtual camera and of the billboard where the video is going to be played to the View Renderer. The View Render then synthesizes a view from a camera that is positioned on the surface of a sphere around the scene, pointing to its center.

The rendered images are encoded as 2D video and transmitted to the WebRTC server.

C. WebRTC Server

The WebRTC Server is the component that manages the connection between the remote user and the View Renderer. The main streams that it handles are the data channel carrying the position information from the Unity client and the synthesized video stream from the View Renderer.

In general, the WebRTC protocol takes care of the decision on encoding and transmission parameters. To achieve a lower latency, the WebRTC Server uses a custom configuration that avoids decoding the video stream coming from the View Renderer and simply relays the stream to the remote user. In this scenario, the encoding process takes place in the View Renderer, utilizing hardware acceleration with controlled parameters.

In addition, it can provide a simple return channel for video and audio. This feature allows local users to see and hear the remote one through a screen.

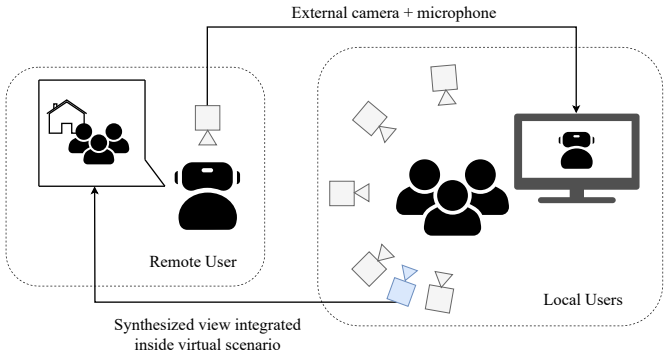

Fig. 4. Diagram of the communication channels in the demo. In both cases, users receive audio and video feedback from the other side.

D. Immersive visualization application

The immersive visualization application consists of a virtual scenario built on Unity. To integrate the avatars captured by the *FVV Live* system, a billboard is placed within the scenario. This billboard will display the video transmitted from the View Renderer while always facing the camera.

Fig. 3 shows a diagram explaining how the View Renderer decides on the virtual viewpoint to render the avatars. The Unity application communicates the position of the camera and the billboard encoded in a JSON message. The View Renderer then uses that information to compute the direction from where the Unity application camera is watching the billboard, and places the *FVV Live* virtual camera on the surface of a sphere around the scene. The sphere is defined approximately following the arc described by the physical cameras.

This approach has two main advantages:

- The *FVV Live* viewpoint is independent of the camera orientation (rotation). *FVV Live* always renders the avatars standing in place, the Unity engine is the one in charge of handling their location and orientation based on where the billboard is.
- Having the avatars fixed in place greatly reduces the effect of motion-to-photon (M2P) latency. Since the location of the avatars does not depend on the transmitted video, the effects of delay are much less noticeable. This was one of the main problems observed in our previous work [9].

III. Immersive Communication Demo

The proposed demo consists of an immersive videoconference setup with bidirectional communication between local users (people recorded by the *FVV Live* system) and one remote user.

Fig. 4 presents a diagram of the proposed bidirectional communication system. Local users are recorded by the *FVV Live* system and visualize the remote user through a screen. The remote user is recorded from an external camera and microphone setup (connected to a laptop) and visualizes the local users integrated inside an immersive scene.

A. Local deployment

The local setup involves the following components:

- Nine Stereolabs ZED cameras and a microphone, managed by three capture servers.
- One server working as the View Renderer.
- An additional server managing the WebRTC connections.
- A screen with speakers to visualize the remote user, managed by one of the capture servers.
- A stage with a green screen setup.

The capture servers will have a direct wired connection to the View Rederer to avoid any bandwidth problems when transmitting the depth information. The connection between the View Renderer and the WebRTC Server will also be wired.

B. Remote equipment

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

- A Meta Quest 3 HMD, which will run the visualization application.
- A simple camera and microphone setup managed by a laptop. Used to record the remote user.

Both components communicate with the local deployment through the Internet, connecting to the WebRTC Server.

IV. Conclusions

In this work, we propose a hybrid videoconferencing demo that uses FVV technology to allow users to visualize a group of people integrated into a virtual immersive environment. Additionally, the system will provide a return video channel and a bidirectional audio channel, enabling the interaction between local and remote users.

The system is designed to transmit a synthesized virtual view to the user, taking advantage of the existing video transmission infrastructure. This way, the heavy-processing components can be deployed locally, minimizing the equipment required for the remote user.

On the topic of visualization, the approach proposed to integrate the FVV view into the virtual scene using a billboard helps reducing the impact of the delay in the transmission despite not adding complexity to the visualization application.

As future work, we propose adapting the deployment and the WebRTC Server to handle several simultaneous remote users receiving the same *FVV Live* transmission.

Acknowledgment

This work was supported by the projects: HORIZON-IA-1010702-50 (XRECO) funded by the European Union, PID2020-115132RB (SARAOS) 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, TSI-063000-2021-4 (6G-Openverso - Holo) and TSI-063000-2021-80 (DISRADIO –Pilotos) funded by the Ministry of Digital Transformation of the Spanish Government and NextGenerationEU/PRTR.

REFERENCES

[1] 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.

[2] "Mozilla hubs," https://hubs.mozilla.com/, Accessed: April 15, 2024.

[3] Redouane Kachach, Sandra Morcuende, Diego Gonzalez-Morin, Pablo Perez-Garcia, Ester Gonzalez-Sosa, Francisco Pereira, and Alvaro Villegas, "The owl: Immersive telepresence communication for hybrid conferences," in *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 2021, pp. 451–452.

[4] Oliver Schreer, Ingo Feldmann, Sylvain Renault, Marcus Zepp, Markus Worchel, Peter Eisert, and Peter Kauff, "Capture and 3d video processing of volumetric video," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 4310–4314.

[5] Janto Skowronek, Alexander Raake, Gunilla H. Berndtsson, Olli S. Rummukainen, Paolino Usai, Simon N. B. Gunkel, Mathias Johanson, Emanuël A. P. Habets, Ludovic Malfait, David Lindero, and Alexander Toet, "Quality of experience in telemeetings and videoconferencing: A comprehensive survey," *IEEE Access*, vol. 10, pp. 63885–63931, 2022.

[6] Irene Viola, Jack Jansen, Shishir Subramanyam, Ignacio Reimat, and Pablo Cesar, "Vr2gather: A collaborative, social virtual reality system for adaptive, multiparty real-time communication," *IEEE MultiMedia*, vol. 30, no. 2, pp. 48–59, 2023.

[7] 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.

[8] 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.

[9] Javier Usón, Carlos Cortés, Victoria Muñoz, Teresa Hernando, Daniel Berjón, Francisco Morán, Julián Cabrera, and Narciso García, "Untethered real-time immersive free viewpoint video," in *Proceedings of the 16th International Workshop on Immersive Mixed and Virtual Environment Systems*, New York, NY, USA, 2024, MMVE '24, p. 45–49, Association for Computing Machinery.

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

[11] Teresa Hernando, Daniel Berjón, Francisco Morán, Javier Usón, Cesar Díaz, Julián Cabrera, and Narciso García, "Real-time layered view synthesis for free-viewpoint video from unreliable depth information," in *Proceedings of the 15th International Workshop on Immersive Mixed and Virtual Environment Systems*, New York, NY, USA, 2023, MMVE '23, p. 7–11, Association for Computing Machinery.