

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE  
INGENIEROS DE TELECOMUNICACIÓN



UNDERSTANDING AND ASSESSING  
QUALITY OF EXPERIENCE IN IMMERSIVE  
COMMUNICATIONS

Ph.D. Thesis

MARTA ORDUNA CORTILLAS

MASTER OF SCIENCE IN TELECOMMUNICATION ENGINEERING

2023



**Departamento:** SEÑALES, SISTEMAS Y RADIOCOMUNICACIONES  
Escuela Técnica Superior de Ingenieros de Telecomunicación

**Título:** UNDERSTANDING AND ASSESSING QUALITY OF EXPERIENCE  
IN IMMERSIVE COMMUNICATIONS

**Autora:** MARTA ORDUNA CORTILLAS  
Máster Universitario en Ingeniería de Telecomunicación

**Director:** NARCISO GARCÍA SANTOS  
Doctor Ingeniero de Telecomunicación

**Director:** PABLO PÉREZ GARCÍA  
Doctor Ingeniero de Telecomunicación

2023



TESIS DOCTORAL

**Understanding and Assessing Quality of Experience in Immersive Communications**

Autora: Marta Orduna Cortillas

Director: Narciso García Santos

Director: Pablo Pérez García

Tribunal nombrado por el Mgfco. y Excmo. Sr. Rector de la Universidad Politécnica de Madrid, el día . . . . de . . . . . de 2023.

Presidente: D. . . . .

Vocal: D. . . . .

Vocal: D. . . . .

Vocal: Dña. . . . .

Secretario: Dña. . . . .

Realizado el acto de defensa y lectura de la Tesis el día . . . . de . . . . . de 2023 en . . . . .

Calificación: . . . . .

EL PRESIDENTE

LOS VOCALES

EL SECRETARIO



# Agradecimientos

*A todas las personas que habéis compartido*

*esta etapa o parte de ella,*

*gracias.*

*Ha sido un verdadero placer aprender de vosotras.*



---

# Resumen

El desarrollo del hardware y software que hacen posible las tecnologías de Realidad Extendida (XR, del inglés *eXtended Reality*), también llamada Realidad Mixta (MR, del inglés *MiXed Reality*), es constante, mejorando las experiencias ofrecidas a las personas usuarias. Uno de los grandes avances en XR fue la introducción de información visual real en el entorno virtual, consiguiendo que la interacción con la escena fuera más natural y haciendo que aumentara la aceptación de estas tecnologías XR. Posteriormente, aparecieron los vídeos de 360 grados u omnidireccionales que cubren toda la escena. Estos vídeos son grabados con cámaras con lentes omnidireccionales que cubren los 360 grados de la escena para ser visualizados, mayoritariamente, con gafas de realidad virtual (HMD, del inglés *Head-Mounted Display*). Los HMD permiten que únicamente se vea una parte de la escena que cambia sincrónamente con los movimientos de la cabeza.

Esta tesis va un paso más allá y considera una comunicación en la que el vídeo de 360 grados se captura y transmite en tiempo real. Preveemos que este tipo de comunicación será una realidad en la sociedad en las próximas dos décadas. Nuestro objetivo es investigar la tecnología que podría hacer esto posible y diseñar una metodología de evaluación que sea escalable. Para hacer efectivas las experiencias inmersivas, es necesario garantizar una Calidad de Experiencia (QoE, del inglés *Quality of Experience*) aceptable, definida como el nivel de agrado o molestia de la persona usuaria con una aplicación o servicio. En base a este contexto, esta tesis presenta una investigación transversal que busca evaluar aspectos técnicos y socioemocionales en el paradigma de las comunicaciones inmersivas, en concreto, con vídeos de 360 grados.

La investigación sigue un recorrido incremental. Partiendo de una configuración de referencia, se va modificando y evaluando para comprender los desafíos de las tecnologías XR en términos de evaluación de QoE. Asimismo, teniendo la información visual y, por tanto, la calidad del vídeo como uno de los factores más influyentes sobre la QoE, validamos una de las métricas objetivas más robustas desarrollada y utilizada sobre contenidos 2D, Video Multimethod Assessment Fusion (VMAF), en vídeos de 360 grados. Para evaluar la calidad de vídeo en pruebas de evaluación subjetivas, validamos una metodología, Stimulus Discrete Quality Evaluation (SSDQE), con el objetivo de aumentar la validez ecológica de los experimentos ya que permite su aplicación con contenidos de larga duración que dan cabida a una narrativa y contexto. Gracias a la posibilidad de que haya una narrativa y un contexto, validamos que no únicamente permite la evaluación de aspectos técnicos, sino que también permite la evaluación de aspectos socioemocionales que afectan a la QoE.

El sistema de comunicación inmersivo se explora principalmente desde la perspectiva de la persona que está en remoto, con conclusiones de bajo nivel como por ejemplo, la capacidad de visualizar las manos o la naturalidad de utilizar el táctil del HMD o los mandos para interactuar con el entorno virtual durante la experiencia. De la misma manera, presentamos conclusiones de más alto nivel, como por ejemplo, en relación a la perspectiva de adquisición del contenido. Mediante el diseño y realización de pruebas subjetivas en las que se simula la comunicación hasta en pruebas con comunicación interactiva, evaluamos diferentes factores que influyen sobre la QoE, mejorando su comprensión. Comprender el paradigma de la evaluación de QoE nos permite presentar una guía de buenas prácticas para diseño de experimentos. Lo presentamos desde un marco común para

los dos puntos de vista principales de la evaluación de QoE, las telecomunicaciones y el área de interacción hombre-máquina (HCI, del inglés *Human Computer Interaction*), de manera que sea una herramienta útil para personas con distinto perfil y experiencia.

Como caso de uso de aplicación, se analiza el caso de tele-educación. Al prototipo de referencia le añadimos un módulo de análisis de vídeo para detectar eventos de interés en la escena de 360 grados para que sean notificados a los/as estudiantes que siguen la clase en remoto y guiar así su atención. Según el análisis realizado, las notificaciones y el sistema como solución para tele-educación tienen una gran aceptación por parte de los/as estudiantes. Además, se ha hecho pública una base de datos de vídeos de 360 grados que hemos generado y anotado a partir de clases reales para que pueda utilizarse en el entrenamiento de algoritmos de aprendizaje máquina y en pruebas de evaluación subjetiva.

Esta tesis es una contribución para entender el paradigma de las comunicaciones inmersivas para que gracias a su continuo desarrollo y evaluación lleguen a ser una realidad en la sociedad.

---

# Abstract

eXtended Reality (XR) technology, also called Mixed Reality (MR), is in constant development and improvement in terms of hardware and software to offer relevant experiences to users. One of the advances in XR has been the introduction of real visual information in the virtual environment, offering a more natural interaction with the scene and a greater acceptance of technology. Another advance has been achieved with the representation of the scene through a video that covers the entire environment, called 360-degree or omnidirectional video. These videos are acquired by cameras with omnidirectional lenses that cover the 360-degrees of the scene and are generally viewed by users through a head-tracked Head Mounted Display (HMD). Users only visualize a subset of the 360-degree scene, called viewport, which changes with the variations of the viewing direction of the users, determined by the movements of the head.

This thesis goes one step further and considers a real-time 360-degree video communication for teleconferencing purposes. We envision that this kind of communication will become mainstream within the next couple of decades. Our target is to research the technology that could make this possible and design a proper assessment methodology that scales for massive usage. Therefore, it is necessary to guarantee an acceptable Quality of Experience (QoE), defined as the degree of delight or annoyance of the user with an application or service, to increase the use of immersive communications. Based on this, this thesis presents a cross-sectional research to include the assessment of technical and socioemotional aspects in the 360-degree video communications paradigm.

The research follows an evolutionary approach, modifying different conditions of the reference configuration of a 360-degree video communication prototype to understand the challenges of XR technologies in terms of QoE assessment. Starting from video quality, as a significant factor impacting QoE, we validate the Video Multimethod Assessment Fusion (VMAF) objective metric on 360-degree video, designed and developed for 2D content by Netflix, saving time and resources. To evaluate video quality in subjective assessments, we validate the Stimulus Discrete Quality Evaluation (SSDQE) methodology, which can be used with contents of long duration, allowing narrative. Then, we validate the fact that SSDQE allows the simultaneous evaluation of socioemotional and technical aspects, increasing the ecological validity of the experiments.

The immersive communication system is mainly explored from the perspective of the remote user, with conclusions drawn on low-level (e.g., possibility of visualizing the hands or using the touchpad or the handheld controller to interact with the virtual environment) and high-level of the scenario (e.g., acquisition perspective). By conducting assessments based on both simulated and interactive communications, valuable insights have been concluded. Furthermore, what we have learned about design of experiments is summarized as a best practices guide for developers and researchers. Due to the transversal research, the guidelines are proposed from a common framework for two of the main viewpoints in QoE assessment, telecommunications and human computer interaction areas.

The use case of tele-education is analyzed, including a video analysis module added to detect events of interest around the 360-degree scene and notify them to the remote students, helping to guide their attention. The notifications and the system as solution for tele-education are highly accepted by students. Additionally, we provide a database of 360-degree videos of real lessons with

annotated events of interest, which is publicly available for training machine learning algorithms and subjective assessments.

This thesis is a contribution to understand the paradigm of immersive communications to continue developing and evaluating them until they become a reality in society.

# Contents

|  |            |
|--|------------|
| <b>Resumen</b>   | <b>ii</b>  |
| <b>Abstract</b>  | <b>iv</b>  |
| <b>List of Figures</b>   | <b>ix</b>  |
| <b>List of Tables</b>  | <b>xii</b> |
| <b>Acronyms</b>  | <b>xiv</b> |
| <b>1 Motivation</b>  | <b>1</b>   |
| 1.1 Introduction . . . . .   | 1          |
| 1.2 Research objectives and contributions . . . . .  | 6          |
| <b>2 Objective Video Quality Assessment</b>  | <b>10</b>  |
| 2.1 Introduction . . . . .   | 10         |
| 2.2 Related work . . . . .   | 11         |
| 2.3 Video Multimethod Assessment Fusion (VMAF) on 360-degree videos . . . . .                  | 12         |
| 2.3.1 Video source characterization . . . . .  | 14         |
| 2.3.2 Experimental results . . . . .   | 16         |
| 2.4 Validation of VMAF for 360-degree videos through a subjective quality assessment . . . . . | 17         |
| 2.4.1 Stimuli . . . . .  | 18         |
| 2.4.2 Methodology . . . . .  | 18         |
| 2.4.3 Equipment and environment . . . . .  | 19         |
| 2.4.4 Test session . . . . .   | 19         |
| 2.4.5 Participants . . . . .   | 20         |
| 2.4.6 Experimental results . . . . .   | 20         |
| 2.5 Comparison of VMAF with other objective metrics . . . . .                                  | 23         |
| 2.6 Conclusions . . . . .  | 24         |
| <b>3 Subjective Video Quality and Presence Assessment</b>                                      | <b>26</b>  |
| 3.1 Introduction . . . . .   | 26         |
| 3.2 Related work . . . . .   | 26         |
| 3.3 Experiment design . . . . .  | 27         |
| 3.3.1 Stimuli . . . . .  | 27         |
| 3.3.2 Methodology . . . . .  | 31         |
| 3.3.3 Equipment and environment . . . . .  | 32         |
| 3.3.4 Test session . . . . .   | 32         |
| 3.3.5 Participants . . . . .   | 33         |
| 3.3.6 Hypotheses . . . . .   | 33         |
| 3.4 Experimental results . . . . .   | 33         |
| 3.5 Conclusions . . . . .  | 34         |
| <b>4 Subjective Video Quality and Socioemotional Aspects Assessment</b>                        | <b>36</b>  |
| 4.1 Introduction . . . . .   | 36         |
| 4.2 Related work . . . . .   | 36         |

