XV Jornadas de Ingeniería Telemática.
JITEL 2021.
Universidad de A Coruña.

Actas de las XV Jornadas de Ingeniería Telemática (JITEL 2021), A Coruña (España), 27-29 de octubre de 2021.

ISBN

Holo-conferencias 3D multi-usuario: hacia una nueva generación de reuniones virtuales

Sergi Fernández[1], Mario Montagud[1], Gianluca Cernigliaro[1], Marc Martos[1], David Rincón[2]

[1]Media & Internet Area
Fundación i2CAT
C/ Gran Capità 2-4 Edifici Nexus I, Barcelona

[2]Departament d'Enginyeria Telemàtica
Universitat Politècnica de Catalunya (UPC)
Edifici C4C, Castelldefels (Barcelona)

{sergi.fernández; mario.montagud; gianluca.cernigliaro; marc.martos}@i2cat.net; david.rincon@upc.edu

Los sistemas de videoconferencia multi-usuario han ganado mucha relevancia en la sociedad. Aunque estos hayan evolucionado considerablemente en cuanto a escalabilidad, interoperabilidad y funcionalidades de interacción, todavía presentan limitaciones importantes en cuanto a realismo, calidad de interacción y confort, debido principalmente al uso de representaciones de bustos 2D encuadrados en una matriz. Este artículo presenta una plataforma extremo-a-extremo que posibilita una nueva era para las reuniones virtuales mediante la integración en tiempo real de representación volumétricas holográficas de los usuarios en entornos 3D compartidos, mediante tecnologías compatibles a los estándares actuales y utilizando equipamiento de bajo coste. El artículo además reporta sobre resultados obtenidos, tanto a nivel de rendimiento y consumo de recursos para escenarios típicos, como en lo que se respecta a la experiencia de usuario en sesiones de holo-conferencia interactivas de cuatro participantes.

***Palabras Clave*- Holo-Conferencias, Hologramas, QoS, QoE, Realidad Virtual Social, Video Volumétrico**

I. INTRODUCCIÓN

Los avances tecnológicos en las últimas dos décadas han permitido la proliferación de herramientas de comunicación, en tiempo real y multiusuario, como son los sistemas de videoconferencia que permiten reuniones remotas e incluso el tele-trabajo, a un coste mínimo y suponiendo una alternativa real a la presencialidad.

En el campo de la investigación se han realizado esfuerzos considerables con tal de optimizar los sistemas de videoconferencia (ej. reducción de latencias, compresión de video y audio, escalabilidad para sesiones multiusuario…) [1], así como para determinar el impacto sobre la calidad de la experiencia (QoE) que provocan ciertos umbrales en cuanto a parámetros de calidad de servicio (QoS) [2], que en gran medida vienen a determinar el coste del servicio. A su vez, estos sistemas de videoconferencia se han ido complementando con funcionalidades interactivas que mejoran y enriquecen la experiencia compartida [3], como serían la compartición de ficheros y/o pantalla, el ajuste del fondo y composición de la representación 2D de los participantes, o incluso sincronizar la reproducción de contenido compartido.

Más recientemente, la madurez de los sistemas de Realidad Virtual (RV) y su convergencia con los sistemas de audio-videoconferencia están permitiendo el desarrollo de una nueva generación de plataformas de Social VR (ej. Mozilla Hubs, Altspace, Facebook Horizon, Spatial, Glue…) que permiten la interacción y la comunicación entre múltiples usuarios remotos inmersos en entornos virtuales compartidos y, aunque no es condición *sine qua non*, a través de cascos de Realidad Virtual o Aumentada (RV o RA). Gracias a estas plataformas, el lienzo donde se plasma la comunicación ya no está necesariamente restringido a una pantalla rectangular donde se representan múltiples usuarios y contenidos, sino que el entorno (3D) digital con libertad de exploración, o incluso el entorno real en el caso de la RA, se convierte en dicho lienzo, con las nuevas oportunidades y retos que ello supone. Ejemplos claros de estos retos son el método de representación de los usuarios y cómo se interactúa con otros usuarios y el entorno.