---

|          |   |           |
|----------|---|-----------|
| 4.3      | Experiment design . . . . .   | 38        |
| 4.3.1    | Research questions . . . . .  | 38        |
| 4.3.2    | Experimental conditions . . . . .   | 39        |
| 4.3.3    | Stimuli . . . . .   | 40        |
| 4.3.4    | Methodology . . . . .   | 41        |
| 4.3.5    | Equipment and environment . . . . .   | 44        |
| 4.3.6    | Test session . . . . .  | 45        |
| 4.3.7    | Participants . . . . .  | 46        |
| 4.4      | Experimental results . . . . .  | 46        |
| 4.5      | Conclusions . . . . .   | 53        |
| <b>5</b> | <b>Interactive Communication Assessment</b>                                       | <b>55</b> |
| 5.1      | Introduction . . . . .  | 55        |
| 5.2      | Related work . . . . .  | 56        |
| 5.3      | Experiment design . . . . .   | 57        |
| 5.3.1    | Research questions . . . . .  | 57        |
| 5.3.2    | Equipment . . . . .   | 59        |
| 5.3.3    | Experimental conditions . . . . .   | 59        |
| 5.3.4    | Stimuli . . . . .   | 60        |
| 5.3.5    | Experiment setup . . . . .  | 61        |
| 5.3.6    | Methodology . . . . .   | 62        |
| 5.3.7    | Test session . . . . .  | 64        |
| 5.3.8    | Participants . . . . .  | 65        |
| 5.4      | Experimental results . . . . .  | 65        |
| 5.4.1    | Quantitative results . . . . .  | 65        |
| 5.4.2    | Qualitative results . . . . .   | 69        |
| 5.5      | Discussion . . . . .  | 71        |
| 5.6      | Conclusions . . . . .   | 74        |
| <b>6</b> | <b>Tele-education Use Case</b>  | <b>76</b> |
| 6.1      | Introduction . . . . .  | 76        |
| 6.2      | Prototype . . . . .   | 76        |
| 6.3      | EVENT-CLASS database . . . . .  | 78        |
| 6.3.1    | Specifications . . . . .  | 79        |
| 6.4      | Evaluation of the performance of an immersive system for tele-education . . . . . | 80        |
| 6.4.1    | Related work . . . . .  | 80        |
| 6.4.2    | Research questions . . . . .  | 81        |
| 6.4.3    | Experimental conditions . . . . .   | 82        |
| 6.4.4    | Stimuli . . . . .   | 83        |
| 6.4.5    | Methodology . . . . .   | 84        |
| 6.4.6    | Equipment and environment . . . . .   | 86        |
| 6.4.7    | Test session . . . . .  | 86        |
| 6.4.8    | Participants . . . . .  | 87        |
| 6.4.9    | Experimental results . . . . .  | 88        |
| 6.4.10   | Conclusions . . . . .   | 93        |
| <b>7</b> | <b>Lessons Learned for the QoE Assessment of Immersive Communications</b>         | <b>95</b> |
| 7.1      | Introduction . . . . .  | 95        |
| 7.2      | Source . . . . .  | 96        |

---

|          |  |            |
|----------|--|------------|
| 7.3      | Conditions . . . . .                   | 98         |
| 7.3.1    | Continuous conditions . . . . .        | 99         |
| 7.3.2    | Categorical conditions . . . . .       | 101        |
| 7.4      | Feedback . . . . .                     | 102        |
| 7.4.1    | Questionnaires . . . . .               | 102        |
| 7.4.2    | Method of collecting ratings . . . . . | 104        |
| 7.5      | Participants . . . . .                 | 105        |
| 7.6      | Conclusions . . . . .                  | 105        |
| <b>8</b> | <b>Conclusions and Future Work</b>     | <b>108</b> |
| 8.1      | Conclusions . . . . .                  | 108        |
| 8.2      | Future work . . . . .                  | 110        |
| <b>A</b> | <b>Scientific Contributions</b>        | <b>111</b> |
| <b>B</b> | <b>Outreach Activity</b>               | <b>114</b> |
|          | <b>Bibliography</b>                    | <b>116</b> |



# List of Figures

|     |  |    |
|-----|--|----|
| 1.1 | Reality–Virtuality Continuum [1]. . . . .  | 1  |
| 1.2 | Example of a real-time 360-degree video communication environment. Three local participants share the same physical space while there is one remote user who is attending the meeting through an HMD. . . . .  | 2  |
| 1.3 | The degrees of freedom that an HMD can have in the virtual environment, corresponding to movements in X, Y, and Z and rotational movements in pitch, yaw, and roll. . . . .  | 2  |
| 1.4 | Equirectangular projection example. In this figure the peculiarity of this projection in which the sphere of radius r is stretched to fit the XY plane can be observed. . . . .  | 3  |
| 2.1 | Video sources screenshots in equirectangular projection. . . . .   | 14 |
| 2.2 | Spatial (x-axis) and Temporal Information (y-axis) indicators for the 360-degree videos considered in the test. . . . .  | 16 |
| 2.3 | VMAF (y-axis) vs QP (x-axis) curve for all SRCs. The VMAF anchor values used for the validation are represented with solid black lines (y-axis) which correspond to a range of QP values (x-axis). . . . .   | 17 |
| 2.4 | Test session structure. . . . .  | 19 |
| 2.5 | MOS and DMOS (y-axis) on a five-level scale obtained from 23 participants, including CIs. Participants evaluated clips of short duration (10s) encoded with fixed quantization parameters which determined Qualities A, B, B, D, E, and F (x-axis). . . . .  | 21 |
| 2.6 | Evolution of the VMAF scores and the normalized DMOS on a 100-level scale with the associated CI (y-axis) with the QP value for each content (x-axis). . . . .   | 22 |
| 2.7 | Mapping of DMOS ratings (y-axis) to objective scores (x-axis). Solid line represents the best fitting by a third degree polynomial curve. . . . .  | 25 |
| 3.1 | Video sources screenshots in equirectangular projection. . . . .   | 28 |
| 3.2 | Spatial (x-axis) and Temporal Information (y-axis) indicators for all contents. . . . .  | 29 |
| 3.3 | Test session structure. . . . .  | 32 |
| 3.4 | Average of the ratings of the participants of TPI, sPQ, Usability, and video quality (y-axis) with the associated 95% CI on a seven-level scale for each condition (x-axis). . . . .   | 33 |
| 4.1 | Simulated immersive communication environment of the experiment. 360-degree video and audio recorded on the provider side is visualized by the remote client. Participants assigned to condition C can see their hands and take notes on a physical whiteboard (Augmented VR). . . . .   | 39 |
| 4.2 | Participant of condition C taking notes on a physical whiteboard. The photo was taken at the environment of the experiment. . . . .  | 40 |
| 4.3 | Video sources screenshots in equirectangular projection [2]. . . . .   | 41 |
| 4.4 | Structure of the test sequences used with SSDQE methodology. . . . .   | 43 |
| 4.5 | Test session structure. . . . .  | 45 |
| 4.6 | Observers distribution in conditions A, B, and C taking into account the gender and international experience. . . . .  | 46 |
| 4.7 | The mean opinion scores (y-axis) on a five-level scale obtained from 17 participants assigned to condition A who evaluated the perceived video quality in the processed segment, encoded with specific QP, every 20 seconds while visualizing each of the three contents (x-axis), following the SSDQE methodology. Error bars represent 95% CI. . . . . | 47 |

|     |  |    |
|-----|--|----|
| 4.8 | Comparison of DMOS (y-axis) on a five-level scale obtained from 17 participants assigned to condition A, and from participants from the pilot study. Both participants evaluated clips of short duration encoded with fixed quantization parameters (x-axis). However, participants assigned to Condition A evaluated perceived quality following SSDQE methodology and participants from the pilot study evaluated it following ACR methodology. . . . .  | 48 |
| 5.1 | Prototype of 360-degree video communication considered in this experiment. The remote user visualizes the 360-degree of the local side through the HMD. Local participants visualize the synchronized head movements of the remote user in a cartoon avatar on a tablet. Additionally, the lips of the avatar move with the audio. . . . .   | 58 |
| 5.2 | Interactive communication scenarios considered in the experiment based on the experimental conditions: in-person and mixed technology mediated discussion between three participants. . . . .  | 59 |
| 5.3 | Remote participant attending the hybrid meeting through the HMD. Local participants visualizing the remote participant at the tablet (cartoon avatar) and following the indications of the cards on the table. . . . .   | 61 |
| 5.4 | Material used to guide the test session: timer, cards of six colored hats, summary of the representation of the color of each hat, and tasks of the test session. . . . .  | 62 |
| 5.5 | Test session structure. . . . .  | 64 |
| 5.6 | Mean opinion scores (y-axis) on a five-level scale obtained from 27 participants in each of the tasks (x-axis) and presented by experimental conditions. . . . .   | 66 |
| 5.7 | Mean scores (y-axis) of quality, spacial, and social presence (x-axis) collected in this experiment (Hybrid meetings) and collected in the previous test presented in Chapter 4. All ratings were provided by remote participants. Note that quality was rated on a 5-level scale, while social and spatial presence were rated on a 7-level scale. The number of participants of this experiment is not the same of the previous one. . . . .   | 66 |
| 5.8 | Reported emotions (I-PANAS-SF) divided into the experimental conditions: in-person and hybrid meetings. . . . .  | 67 |
| 6.1 | Architecture of the tele-education scenario. . . . .   | 77 |
| 6.2 | Example of accepted notifications of events of interest considered in tele-education scenario. The bounding box frames the student who has been detected raising her hand. At the top appears, the number of notifications waiting to be accepted, an arrow that indicates the direction in which the next event of interest has occurred, and a check that indicates that the notification has been accepted. This is also indicated by the color of the bounding box, yellow if user has not centered the viewport on the event, or green when user has already looked at and therefore accepted the notification. . . . . | 78 |
| 6.3 | Examples of people detection obtained in equirectangular frames. . . . .   | 80 |
| 6.4 | Screenshots of video sources considered in the study. . . . .  | 84 |
| 6.5 | Diagram of the test session structure. Each participant was randomly assigned to two different conditions, so each participant carried out this part of the test twice. Then, the duration of the test session was around 50 minutes. . . . .  | 87 |
| 6.6 | Characterization of the 39 participants of the study in terms of their experience in VR and attending online lessons. . . . .  | 88 |
| 6.7 | The mean opinion scores (y-axis) on a five-level scale obtained from 39 participants, taking into account that each one rated two conditions, who evaluated the perceived video quality at the end of each sequence (x-axis). Error bars represent 95% CI. . . . .   | 89 |