En 2020, con la llegada de la pandemia mundial, los sistemas de videoconferencia se han convertido en una herramienta fundamental, extendiendo su uso ampliamente y en ámbitos que trascienden el profesional, permitiendo la socialización en un contexto de distanciamiento social obligatorio. Del mismo modo, las plataformas de Social VR se han ido popularizando y su uso se ha visto acelerado por la necesidad de una mayor y mejor interacción que la que ofrecen hoy en día los sistemas de videoconferencia 2D. A pesar del potencial que estas plataformas pueden ofrecer para una comunicación más natural y realista, hay pocos estudios al respecto, y los que hay se enfocan principalmente en aspectos tecnológicos como la trasmisión de contenido volumétrico [4], el consumo de contenidos compartidos en

entornos virtuales [5], o el nivel de identificación con avatares virtuales (Sección II), pero dejando de lado aspectos clave como son la representación de los usuarios y cómo esta afecta al proceso de comunicación entre ellos, cuando este proceso es, precisamente, el principal objetivo en experiencias de Social VR.

Este articulo presenta una evolución de una plataforma Social VR [4, 5] de bajo coste en el cual los usuarios son capturados por una o múltiples cámaras con sensores de profundidad, esto es RGB-D (ej. Azure Kinect), e integrados en tiempo real y en un formato fotorrealista y volumétrico (nubes de puntos 3D) en entornos virtuales compartidos. Tras presentar los componentes tecnológicos que componen la plataforma, se evalúa su uso cuando se utilizan representaciones de usuarios capturados por una simple cámara RGB-D tanto a nivel QoS como QoE en sesiones interactiva en grupos de cuatro participantes.

En la Figura 1 se representa un ejemplo de reunión virtual por holo-conferencia utilizando la plataforma Social VR desarrollada, así como del equipamiento utilizado por usuarios en los tests realizados. Los resultados obtenidos sirven para determinar los requisitos y costes de implantación de este tipo de servicios, así como demostrar su potencial impacto debido al alto interés que suscitan y a los prometedores niveles de (co-)presencia y calidad de interacción que este nuevo *medium* de comunicación es capaz de proporcionar.

II. Estado del Arte

Las soluciones de Social VR pretenden proporcionar experiencias similares a las reuniones presenciales, tratando de maximizar la sensación de presencia, co-presencia, así como la calidad de la interacción y la plausibilidad del conjunto. En [6] se analizan los factores que tienen impacto en la sensación de identificación con un avatar (*embodiment*). En [7] se demuestra como una sincronización entre acciones del propio cuerpo y su representación mediante un avatar incrementa la sensación de presencia en un entorno virtual. En [8] se demuestra que la representación mediante avatares realistas mejora la sensación de *embodiment* y presencia, quedando confirmado en otros estudios posteriores (ej. [9, 10]). El estudio en [11] aporta evidencias sobre un incremento de presencia, emoción y reconocimiento de los usuarios cuando se usan avatares a partir de reconstrucciones 3D realistas obtenidas en tiempo real, en comparación al uso de avatares sintéticos animados. Asimismo, en [12] se aportan evidencias preliminares sobre los potenciales beneficios que pueden aportar las capturas volumétricas mediante sensores RGB-D y cascos de RV en el ámbito de las experiencias de consumo multimedia compartidas. En [13] se presenta una plataforma Social VR en la que los usuarios se representan en formato vídeo RGB-D mediante el uso de cámaras únicas, y se integran en un entorno estático formado por una imagen 360º. En [14] se desarrolla un cuestionario para experiencias Social VR y se utiliza para comparar experiencias de visionado de fotos compartidas en grupos de 2 usuarios, utilizando tres condiciones de test: a) escenario físico real; b) uso de Skype; y c) plataforma Social VR desarrollada en [13]. Se concluye que el uso de Social VR con representaciones realistas de los usuarios mejora la experiencia de usuario (presencia, co-presencia y calidad de interacción) en comparación al uso de Skype, así como proporciona experiencias comparables a entornos físicos. En [5] se confirman estos beneficios en un escenario de visionado de vídeo compartido, también en grupos de 2 personas, cuando se utiliza Facebook Spaces (plataforma Social VR en la que los usuarios se representan como avatares) como condición de test de comparación. Asimismo, en [5] se presenta una versión evolucionada de la plataforma en [13], incorporando representaciones volumétricas de los usuarios, así como entornos virtuales compartidos en 3D, tras haber realizado un análisis del estado del arte y haber comprobado que existen únicamente un par de plataformas con dichas características, pero requieren equipamiento complejo y caro, y no tienen un soporte multiusuario claro. Finalmente, en dicho estudio se demuestra los potenciales beneficios que esta tecnología de holo-portación puede aportar en una gran variedad de casos de uso. Finalmente, en [4] se presenta una versión evolucionada de la plataforma en [5] con soporte para más de 2 usuarios, y para representaciones volumétricas de los usuarios mediante nubes de puntos (*Point Clouds*). Esta versión de la plataforma Social VR se adopta en este trabajo para investigar el potencial y nivel de madurez de esta tecnología puntera y de bajo coste para posibilitar servicios de holo-portación (esto es, tele-transportación holográfica) multiusuario en tiempo-real, no centrándose únicamente en entornos de consumo de contenidos compartidos como se ha hecho en trabajos anteriores, sino poniendo el foco en la calidad de los usuarios y de la comunicación/interacción entre los mismos.

Fig. 1. Escenarios de holo-conferencia desarrollados y evaluados en el artículo: a) captura de pantalla de sala de reuniones virtual con cuatro participantes; b) y c) equipamiento por cada usuario, con sistema de captura con cámara RGB-D única y casco RV.

Actas de las XV Jornadas de Ingeniería Telemática (JITEL 2021), A Coruña (España), 27-29 de octubre de 2021.

ISBN

III. PLATAFORMA SOCIAL VR

Esta sección presenta la plataforma desarrollada a partir de [4] y [5]. A diferencia de dichos trabajos previos, la plataforma que se presenta soporta más de dos usuarios capturados en tiempo real e insertados en el mismo entorno virtual compartido, amplía el abanico de tipologías de representaciones de usuarios soportadas (aunque este artículo se centre en nubes de puntos capturadas por sensores RGB-D únicos), permite la ingesta y reproducción de contenidos pre-grabados y en vídeo, y soporta consumo en modo pantalla 2D y en cascos RV. Además, incorpora una gestión de sesiones concurrentes y una gestión de eventos que permite la interactividad con el entorno más allá de la visualización de contenidos. En la Figura 2 se muestra un diagrama aproximado sobre la arquitectura del sistema. Las siguientes subsecciones describen más en detalle sus componentes principales.

A. Captura y transmisión de nubes de puntos

El cuerpo de los usuarios se captura mediante 1 o varias cámaras RGB-D (ej. Kinect Azure), basándose en el sistema presentado en [4], que convierte las diferentes imágenes de color y profundidad capturadas por el sensor RGB-D en una nube de puntos 3D (fusionada, en el caso de utilizar sistemas de captura de varios sensores). Por un lado, la nube de puntos se renderiza en el reproductor (Sección III.C) que se ejecuta en el mismo entorno local que el sistema de captura para posibilitar la representación del usuario local y que, por tanto, se vea su propio cuerpo (*self-representation*). Por otro lado, la nube de puntos se comprime usando el códec propuesto en [15] y se transmite vía un Orquestador (Sección III.B), utilizando bien *Dynamic Adaptive Streaming over HTTP* (DASH) o una comunicación basada en sockets mediante *socket.io*. El códec adoptado [15] permite la compresión de nubes de puntos usando la ocupación de los nodos en *octrees* para representar la geometría y proyectando el color de cada nodo en un *grid* 2D. A pesar de que el códec permite explotar las inter-dependencias temporales entre tramas, en este trabajo se utilizan únicamente tramas *intra* para no comprometer la latencia del sistema.

B. Orquestador

Los componentes de orquestación son comúnmente usados en sistemas de videoconferencia para gestionar sesiones y flujos de datos [16]. El orquestador desarrollado maneja la información relativa a cada participante de la sesión, así como de los flujos de datos que intervienen en la misma (audio, video 2D / volumétrico, eventos, etc.), actuando como un reflector. Asimismo, se encarga de notificar sobre cambios en el estado de la sesión (ej. cambios de posición). Finalmente, el orquestador se encarga de gestionar la distribución interactiva de los contenidos asociados a cada sesión (ej. un video a presentarse en una pantalla virtual en el entorno 3D).