---

|     |  |    |
|-----|--|----|
| 6.8 | The aggregate Spatial Presence and Social Presence (y-axis) on a seven-level scale obtained from 39 participants after the visualization of each sequence (x-axis) in three conditions. Error bars represent 95% CI. . . . . | 91 |
| 6.9 | Simulator Sickness results obtained from the evaluations of the 39 participants at different time points during the test session. . . . .  | 92 |
| 7.1 | Theoretical framework of the stages necessary to design an experiment. . . . .   | 95 |
| 7.2 | Example of the condition-stimulus applied spatially on a 360-degree scene and the condition-stimulus applied temporally. . . . .   | 99 |



# List of Tables

|     |   |    |
|-----|---|----|
| 2.1 | Semantic characterization of the 360-degree videos considered in the test. . . . .  | 15 |
| 2.2 | Characteristics of the dataset of the 360-degree videos selected for the test. . . . .  | 16 |
| 2.3 | Pearson correlation, RMSE and Spearman’s rank correlation between VMAF and DMOS for all contents. . . . .   | 23 |
| 2.4 | Pearson correlation, RMSE, and coefficient of determination $R^2$ of fitting curves and DMOS for all analysed metrics. . . . .  | 24 |
| 3.1 | Semantic characterization of the 360-degree videos considered in the experiment. . .  | 28 |
| 3.2 | Characteristics of the 360-degree videos considered in the test. . . . .  | 29 |
| 3.3 | PSNR, WS-PSNR, CPP-PSNR, VMAF, SSIM, and MS-SSIM results of PVSs used in the work. . . . .  | 30 |
| 3.4 | Overview of the two tests depending on the order of the experimental conditions assessed in the experiment. . . . .   | 32 |
| 4.1 | Overview of the three experimental conditions with the associated interactive element and features assessed in the experiment. . . . .  | 40 |
| 4.2 | Semantic characterization of the 360-degree videos considered in the experiment. . .  | 41 |
| 4.3 | Structure of the test session questionnaires. . . . .   | 42 |
| 4.4 | Attention survey: True/False statement, short answer, and multiple choice question for each of the 360-degree videos. . . . .   | 43 |
| 4.5 | Difference in aggregate quality and socioemotional features between the three conditions. . . . .   | 49 |
| 4.6 | Difference in aggregate quality and socioemotional features between the three contents. . . . .   | 49 |
| 4.7 | Cronbach’s $\alpha$ obtained for the questionnaires used in the experiment about spatial presence, social presence, and attitude. . . . .   | 50 |
| 4.8 | The mean and standard deviations on a seven-level scale of the items of the attitude survey: interest, respect, tolerance, and social sensitivity in the three experimental conditions. . . . .   | 52 |
| 5.1 | Overview of the questionnaire items asked depending on the participant condition and the assigned role. . . . .   | 63 |
| 5.2 | Summary of the results obtained from remote participants in the items: aggregate quality and social and spatial presence. . . . .   | 67 |
| 5.3 | Summary of the aggregate quality and social presence evaluated by local participants in the in-person and hybrid meetings. The results presented are independent of the role of the participants. . . . .                               | 68 |
| 5.4 | Summary of the results obtained from participants with the role of speaker in the questionnaires about: liking, self-other overlap, and perceived empathy. These items were asked in relation to the moderator role (blue hat). . . . . | 68 |
| 6.1 | Technical properties of the stitched sequences recorded with each camera. . . . .   | 79 |
| 6.2 | Main specifications of the video sources considered in the study in terms of the scene (classroom type, acquisition perspective and camera location) and camera specifications (camera, resolution and framerate, and bitrate). . . . . | 83 |
| 6.3 | Number of notification events (raised hands and slide changes) that occur in the video sources considered in the study and number of people in the scene. . . . .   | 84 |

---

|     |   |    |
|-----|---|----|
| 6.4 | Structure of the test session questionnaires divided into three points in time within the test session: pre-questionnaire, before the start of the test; post-seq questionnaire, at the end of the visualization of each sequence; post-cond. questionnaire, at the end of each condition evaluated by the participant. . . . . | 84 |
| 6.5 | Difference in aggregate quality, spatial and social presence, and usability between the three conditions. . . . .   | 88 |
| 6.6 | Notifications scores obtained from the evaluations of the 28 participants assigned to VR+notifications condition at the end of the test session . . . . .   | 92 |

# Acronyms

|                   |  |
|-------------------|--|
| <b>ABR</b>        | <b>Adaptive Bit Rate</b>   |
| <b>ACR</b>        | <b>Absolute Category Rating</b>  |
| <b>ACR-HR</b>     | <b>Absolute Category Rating with Hidden Reference</b>                        |
| <b>AI</b>         | <b>Artificial Intelligence</b>   |
| <b>ANOVA</b>      | <b>ANalysis Of VAriance</b>  |
| <b>AR</b>         | <b>Augmented Reality</b>   |
| <b>AV</b>         | <b>Augmented Virtuality</b>  |
| <b>AVC</b>        | <b>Advanced Video Coding</b>   |
| <b>CI</b>         | <b>Confidence Interval</b>   |
| <b>CNN</b>        | <b>Convolutional Neural Network</b>  |
| <b>CPP-PSNR</b>   | <b>Craster Parabolic Projection PSNR</b>                                     |
| <b>CVE</b>        | <b>Collaborative Virtual Environment</b>                                     |
| <b>DMOS</b>       | <b>Differential Mean Opinion Score</b>                                       |
| <b>DoF</b>        | <b>Degrees of Freedom</b>  |
| <b>EC</b>         | <b>Empathic Concern</b>  |
| <b>FS</b>         | <b>Fantasy Scale</b>   |
| <b>FoV</b>        | <b>Field of View</b>   |
| <b>FR</b>         | <b>Full Reference</b>  |
| <b>GDPR</b>       | <b>General Data Protection Regulation</b>                                    |
| <b>HCI</b>        | <b>Human Computer Interaction</b>  |
| <b>HEVC</b>       | <b>High Efficiency Video Coding</b>  |
| <b>HMD</b>        | <b>Head Mounted Display</b>  |
| <b>HRC</b>        | <b>Hypothetical Reference Circuit</b>  |
| <b>HVS</b>        | <b>Human Visual System</b>   |
| <b>IOS</b>        | <b>Inclusion of Other in the Self</b>  |
| <b>IRI</b>        | <b>Interpersonal Reactivity Index</b>  |
| <b>ITU</b>        | <b>International Telecommunication Union</b>                                 |
| <b>I-PANAS-SF</b> | <b>Internationally Reliable Short-Form Positive Negative Affect Schedule</b> |
| <b>JND</b>        | <b>Just Noticeable Difference</b>  |
| <b>LPCC</b>       | <b>Linear Pearson Correlation Coefficient</b>                                |
| <b>MR</b>         | <b>Mixed Reality</b>   |
| <b>MOS</b>        | <b>Mean Opinion Score</b>  |

---

|                 |  |
|-----------------|--|
| <b>MS-SSIM</b>  | <b>M</b> ulti <b>S</b> cale - <b>SSIM</b>  |
| <b>PD</b>       | <b>P</b> ersonal <b>D</b> istress  |
| <b>PLCC</b>     | <b>P</b> earson's <b>L</b> inear <b>C</b> orrelation <b>C</b> oefficient                     |
| <b>PQ</b>       | <b>P</b> resence <b>Q</b> uestionnaire   |
| <b>PSNR</b>     | <b>P</b> eak <b>S</b> ignal to <b>N</b> oise <b>R</b> atio                                   |
| <b>PT</b>       | <b>P</b> erspective <b>T</b> aking   |
| <b>PVS</b>      | <b>P</b> rocessed <b>V</b> ideo <b>S</b> equence   |
| <b>QoE</b>      | <b>Q</b> uality of <b>E</b> xperience  |
| <b>QoS</b>      | <b>Q</b> uality of <b>S</b> ervice   |
| <b>QP</b>       | <b>Q</b> uantization <b>P</b> arameter   |
| <b>RQ</b>       | <b>R</b> esearch <b>Q</b> uestions   |
| <b>SD</b>       | <b>S</b> tandard <b>D</b> eviation   |
| <b>SSQ</b>      | <b>S</b> imulator <b>S</b> ickness <b>Q</b> uestionnaire                                     |
| <b>sPQ</b>      | subsampling of the <b>P</b> resence <b>Q</b> uestionnaire                                    |
| <b>S-PSNR</b>   | <b>S</b> phere based <b>PSNR</b>   |
| <b>SRC</b>      | <b>SouRCe</b>  |
| <b>SSIM</b>     | <b>S</b> tructural <b>S</b> iMilarity <b>I</b> ndex  |
| <b>SSQ</b>      | <b>S</b> imulator <b>S</b> ickness <b>Q</b> uestionnaire                                     |
| <b>SROCC</b>    | <b>S</b> pearman's <b>R</b> ank- <b>O</b> rded <b>C</b> orrelation <b>C</b> oefficient       |
| <b>SS</b>       | <b>S</b> ingle <b>S</b> timulus  |
| <b>SSCQE</b>    | <b>S</b> ingle <b>S</b> timulus <b>C</b> ontinuous <b>Q</b> uality <b>E</b> valuation        |
| <b>SSDQE</b>    | <b>S</b> ingle <b>S</b> timulus <b>D</b> iscrete <b>Q</b> uality <b>E</b> valuation          |
| <b>ST-VMAF</b>  | <b>S</b> patio <b>T</b> emporal- <b>VMAF</b>   |
| <b>SUS</b>      | <b>S</b> ystem <b>U</b> sability <b>S</b> cale   |
| <b>TPI</b>      | <b>T</b> emple <b>P</b> resence <b>I</b> nventry   |
| <b>USS-PSNR</b> | <b>U</b> niformly <b>S</b> ampled <b>S</b> pherical <b>PSNR</b>                              |
| <b>UTAUT</b>    | <b>U</b> nified <b>T</b> heory of <b>A</b> cceptance and <b>U</b> sage of <b>T</b> echnology |
| <b>VDK</b>      | <b>VMAF</b> <b>D</b> evelopment <b>K</b> it  |
| <b>VMAF</b>     | <b>V</b> ideo <b>M</b> ultimethod <b>A</b> ssessment <b>F</b> usion                          |
| <b>VQEG</b>     | <b>V</b> ideo <b>Q</b> uality <b>E</b> xperts <b>G</b> roup                                  |
| <b>VQM</b>      | <b>V</b> ideo <b>Q</b> uality <b>M</b> etric   |
| <b>VR</b>       | <b>V</b> irtual <b>R</b> eality  |
| <b>RMSE</b>     | <b>R</b> oot <b>M</b> ean <b>S</b> quare <b>E</b> rror                                       |
| <b>WS-PSNR</b>  | <b>W</b> eighted to <b>S</b> pherically <b>PSNR</b>  |
| <b>XR</b>       | <b>e</b> Xtended <b>R</b> eality   |

# Chapter 1

## Motivation

### 1.1 Introduction

**EXtended Reality (XR)** technology, also called Mixed Reality (MR), is an umbrella term referring to technologies that offer an alternate view of reality [3]. It can be understood as a continuous scale from the real world to a fully synthetic world, as presented in Figure 1.1 [1]. Following the figure, XR can be a combination of Augmented Reality (AR), Augmented Virtuality (AV), and/or Virtual Reality (VR). AR allows the integration of virtual objects in reality. AV allows the integration of real objects into the virtual environment. VR creates an immersive environment around users, allowing the interaction with the objects in it.