C. Reproductor Multimedia

El reproductor (*player*) de la experiencia RV se ha desarrollado en Unity (ejecutable Windows) y es el responsable de la interacción con el Orquestador, instanciar los sistemas de captura y de transmisión/recepción, así como de representar todos los flujos que componen la experiencia adecuadamente, incluyendo el entorno virtual compartido, que puede estar alojado en el mismo *player* o bien descargarse de un servidor de contenidos al inicio de la sesión.

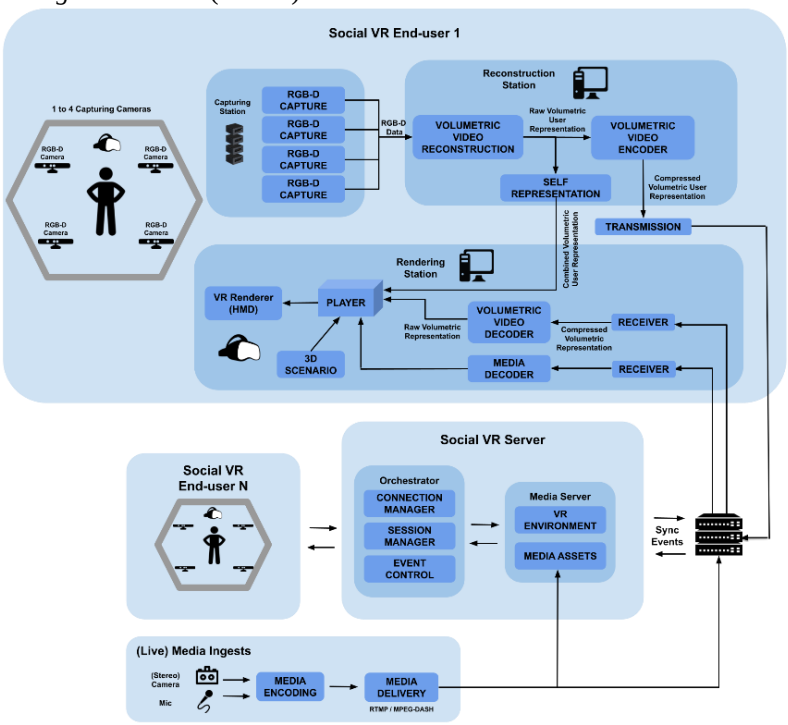

Fig. 2. Arquitectura y diagrama de flujo de alto nivel de la plataforma Social VR desarrollada

IV. Evaluación y Resultados

Esta sección reporta sobre resultados de evaluaciones objetivas (QoS) y subjetivas (QoE) realizadas, con el objetivo de determinar los requisitos y rendimiento de la plataforma, así como la receptividad, calidad percibida e interés en estos escenarios, respectivamente.

Las pruebas reportan sobre experimentos con grupos de 4 usuarios, capturados por cámaras RGB-D únicas, a 15 fps y aproximadamente 50000 puntos por trama, y utilizando *socket.io* para su distribución. En cuanto a los parámetros QoS, se midieron en PCs con los siguientes recursos: CPU Intel(R) Core(TM) i7-10750H @ 2.60GHz, 16GB de RAM y una GPU NVIDIA GeForce RTX 2070.

A. Evaluación Objetiva

En primer lugar, se midió el uso recursos computacionales en uno de los PC clientes (a temperatura ambiente), utilizando un herramienta desarrollada en [17]: CPU: 22.86 %; GPU: 34.53 %; y RAM: 630.11 MB. Para el número fijado de puntos por trama y tramas por segundo por cada usuario, se estima que un PC sin gráfica dedicada suportaría hasta dos usuarios en un escenario simple. En segundo lugar, se midió el consumo de ancho de banda correspondiente a los flujos de nubes puntos. Para los parámetros fijados, cada flujo consumió alrededor de 6.2 Mbps (stdv=1.1 Mbps). En tercer lugar, se midió el retardo extremo-a-extremo (esto es, desde captura hasta renderizado) para cada flujo de nubes puntos, en un escenario en el Orquestador está desplegado en una ciudad situada a 40Km de distancia de los PCs de los clientes finales. Para los parámetros fijados, el retardo para cada flujo fue de 211.22 ms (stdv=10.3 ms). Estos resultados se refieren al promedio para 5 sesiones de unos 8 minutos.

B. Evaluación Subjetiva

Se realizaron tests en grupos de 4 usuarios (*N=32* participantes, edad entre 18-40 años, 17 mujeres), integrándolos en una sala de reuniones virtual, equi-espaciados alrededor de una mesa redonda (Figura 1). Para estimular la interacción, se les instruyó para que realizasen tareas gamificadas, como juegos de adivinanzas (oficios, ciudades, películas…) utilizando gestos no-verbales. La duración de cada sesión fue de 8-10 minutos.

Se adoptó la metodología de evaluación propuesta en [14, 5], incluyendo cuestionarios sobre (co-)presencia, calidad de interacción, así como sobre usabilidad, mareo, cansancio y dificultad. También se realizaron entrevistas con los usuarios. Por motivos de extensión, este artículo se centra en los resultados obtenidos para cuestionarios adicionales diseñados específicamente para capturar la percepción y opinión de los usuarios en cuanto a la calidad audiovisual, comparación con experiencias reales, así como a potencial e impacto de la tecnología que se presenta. Las Tablas I-III reportan los resultados obtenidos, que son muy satisfactorios y prometedores.

V. CONCLUSIONES Y TRABAJO FUTURO

Este trabajo ha presentado una versión evolucionada de una plataforma Social VR y evaluado su potencial para / en escenarios de holo-conferencia o reuniones virtuales multiusuario. A pesar de margen de mejora (ej. en cuanto a la calidad de la reconstrucción gráfica, la falta de un volumen completo en sistemas de captura de cámara RGB-D única, escalabilidad…), los usuarios se han mostrado muy satisfechos con la experiencia, reportando alto niveles de (co-)presencia y calidad de interacción, y sin experimentar mareos ni cansancio. Además, el sistema se comporta de forma satisfactoria y robusta, evidenciando madurez tecnológica.

En cuanto a la experiencia de usuario, en futuros trabajos cabe esperar una comparación con condiciones *baseline* (reunión física, solución de videoconferencia tradicional u otras plataformas de Social VR). Aun así, esta es una primera aproximación que demuestra el potencial de este *medium* es escenarios multi-usuario donde el foco no está tanto en el entorno o en un consumo compartido, sino en los usuarios mismos y en la interacción entre los mismos. A pesar de que en el presente trabajo se ha utilizado un sistema de captura sencillo basado en una sola cara RGB-D frontal, los resultados demuestran que esta configuración puede ser suficiente para entornos en los que los usuarios no se mueven libremente por el entorno, incluso cuando estos están localizados en vistas laterales.

Como trabajo futuro, también se plantea la mejora de las prestaciones del sistema, a varios niveles. Primero, se persigue incrementar la calidad de la representación visual de los usuarios. Segundo, se persigue incrementar la escalabilidad del sistema. Tercero, se pretende habilitar sesiones compartidas con usuarios utilizando diferentes formatos de representación y tipos de dispositivos..

Tabla I
Calidad Audiovisual (1=muy mal; 5=Excelente)

Pregunta / Puntuación	1	2	3	4	5
The visual quality of the virtual scenario	-	-	1	22	9
The visual quality of my representation	-	2	13	15	2
The visual quality of the representation of the user(s) next to me	-	4	16	12	-
The visual quality of the representation of the user(s) in front of to me	-	-	7	21	4
The audio quality from the user(s)	-	-	-	20	12

Tabla II
Comparación a Escenario Real (1=Mucho Peor; 3=Igual; 5=Mucho Mejor)

Pregunta / Puntuación	1	2	3	4	5
The overall virtual experience with one in real life	-	2	19	3	1
The visual representation of users in the virtual experience compared to a real scenario	2	19	11	-	-
The audio quality in the virtual experience compared to the one in a real scenario	-	5	20	7	-
The naturalness of the gestures in the virtual scenario, compared to a real scenario	-	8	22	2	-
The overall communication quality in the virtual scenario, compared to a real scenario	-	5	19	7	1

Tabla III
Potencial e Impacto (1=Totalmente en Desacuerdo; 5=totalmente de Acuerdo)

Pregunta / Puntuación	1	2	3	4	5
This system is effective to hold virtual meetings	-	-	-	6	26
The quality of the users' representation is enough to enable effective and comprehensive interactions and collaborations	-	1	1	21	9
I would use a system like this one for meetings and collaborative tasks in virtual scenarios	-	-	-	10	22
These kind of systems can contribute to a more sustainable environment	-	-	5	12	15

AGRADECIMIENTOS

Este trabajo ha sido financiado parcialmente por el programa H2020 de la Unión Europea, en el marco del proyecto VR-Together (ID 762111) y por ACCIÓ (RIS3CAT, Generalitat de Catalunya), en el marco del proyecto VIVIM (Ref. COMRDI18-1-0008). Asimismo, el trabajo de Mario Montagud ha sido financiado por una Beca JdC-Incorporación (MICINN, IJCI-2017-34611).