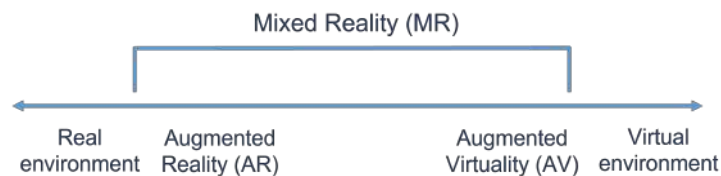


FIGURE 1.1: Reality–Virtuality Continuum [1].

Initially, XR generated great interest in the entertainment sector. However, it has evolved a lot since then, offering better visualization conditions and more affordable consumer devices. From there, use cases have been explored towards other areas, such as education [4–6], health [7, 8], industry [9–12], or tourism [13], whose impact on day-to-day life can be even greater. In fact, XR is expected to change the way people communicate remotely [14], replacing or complementing video conferencing as it is today. The basic idea is to convey a more realistic experience, making the interaction more similar to face-to-face. The potential lies in supporting broader nonverbal communication, defined as the behavior that goes beyond linguistic language, such as the gaze, body language, and spatial distance [15, 16].

This work follows one of the advances of XR technologies, the representation of the scene through a video acquired by omnidirectional cameras that cover the 360 degrees of the scene, called *omnidirectional*, *360VR*, or *360-degree* video. It is acquired and transmitted in real-time, allowing a real-time 360-degree video communication [17, 18].

In this work, we consider a communication scenario, such as the one presented in Figure 1.2, where unidirectional real-time 360-degree video is sent from the local participants (onsite) to the remote participant, who attends the conversation with a Head Mounted Display (HMD). Also, bidirectional audio is sent between them. The peculiarity here, compared to other XR communication solutions such as social VR [18], is that the remote attendant visualizes through a head-tracked HMD a realistic virtual environment.

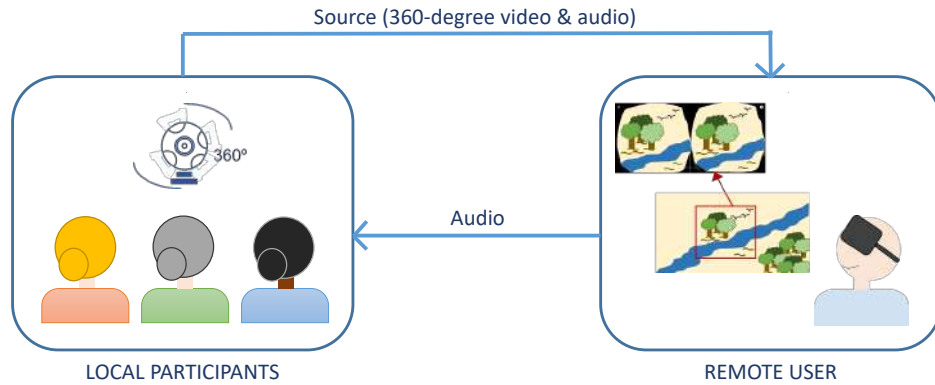


FIGURE 1.2: Example of a real-time 360-degree video communication environment. Three local participants share the same physical space while there is one remote user who is attending the meeting through an HMD.

The HMD allows the visualization of a subset of the 360-degree scene, called *viewport*, and this viewport varies with the viewing direction of the remote user [19]. As presented in Figure 1.3, if the HMD has three Degrees of Freedom (DoF), the viewport changes with the rotational movements of the user’s head, interpreted as the Euler angles, *yaw*, *pitch*, and *roll*. Although not used in 360-degree video, the HMD can have six DoFs. In this case, the viewport changes with the three rotational movements and the three movements in *X-axis* (*left/right*), *Y-axis* (*up/down*), and *Z-axis* (*forward/backward*).

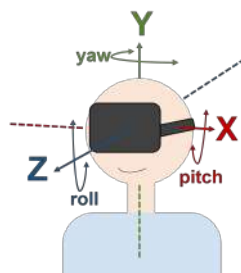


FIGURE 1.3: The degrees of freedom that an HMD can have in the virtual environment, corresponding to movements in X, Y, and Z and rotational movements in pitch, yaw, and roll.

The 360-degree cameras used to capture the scene are composed of at least two lenses. The lenses cover the 360-degree scene with enough overlap to provide an acceptable *stitching*, which is in charge of converting synchronized streams from the lenses into a panoramic video. Then, the 360-degree

scene is mapped into a planar format. The most common mapping and used in this research is the *quirectangular projection* [20]. Figure 1.4 presents an example of a frame in equirectangular projection. Once the content is in this projection, it is encoded and prepared to be transmitted. The standardized encoders: ITU-T Rec. H.265/High Efficiency Video Coding (HEVC) [21] and ITU-T Rec. H.264/MPEG-4 Part 10 Advanced Video Coding (AVC) [22] are used to encode 360-degree videos. The encoding can be controlled with the *Quantization Parameter (QP)* or the *bitrate*. Although it is explained in detail later, the relationship between them is that a higher QP implies a lower bitrate and a lower quality while a lower QP implies a higher bitrate and a higher quality.



FIGURE 1.4: Equirectangular projection example. In this figure the peculiarity of this projection in which the sphere of radius  $r$  is stretched to fit the  $XY$  plane can be observed.

Relevant technical limitations found in this scenario are associated with the encoding and transmission of high-quality 360-degree video [23]. The pixels are distributed in a 360-degree scene, which implies higher bitrates than 2D video to offer an acceptable quality [19]. Despite the great effort of the research community to offer the best quality to remote users, it is still a limitation if we compare it with the quality perceived on a screen.

Since **video quality** is a crucial sensory feedback to create the impression of a realistic virtual environment, and therefore a crucial aspect and a fundamental pillar of this work, measuring it is the first goal tackled. As with 2D content, the evaluation of video quality can be divided into two main groups: video quality objective metrics and subjective quality assessments. Objective quality metrics provide video quality scores that should estimate the quality perceived by users. The success of this estimation is achieved if the scores obtained from one content with an objective metric correlate with the ratings provided by users for the same content. For the implementation of objective metrics, psychophysical, engineering designs, or a combination of both can be used. Psychophysical design models components of the Human Visual System (HVS) and engineering design take into account the measurable distortions on the content and the impact on the perception by users [24]. On the other hand, subjective video quality assessments are based on test sessions in which participants visualize videos and rate the perceived quality [25]. The International Telecommunication Union (ITU) provides recommendations to standardize objective and subjective assessments. These recommendations define detailed procedures to execute the assessments,

including characterization of the source contents and the impairments, structure of the test sessions, rating scales, and statistical analysis methods for the results, among others.

The ITU recommendations evolve hand in hand with technology and research, and this has two main implications. In terms of technology, recommendations continually adapt to new challenges: ITU-R Rec. BT.500-14 (television) [25] or ITU-T Recs. P.800 (voice) [26], P.913 (2D video) [27], and P.919 (360 degree video) [28]. In terms of research, although they began by evaluating video quality or other aspects related to Quality of Service (QoS) in isolation, several works demonstrated the influence on the results of other aspects that should be characterized, such as subjects or environments [29, 30]. To address these aspects, ITU-T Rec. P.10/G.100 [31] defined the term **Quality of Experience (QoE)**, the degree of delight or annoyance of the user with an application or service [32]. The aspects that influence QoE in the XR environment can be divided into three categories: human (vision and hearing, simulator sickness, immersion, and expectations and expertise), system (content, network conditions, and encoding parameters), and context (physical, temporal, social, and task context) [33]. Despite the constant evolution of the recommendations, each individual one is tailored to specific test conditions, and generalization outside them is not simple. Even though ITU-T is now introducing these aspects in its schedule (e.g., in the ITU-T Rec. P.1320 “QoE assessment of eXtended Reality (XR) meetings” [34]), this work is far from being complete.

QoE is the second fundamental pillar of this work. This context and the evolution of QoE assessment guide the research line of this work. In fact, some results from this work are contributions to ITU through Video Quality Experts Group (VQEG) [35], in charge of defining methods and tools that address new technologies and result in ITU recommendations. Experiment designs are based on the knowledge acquired with traditional contents [25, 27, 36] and recently adapted to omnidirectional video [37]. Therefore, the inheritance of the ITU origin remains latent when evaluating the perceived effect of the impairments caused along the signal processing chain (acquisition, compression, transmission, display, etc.) as the main target. For example, to evaluate video quality, short-duration videos (10 s-20 s) encoded with different qualities are displayed in random order to be evaluated (on a five-level scale) by participants [38]. However, short duration contents are not suitable for assessments that consider narrative or context. Therefore, these methodologies are not applicable to environments in which communication exists and requires a narrative, losing information on the effects that impairments can have on the experience. To go a step beyond technology when evaluating XR communications, **socioemotional aspects** are defined as that features related to human and context categories in QoE, mainly determined by psychological effects induced in the users [39]. For example, presence, “the sense of being there” [40], empathy, “the ability to view the world from another person’s perspective combined with an emotional reaction to that

perspective, including feelings of concern for others” [41], or attitude, “an individual’s evaluative judgment of the target behavior on some dimension” [42].

A great variety of works can be found in the literature that demonstrate the effectiveness of XR technology for achieving a higher enjoyment and involvement of users than traditional technologies [43, 44] and even the benefits on memory and attention assessments [45, 46]. Hence, it is possible to find questionnaires, environments, and applications developed considering the needs of each use case and experiment, usually in controlled environments and without considering the influence of technical aspects.

There are here two main limitations: reproducibility and ecological validity, understood as how well the experiment replicates the reality of the research [47]. First, some works in the literature are focused on methodologies and experiments on specific use cases and with participants that fit those use cases (e.g., VR music therapy for the elderly). This fact makes the replication of the experiments difficult and, therefore, can limit the validation and generalization of the results. Second, the controlled environment limits the generalization of the findings in experiments to the real world outside the laboratory [48, 49]. This is aggravated in interactive experiments, since the behavior of the participants is not totally natural [16].

Following the Unified Theory of Acceptance and Usage of Technology (UTAUT) [42], the acceptance of a technology by population is influenced by the performance expectancy and the effort expectancy. Based on the gaps in the literature regarding the assessment of this technology, consistent methodologies, more realistic scenarios, and a greater diversity of samples are needed to ensure that this technology is accepted in society.

Based on this analysis, it is assumed that XR communications can be a solution for teleconferencing, remote collaboration, or teleoperation purposes, but the reliability of the assessments should be increased with:

- **methodologies that consider technical and socioemotional aspects** and the influences between them
- **methodologies that consider different conditions** (e.g., type of conversation, the possibility to see your hands while using the HMD)
- the **validation of these methodologies in real scenarios**

The main objective of this work is to understand immersive communications from the user-centered point of view. Given the diversity of options offered by XR technologies, we focus the research on

a real-time 360-degree video communication, analyze it considering technical and socioemotional aspects, and validate the methodology and findings in a reference use case.

## 1.2 Research objectives and contributions

To evaluate and understand interactive immersive communications in real scenarios, we propose the design of suitable methods that scale for massive use. It means that the findings should be applicable to different communication environments and test conditions, considering technical and socioemotional aspects. As the final step of the research, the validation of the work in real communications, understood as those tests that allow participants to speak naturally, is proposed.

The use case of reference selected for this research is **tele-education**. One of the main motivations is that in education it is key to socialize with classmates since it is at an early age when we learn about relationships. Therefore, immersive communications can be an alternative for those students who cannot attend lessons in person, alleviating the isolation. Also, tele-education is a known environment for us and we have access to carry out experiments and create material (e.g., databases) at the university.

So, the main challenges tackled in this thesis are:

- How can we jointly evaluate technical and socioemotional aspects?
- What conditions influence on the socioemotional aspects in immersive communications?
- What tools can be added to improve the experience in immersive communications?
- How to design a methodology that can be applied in real-time 360-degree video communications? How could it be validated?
- How can the methodologies and findings be validated in real uses cases, such as tele-education?

To face these challenges, the system presented in Figure 1.2 is the reference configuration on which we work. However, in each chapter, different conditions are evaluated, giving rise to the phases of this research and the structure of this document. Furthermore, this work is an easily adaptable contribution to other immersive communication technologies.

**Chapter 2** focuses on the **objective video quality assessment** of 360-degree videos. At a time when Video Multimethod Assessment Fusion (VMAF) [50], a full reference objective metric developed by Netflix, is being widely used with 2D video, we propose its application on 360-degree content [51]. Additionally, we perform an exhaustive analysis of the traditional objective

metrics and the adaptations of them for omnidirectional contents. We conclude that VMAF works correctly with 360-degree videos homogeneously encoded, without specific adaptations of the VMAF implementation for this type of video. This result is an important contribution since providing an acceptable video quality is necessary to offer users a good QoE. This also means that video quality assessment procedures, based on traditional methodologies, are suitable for this type of content. Additionally, this work is a very useful tool for the research community focused on the development of techniques and encoding schemes to save bitrate by offering the best possible quality. This work has given rise to the contribution [52].

**Chapter 3** is a first approximation of the subjective assessment of technical and socioemotional aspects, focused on **subjective video quality and presence assessment**. The conclusions of this study are a relevant input for the decisions made in the next steps. Based on methodologies highly tested on 2D content, we carry out a subjective assessment where video quality (technical aspect) and presence (socioemotional aspect) are evaluated in two of the most popular HMDs at that moment. The visualization of repeated short-duration clips decreases the sense of presence motivated by what we called fatigue effect. This is a relevant motivation to explore in depth methodologies that cover the evaluation of video quality and socioemotional aspects and the interaction between them. Additionally, thanks to the informal inputs about the influence of the 360-degree video on the sense of presence, we decide to explore higher-level aspects in the following experiments. Also, we find that a handheld controller is a suitable interaction device for subjective evaluation tests asked in the virtual environment. As member of VQEG, the conclusions of this pilot study have been an instrumental contribution to create ITU-T Recommendation P.919 for the quality evaluation of VR, a new standard for subjective test methodologies for 360-degree video on HMDs [37]. This work has given rise to the contribution [53].

**Chapter 4** addresses the **subjective video quality and socioemotional aspects assessment** from low level aspects, such as acquisition and encoding of the videos or test conditions, to high level aspects, such as the context and the type of conversation. We propose a methodology to assess video quality, spatial and social presence, empathy, attitude, and attention in 360-degree videos for immersive communications. We validate it in an experiment where the immersive communication is simulated. Then, participants visualize three contents, designed and acquired for the experiment, considering conversations of different genre (everyday conversation, educational, and discussion) and actor and observer acquisition perspectives. Videos are encoded with quality fluctuations, simulating a streaming session with unstable network conditions. The conversations are placed in the context of international experiences to have a diverse sample of participants with experiences working or studying in a foreign country. We consider three experimental conditions: (A) visualizing and rating the perceptual quality of contents in an HMD, (B) visualizing the contents in an HMD,

and (C) visualizing the contents in an HMD where participants can see their hands and take notes. As a contribution, we highlight the proposal of a methodology to assess socioemotional aspects with the additional task of evaluating the video quality. Also, the fact that the evaluation of the video quality and the socioemotional aspects influence the results motivates the validation of the methodology in realistic scenarios. This work has given rise to the contribution [54].

**Chapter 5** presents an **interactive communication assessment** that considers in-person or hybrid meetings with one remote participant attending the meeting with the HMD and two participants sharing the same physical space. We propose a methodology to assess socioemotional and technical aspects. The test session is based on a decision-making technique in the context of education, which allows a natural flow of the conversation between the participants. We assess aggregate quality, spatial and social presence, and socioemotional aspects evaluated with questionnaires based on previous experiments. This design and the results can be a framework easily extended to other XR technologies used for hybrid meetings, allowing comparisons between them and generalization of the results. This work has given rise to the contribution [55].

**Chapter 6** seeks the validation of the technology and the proposed methodologies in a **tele-education use case**. We propose a system for tele-education streams in real time a class using a 360-degree camera, allowing remote students to explore the whole scene and improving the feeling of being in the classroom with their colleagues, presented in detail in [4]. It is powered by a set of Artificial Intelligence (AI) deep learning algorithms that work on the 360-degree frames to detect the events of interest. Then, these events are sent to the remote participant application and the application presents the notifications in the virtual scene. We create EVENT-CLASS, a dataset of 360-degree videos with annotated events in the classroom [56], to be used in quality assessments and in the implementation of machine learning techniques.

To validate the prototype, socioemotional aspects, such as presence, perceived quality, usability, and usefulness of the notifications, are evaluated using questionnaires. The obtained results show that using immersive tele-education systems can improve presence, as well as the benefits of the notifications on the experience of the remote students. Thanks to the experience acquired in the use case of tele-education, we organized the workshop “Emerging Telepresence Technologies in Hybrid Learning Environments” at ACM Conf. on Human Factors in Computing Systems, CHI 2022 [57]. This work has given rise to the contribution [58].

**Chapter 7** poses the **lessons learned for the QoE assessment of immersive communications** during this research about designing an experiment. Our aim is to share the relevant features that researchers, developers, and service providers from different research areas and degree of expertise should keep in mind. To do that, we present an analysis of the stages necessary to design

subjective assessment tests. It is carried out from a common framework to merge the two visions from telecommunication and HCI research for XR assessments. Additionally, the highlights of each stage are summarized in guidelines. This work has given rise to the contribution [59].

**Chapter 8** provides a summary of the challenges tackled and highlights the main **conclusions** of this research. Additionally, **future research** directions and potential areas for further investigation are proposed. Overall, this chapter aims to provide a comprehensive overview of this research efforts and to emphasize the significance of our findings within the broader context of the field.

**Appendix A** summarizes the **scientific contributions** that have result from this research. It is divided into four categories: journals, conferences, contributions to standards, and public available data.

**Appendix B** presents the presents the **outreach activities** of the research that have been carried out throughout the development of the thesis. It is divided into four categories: press articles, audiovisual multimedia, and talks.



## Chapter 2

# Objective Video Quality Assessment

### 2.1 Introduction

Although different XR applications consider content locally hosted, real-time 360-degree video communications require content streamed to the client whenever required. The delivery of omnidirectional content with acceptable quality is still a challenge due to the amount of resources it demands. To relax these strict conditions, different approaches can be considered. First, the design of new quality ladders leading to different perceptible levels of quality in 360-degree contents. Second, efficient delivery schemes that take advantage of the specific characteristics of 360-degree videos visualization. In particular, existing schemes are typically based on the fact that only the viewport is viewed by users and it varies with the movements of their head with respect to the scene [60]. Therefore, only the area viewed by the user needs to be provided with high quality, which reduces the overall bitrate required. However, with this solution the delay will be a very influential factor in ensuring a good QoE [23]. Moreover, other approaches take into account the users' behavior assuming that users tend to look at certain orientations or elements in the scene with higher probability than others. In this case, the content is prepared considering saliency or attention maps, leading to a better use of the bitrate [61, 62]. Additionally, other proposals exploit the peculiarities of the type of projection (e.g., equirectangular), that each projection impacts in a different way the quality of the different areas of the omnidirectional image, to provide satisfactory quality to users and save bitrate simultaneously [63]. All these approaches require a video quality metric that offers reliable results in the sense that it should be able to capture the quality perceived by users [64].

Given the technical limitations and the relevance of video quality in our scenario, this chapter is focused on objective video quality assessment. Specifically, this work validates one of the most robust objective metrics developed for traditional content, Video Multimethod Assessment Fusion (VMAF) metric [50], on omnidirectional videos.

Here, the structure of the chapter is presented. Section 2.2 presents the main works found in the literature related to objective video quality assesment. Section 2.3 describes VMAF and its application on 360-degree videos. Section 2.4 presents the validation of VMAF for 360-degree videos

through a subjective quality assessment. Section 2.5 includes the performance of VMAF and the comparison with other objective metrics in 360-degree videos. Finally, Section 2.6 presents general conclusions of this chapter.

## 2.2 Related work

Azevedo et al. [65] summarize the main alternatives proposed in the literature to apply objective quality metrics on 360-degree content: a) the application of objective metrics developed for 2D content on the 360-degree video in equirectangular projection, b) the application of objective metrics developed for 2D content on the viewport [66, 67], or c) the application of objective metrics developed or adapted considering the nature of the 360-degree videos. The implementation of option b) applied to immersive communications requires a viewport prediction model that ensures the adaptation of the viewport with low latency. In this way, the update of the quality of the viewport could make an acceptable QoE delivered to users. On the other hand, options a) and c) are the most direct application options. This, added to the appearance of the most robust objective metric for 2D content, makes this work focus on options a) and c).