REFERENCIAS

[1] S. Firestone, T. Ramalingam, S. Fry, "Voice and Video Conferencing Fundamentals", Cisco Press, 397 pages, March 2007, ISBN-10: 1-58705-268-7

[2] M. Schmitt, J. Redi, D.C.A. Bulterman, P. Cesar, "Towards individual QoE for multi-party video conferencing", IEEE Transactions on Multimedia (TMM), 20(7):1781-1795, 2018

[3] F. Boronat, M. Montagud, P. Salvador, J. Pastor, "Wersync: A web platform for synchronized social viewing enabling interaction and collaboration", Journal of Network and Computer Applications, Volume 175, February 2021.

[4] J. Jansen, S. Subramanyam , R. Bouqueau, G. Cernigliaro, M. Martos, F. Pérez, P. Cesar, "A Pipeline for Multiparty Volumetric Video Conferencing: Transmission of Point Clouds over Low Latency DASH", ACM MMSys 2020, Istanbul (Turkey), June 2020

[5] M. Montagud, J. Li, G. Cernigliario, A. El Ali, S. Fernandez, P. Cesar, "Towards SocialVR: Evaluating a Novel Technology for Watching Videos Together", ArXiv (under review in Virtual Reality (Springer)), arXiv:2104.05060, April 2021

[6] K. Kilteni, R. Groten, M. Slater, "The sense of embodiment in virtual reality", Presence: Teleoperators and Virtual Environments, 21, 4 (2012), 373–387.

[7] P. Heidicker, E. Langbehn, F. Steinicke, "Influence of avatar appearance on presence in Social VR", IEEE 3DUI 2017, 233–234

[8] D. Roth, et al, "Avatar realism and social interaction quality in virtual reality", IEEE VR 2016, 277–278

[9] H. J. Smith, M. Neff, "Communication behavior in embodied Virtual Reality", ACM CHI 2018, 289

[10] T. Waltemate, et al., "The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response", IEEE transactions on visualization and computer graphics, 24(4), 1643–1652, 2018

[11] R. Mekuria, P. Cesar, I. Doumanis, A. Frisiello, "Objective and subjective quality assessment of geometry compression of reconstructed 3D humans in a 3D virtual room", Proc. SPIE 9599, Applications of Digital Image Processing XXXVIII, Vol. 95991, September 2015.

[12] M. McGill, J. H. Williamson, S. Brewster, "Examining the role of smart TVs and VR HMDs in synchronous at-a-distance media consumption", ACM TOCHI, 23, 5, 33, 2016

[13] S. Gunkel, M. Prins, H. Stokking, O. Niamut, "Social VR Platform: Building 360-degree Shared VR Spaces", ACM TVX 2017, Hilversum (The Netherlands), June 2017

[14] J. Li, et al., "Measuring and Understanding Photo Sharing Experiences in Social Virtual Reality", ACM CHI 2019, Glasgow (UK), May 2019

[15] R. Mekuria, K. Blom, P. Cesar, "Design, Implementation, and Evaluation of a Point Cloud Codec for Tele-Immersive Video", IEEE Transactions on Circuits and Systems for Video Technology", 27(4), 828–842, 2017

[16] W. Weiss, R. Kaiser, M. Falelakis, "Orchestration for Group Videoconferencing: An Interactive Demonstrator", ACM ICMI '14, Istanbul (Turkey), 2014

[17] M. Montagud, J. Antonio De Rus, R. Fayos-Jordán, M. Garcia-Pineda J. Segura-Garcia, "Open-Source Software Tools for Measuring Resources Consumption and DASH Metrics", ACM MMSYS 2020, Istanbul (Turkey), June 2020