A significant effort has been made to adapt some of the most popular and useful quality metrics of the traditional 2D world to 360-degree scenarios. There is literature referring to modifications of the Peak Signal-to-Noise Ratio (PSNR) metric to fit the specific features of 360VR content [68]. Specifically, Lakshman et al. [69] proposed the Sphere based PSNR computation (S-PSNR), where the distorted frame is projected onto a sphere before computing its distortion. So, for each projected point on the sphere, the associated pixels in the plane domain are calculated to compute the PSNR. Based on the S-PSNR, other methods have targeted the approximation of the average quality of all possible user points of view related to different viewports, weighting them taking into account the attention maps experience. For instance, Sun et al. [70] proposed the use of the Weighted to Spherically PSNR (WS-PSNR) metric, where the weights assigned to an area decreases as this area gets away from the equator. Similarly, Zakharchenko et al. [71] proposed the Craster Parabolic Projection PSNR (CPP-PSNR) metric, where the weights are assigned to different areas based on this projection. In contrast, Ghaznavi et al. [72] introduced the Uniformly Sampled Spherical PSNR (USS-PSNR) metric, which implements a uniform weight sampling of the decoded video on the sphere. Hence, the sample density changes based on latitude and longitude. Anyhow, these adaptations still have the same problem as the original PSNR, they do not take into account any Human Visual System (HVS) characteristics [73].

With the aim of including subjective aspects, the Multi-Scale Structural Similarity (MS-SSIM) index method was proposed by Wang et al. [74]. It extends SSIM by incorporating information

regarding image details at different resolutions and viewing conditions that subsequent works have adjusted for 360-degree content. Although the approximation of the perceived quality carried out by MS-SSIM outperforms the results of PSNR-based methods, the complexity involved complicates its use [68]. None of these Video Quality Metrics (VQMs) adaptations offer a solution for a 360-degree video scenario in terms of reliability and resource consumption. For this reason, we have focused our work on the extension to omnidirectional video of one of the most influential metrics used for 2D contents: Video Multimethod Assessment Fusion (VMAF) metric [50].

### 2.3 Video Multimethod Assessment Fusion (VMAF) on 360-degree videos

VMAF is a Full-Reference (FR) metric which means that one reference sequence is needed to calculate how corrupted the other sequence is. It is based on different elementary metrics combined by a machine-learning algorithm, offering a good prediction of the human quality perception [50]. Its original version was designed to operate with 2D content of up to Full HD (1080p) resolution under limited compression and viewing conditions. However, subsequent proposals have extended its capabilities and range of operating points to include more types of content and displays and extra compression, distribution, and viewing conditions [75]. Furthermore, recent studies have verified its accuracy on environments different from the one it was intended to without any specific training in this sense. In particular, Rassol et al. [76] carried out subjective quality tests to validate the application of VMAF to 4K 2D contents, a resolution the metric was not originally trained for, obtaining good results when trying to predict the VMAF score. Moreover, Barman et al. [77] validated VMAF's performance to assess gaming content quality and Lee et al. [78] proved that it correlates well with the user's perception in ABR environments. Besides, Bampis et al. [79] used the dataset created for VMAF to implement their quality predictor and compare the results obtained by VMAF with other typical metrics. Likewise, Bampis et al. [80] proposed the SpatioTemporal-VMAF (ST-VMAF), a VMAF extension consisting in expanding the analysis of temporal features in video sequences to enhance the metric results. The significantly good results provided by VMAF with several types of non-immersive contents and viewing conditions led to consider its application without making any specific adjustments to assess omnidirectional content. In this way, it is possible to avoid the generation of large and rich specific 360-degree video datasets, the conduction of numerous subjective quality assessments and the performance of corresponding training and testing stages. This new approach, not considered before, allows to save time and resources, and endorses the incorporation of the VMAF metric in the form of embedded software or others in consumer electronic devices to assess the quality of the content provided in 360-degree video systems.

This work sought to apply the FR VMAF metric to 360-degree videos without any special training or modifications [52]. Additionally, we think of applying it across the entire frame without focusing only on the FOV because it would allow for a far more widespread and useful deployment of the technology without the need to monitor user motions and process the necessary data later. Therefore, regardless of the area of the scene that is being observed, we assume that the quality is comparable to the perceived quality if participants see the entire frame. This assumption is based on two factors. First, the spherical image's quality and content are typically similar, producing similar scores of VMAF. Second, there is a significant synchronicity in user behavior—the overlap of FOVs across time—because, as saliency maps demonstrate users frequently focus on the same areas of the scene [81].

Users frequently focus their attention on regions close to the equator [82], where the distortion caused by any projection is smaller, according to saliency or attention maps. Indeed, local image features in these areas are closer to those of 2D contents. So, it stands to reason that a robust metric created for 2D contents can offer significant outcomes for typical 360-degree content. Since the equirectangular projection is the one that is currently most frequently used, we have focused our study on it.

The VMAF metric is tailored to only cover compression and scaling artifacts, as content is assumed to be already edited and finished, and transmission impairments are solved in adaptive bitrate streaming scenarios [50]. With the aim of providing good QoE and thus guaranteeing an immersive and engaging experience, scaling is unusual in XR scenarios [83, 84]. Therefore, our analysis is focused on compression. During the encoding, video suffer a quantization process in which the signal amplitudes of the raw frame are mapped to a finite number of discrete values, compressing the source signal. In this process, the QP determines the size of the quantization step, so the higher the QP value, the less detail is retained in the compressed frame, implying lower bitrate usage but also lower video quality [85]. Then, QP is used to determine the quality of the compression outcome, allowing to establish levels of acceptable and unacceptable quality for most of the users [86]. Besides, QP analysis is useful for the implementation of efficient encoding schemes where multi-quality frames are created.

We hypothesize that there is a similarity relation between the application of VMAF on 2D contents and its proposed new application on 360VR contents. Thus, the VMAF-vs-QP curve for 360VR contents should be monotonically decreasing by the nature of the encoding and, therefore, the validation can be carried out on a reduced set of adequately selected values, without replicating the whole VMAF design on 360VR contents. So, instead of conducting a sweep over the whole range of QP values to search for Just-Noticeable Differences (JND) [87], we consider only a subset of them which correspond to anchor VMAF scores in the curve.

So, after selecting the set of SouRCe sequences (SRCs), the process is performed in two steps. First, to obtain target qualities, we encode each SRC with constant QP covering the whole range of possible QP values. Later on, we apply the original VMAF metric to these Processed Video Sequences (PVSs) to obtain the variation of the score with the encoding parameter. Secondly, we verify through subjective tests that the users' perception fits the VMAF scores obtained in the first step.

Here, we present the reasons why we use this process through which we obtain the reference VMAF-vs-QP curve for 360-degree contents. It is divided into two main parts: Subsection 2.3.1, where the created database and the main features of the SRCs are presented and Subsection 2.3.2, where the VMAF scores are analyzed.

### 2.3.1 Video source characterization

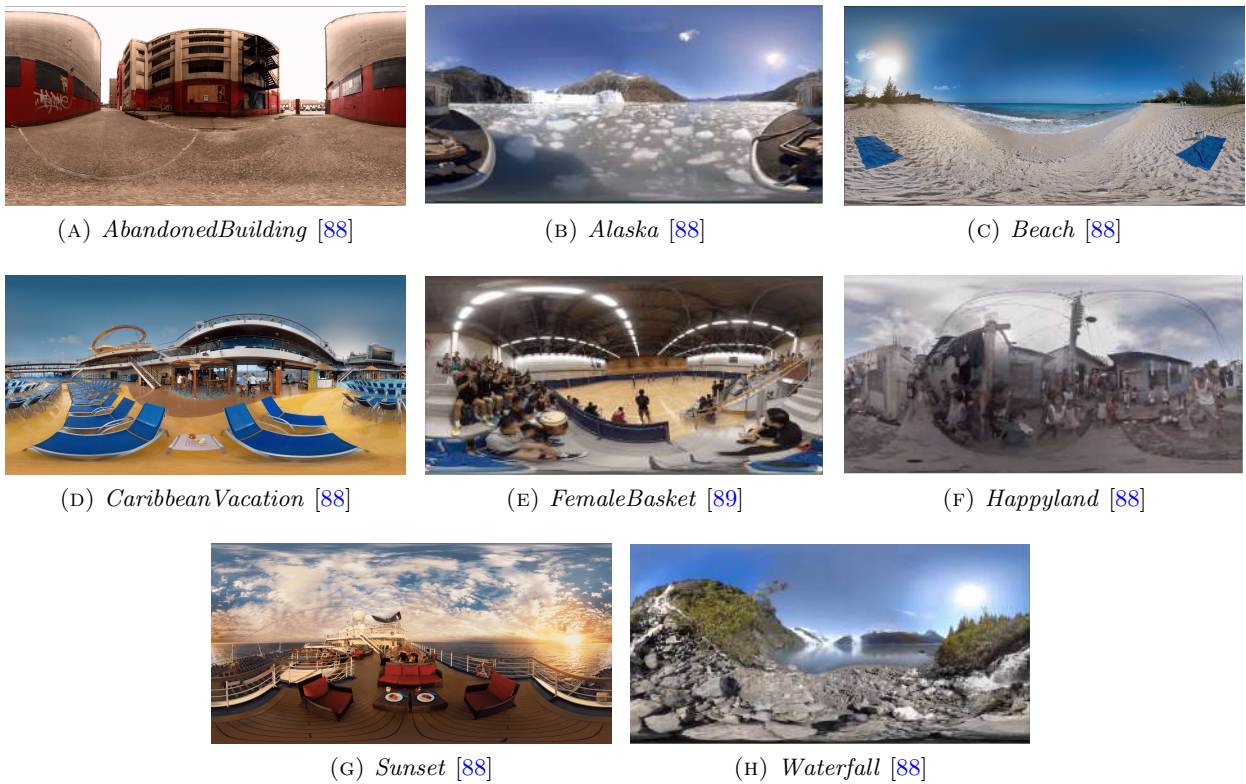


FIGURE 2.1: Video sources screenshots in equirectangular projection.

In compliance with ITU-R Rec. BT.500-14 [25], we have chosen nine clips from a total of 360-degree videos in equirectangular format as SRCs. These clips have different characteristics in terms of color, texture, camera motion, and type of content. The selected clips do not present any relevant changes between frames and guarantee a minimum level of visual comfort to avoid any disturbances that could affect the subjects' rates. Seven of them were obtained from a public database from the Virtual Human Interaction Lab from Stanford University [88] and one came from a dataset produced

by Wu et al. [89]. The last one came from a private source. Figure 2.1 presents screenshots of the first eight sequences.

All nine clips had a duration of 10 seconds, following several works in the literature on subjective quality assessments in 360-degree scenarios [90–92]. Concretely, Singla et al. [90] found that 10 seconds are enough to properly assess homogeneously encoded contents, since this duration allows the participants to find a certain area of the scene and compare it between different qualities. The original resolution of all the sequences was 4K (3840x1920 pixels), maintained throughout the experiment. All clips were set to 25 fps to build a homogeneous dataset. Despite not being a particularly high framerate, we selected it intending to use representative, varied, and habitual sequences of very different nature and complexity. The reason is that most available 360-degree videos (including the ones included in most public databases) are of the same or a very similar framerate, as this is the frequency at which most commercially available cameras capture content [93, 94]. The semantic characterization of the selected contents is presented in Table 2.1.

TABLE 2.1: Semantic characterization of the 360-degree videos considered in the test.

| Name              | Genre       | Acquisition perspective | Description  |
|-------------------|-------------|-------------------------|--|
| AbandonedBuilding | -           | Observer                | Daytime shot of an alley in between two abandoned buildingsstatic content with notable texture.          |
| Alaska            | Nature      | Observer. Motion        | The effects of climate change on Alaska’s glaciers from a sailing boat (camera motion).                  |
| Beach             | Nature      | Observer                | A beach landscape with superimposed titles.  |
| CaribbeanVacation | Documentary | Observer                | People on a cruise deck. An additional video is played back on a screen of the cruise in the background. |
| FemaleBasket      | Sports      | Observer                | A basketball game.   |
| Happyland         | Documentary | Actor                   | Short documentary on a Manila dumpsite where 40,000 people call home.                                    |
| Sunset            | Documentary | Observer                | Camera on a sailing cruise.  |
| Waterfall         | Nature      | Actor. Motion           | A landscape with a large waterfall that is rather close to the camera.                                   |
| Lions             | Animals     | Actor                   | Viewer gets an up close experience with a tiger on a savanna.  |

Additionally, all SRCs were objectively characterized in terms of their spatial and temporal complexity, computing Spatial Information (SI) and Temporal Information (TI) indicators, respectively, as expressed in ITU-T Rec. P.910 and P.913 [27, 36]. Figure 2.2 shows the distribution of the SI and TI values.

To obtain the full range of scores, all SRCs were encoded with ITU-T Rec. H.265/High Efficiency Video Coding (HEVC) using fixed QPs ranging from 1 to 51 [21]. The condition applied on the

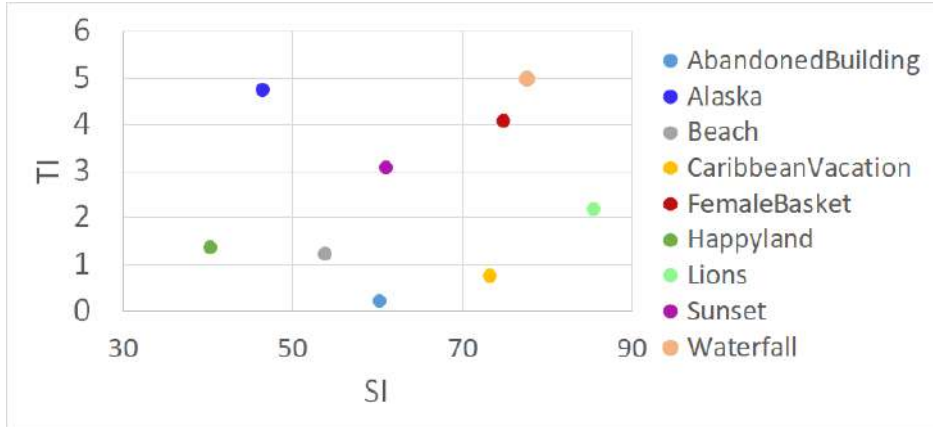


FIGURE 2.2: Spatial (x-axis) and Temporal Information (y-axis) indicators for the 360-degree videos considered in the test.

SRCs to create PVSs in video quality area is called the Hypothetical Reference Circuits (HRCs). As a result, we obtained 51 PVSs per SRC with bitrates ranging from 310 Mbps to 370 kbps. A summary of the created dataset is presented in Table 2.2. This set of 459 (51 times 9) sequences were the inputs to the VMAF computing algorithm.

TABLE 2.2: Characteristics of the dataset of the 360-degree videos selected for the test.

|  |                       |
|--|-----------------------|
| Number of SRC videos                   | 9                     |
| Duration                               | 10 seconds            |
| Encoding                               | H.265/HEVC            |
| Resolution                             | 4K (3840x1920 pixels) |
| Hypothetical Reference Circuits (HRCs) | QP range (1-51)       |
| Framerate                              | 25 fps                |
| <b>Total number of videos: 459</b>     |                       |

### 2.3.2 Experimental results

Here, the results of computing the VMAF metric over the whole set of PVSs are presented. To that end, we used the VMAF Development Kit (VDK) that can be found available in a public repository [95]. Specifically, we employed VDK version 1.3.3 and VMAF version 0.6.1 with the default configuration parameters. This model was selected as it was more stable and better suited to our scenario, as the resolution perceived by users through the HMD is quite lower than 4K. Due to the absence of scene changes in the SRCs, the arithmetic mean was used as a temporal pooling mechanism, since it is a representative value for those sequences.

Figure 2.3 shows the relation of VMAF final scores for all the contents with QP, which is monotonously decreasing. Furthermore, the curve decreases slightly for the highest qualities (low QP values), more sharply for medium qualities (medium QP values), and dramatically for low

qualities (high QP values). As already mentioned, the effect of changing the QP varies with the characteristics of the content, resulting in a different VMAF curve for each of the SRCs.

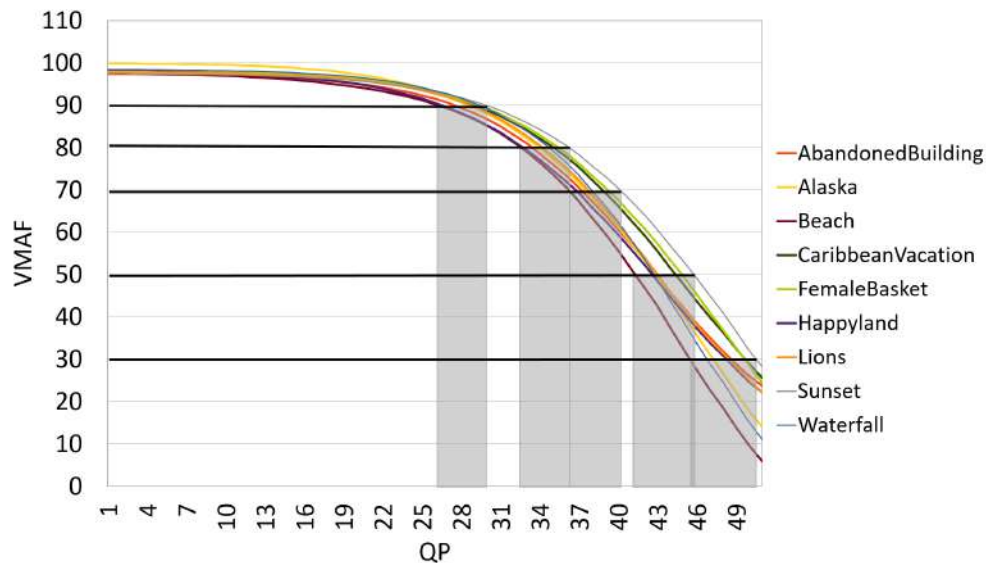


FIGURE 2.3: VMAF (y-axis) vs QP (x-axis) curve for all SRCs. The VMAF anchor values used for the validation are represented with solid black lines (y-axis) which correspond to a range of QP values (x-axis).

## 2.4 Validation of VMAF for 360-degree videos through a subjective quality assessment

In this section, the subjective quality test conducted to validate the results obtained with VMAF is presented. As mentioned above, VMAF is a metric prepared to operate with 2D contents. Here, it is evaluated to what extent VMAF can be used with omnidirectional contents. To that end, we designed an experiment consisting in presenting a subset of the PVSs used in the previous step that are located closest to several strategic VMAF scores to some subjects. For each version, subjects were asked to evaluate the perceived quality. In this way, we obtained subjective quality rates for those strategic points within the QP range. These evaluations were used to check how close the given rates were from the computed VMAF scores for 360-degree videos.

The subjective assessment carried out in this work was based on the information obtained from recommendations related to 2D contents which have been highly tested: ITU-R Rec. BT.500-14 [25], ITU-T Rec. P.910 [36], and P.913 [27]. The reason is that at that time there were no official recommendations for subjective metrics to measure quality in XR scenarios.

### 2.4.1 Stimuli

As stimuli, a subset of the PVSs of the previous step corresponding to six quality levels (five distorted and one reference sequences) was used. So, a total of 54 (six qualities, nine SRCs) were presented to each observer. Concretely, considering the VMAF curve in Figure 2.3, the distorted PVSs selected in the validation step were those closest to the following key VMAF scores:

- VMAF equal to 90. This value is located where the curve begins to decrease slightly.
- VMAF equal to 80 and 70. These values are located where the curve decreases more sharply.
- VMAF equal to 50 and 30. These values are located where the curve decreases more dramatically.

Additionally, concerning the reference sequences, on the one hand, we have no access to the original raw videos. On the other hand, references must comply with the same restrictions as the rest of the sequences in the experiment, namely, that are encoded using a fixed uniform QP value. Therefore, we cannot directly use the available SRCs. So, for each content, we selected a reference that scores higher than 90 in the VMAF scale, since the reference clip needs to offer the best quality presented to the user during the test. In this way, a QP value of 0 is desirable during the encoding of the reference sequences but the high bitrates achieved are not suitable for a real-time 360-degree video communication. Therefore, reference sequences were encoded with a QP value that, when possible, led to a similar bitrate to that of the original video and, as a mandatory restriction, all references provided VMAF scores in the range between 92 and 95. The six qualities are denoted from A to F, where A is the reference (best quality version), and B to F are the five distorted versions associated with the VMAF scores 90, 80, 70, 50, and 30, respectively.

### 2.4.2 Methodology

**Personal information:** For each participant, we collected age, gender, and vision (corrected or normal). This was used to characterize our observers and guarantee diversity.

**Quality.** A Single-Stimulus (SS) method was applied in this experiment, specifically the Absolute Category Rating with Hidden Reference (ACR-HR) [36], where a reference version of each content is randomly presented to subjects, who rate it like any other [27]. This method uses a five-level rating scale: “Bad”, “Poor”, “Fair”, “Good”, and “Excellent” [36].

**Method of collecting ratings.** They used a developed application that allowed for visualizing contents and rating them subsequently without having to remove the HMD or interact outside the

XR environment [96]. This app then enabled a more immersive and engaging experience for the subjects.

### 2.4.3 Equipment and environment

Tests were carried out using a smartphone, Samsung S7, and a mobile VR headset, Samsung GearVR. In any case, the conclusions drawn from the experiments conducted with this device can be extended to the most used HMDs that do not offer significantly better display resolutions [93].

Test area was set according to ITU-R Rec. BT.500-14 [25]. The HMD used by participants tracked the rotational movements of their heads (3 DoF) [84]. Moreover, participants performed the tests seated in a swivel chair in the middle of a room to allow them to spin around freely, facilitating the exploration of 360-degree videos.

### 2.4.4 Test session

Figure 2.4 presents the structure of the test session. First, a quick overview of the experiment was introduced to the participants of the experiment. Also, participants were explained about the HMD they were going to use and the method of collecting ratings. They filled in the consent form, necessary to attend the test and process their data in accordance with the General Data Protection Regulation (GDPR) of the European Union [97]. In this part of the session personal information related to gender, age, and vision was collected.

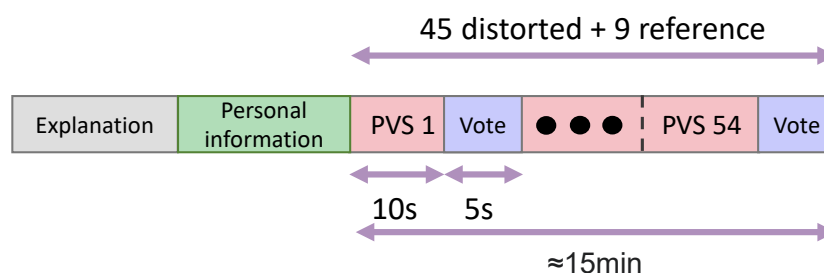


FIGURE 2.4: Test session structure.

Despite designing the experiment following the recommendations presented above, there was no training session in terms of showing the expected maximum and minimum qualities to the subjects. The reason was because we wanted to observe the real absolute quality that users perceive, with no bias. Furthermore, the nature of our experiments would not allow to perform such a session properly, since we could not guarantee a priori that different sources encoded with the same or a very close VMAF score provide an equivalent quality, since validating this link is the objective of

the experiments. Hence, it was not possible to present additional sequences not included in the test session showing representative global maximum and minimum qualities.

Each test session was composed of 54 clips with a duration of 10s (45 distorted and 9 reference videos). All videos were viewed by every subject. The duration of the whole test was around 15 minutes, assuming a period of 5 seconds to vote each PVS. The voting period length was user-driven and so was not limited beforehand. Different randomization of the PVSs was used for each session to reduce contextual effects, following Rec. ITU-R BT.500-14 [25]. Although the same quality could be presented on two consecutive videos, subjects could not watch the same clip with different qualities consecutively.

### 2.4.5 Participants

A total of 24 observers (8 females, 16 males) participated in this experiment, with age ranging from 21 to 36 (average of 26) and normal or corrected vision. An a-posteriori screening was conducted computing the Linear Pearson Correlation Coefficient (LPCC) between the scores of every subject and the average ones of the whole set of observers. Following the guidelines of ITU-T Rec. P.913 [27] for outlier removal, we set a threshold of 0.75, which led to eliminating one subject.

### 2.4.6 Experimental results

Following the ITU-R Rec. BT.500-14 [25], for each video artifact (QP), Mean Opinion Score (MOS) and the Differential Quality Score (DMOS) of the observers with the associated 95% Confidence Intervals (CI) were computed and presented in Figure 2.5. The objective was to analyze the distribution of the means and their cumulative frequency of appearance. The main differences between the MOS and DMOS results are related to high qualities. In the MOS graph, we observe that there is no statistical differences between Quality A, the reference PVS, and Quality B for all the contents. However, for most of the contents, it is possible to observe differences between Quality A and Quality C. This information is lost in the DMOS graph. Regarding medium qualities, the differences between them are more noticeable in general in both cases. Finally, the results show that working with such low qualities as E and F lead to the appearance of artifacts that are virtually equally annoying to the users.

In Figure 2.6, the VMAF and the normalized DMOS curves, with the associated CIs, are presented together for each content to facilitate comparison. The mapping was done considering that VMAF is basically a rescaled DMOS, as acknowledged by the authors [50, 98]. Therefore, for a given SRC, we only needed to connect the reference PVS with the VMAF score associated with the

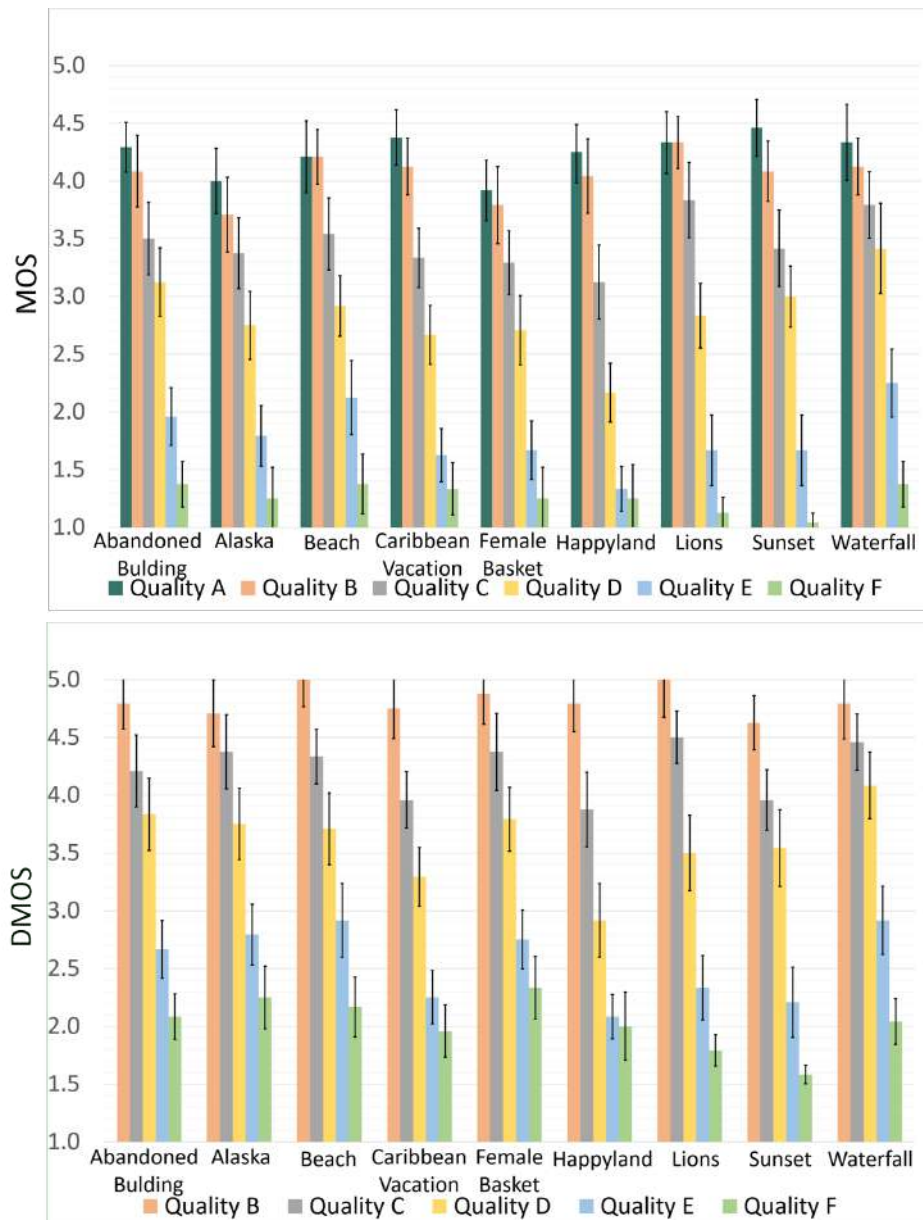


FIGURE 2.5: MOS and DMOS (y-axis) on a five-level scale obtained from 23 participants, including CIs. Participants evaluated clips of short duration (10s) encoded with fixed quantization parameters which determined Qualities A, B, B, D, E, and F (x-axis).

corresponding QP. The rest of the values were obtained preserving the relative differences with respect to the score of the reference video. It is worth mentioning that the absence of raw video sources in our test material influences our analysis in terms of the choice of the references for the subjective assessment and, consequently, the DMOS normalization. However, the alternative of acquiring a new specific database of raw video sources, with its associated problematic acquisition and stitching processes, is beyond the scope of this work.

Through the comparison of the VMAF and DMOS curves for all contents, we can study the performance of the VMAF metric for 360-degree videos. We can see that the shape of the curves is very similar and the gap between both is quite small. Therefore, we can conclude that the subjective rates obtained in our experiment fit the VMAF scores to a great extent for almost the whole range

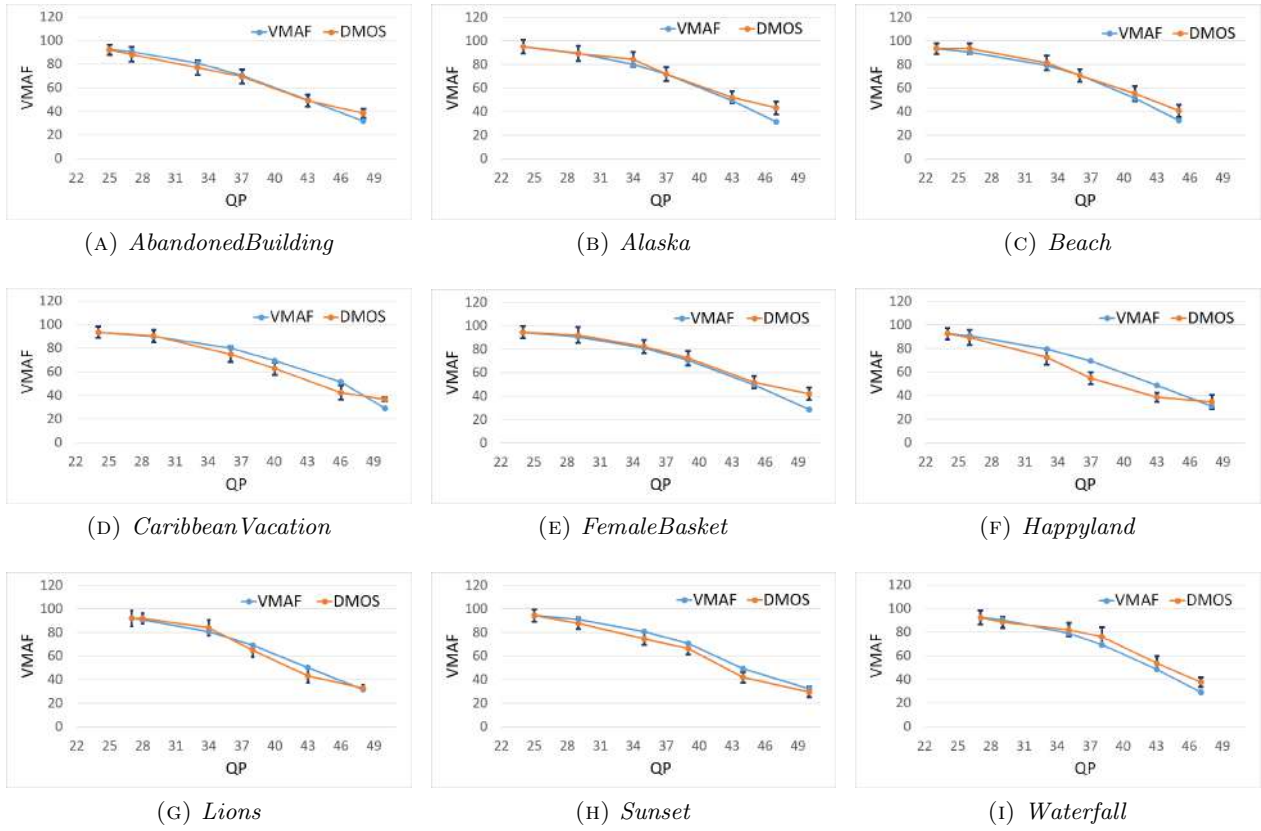


FIGURE 2.6: Evolution of the VMAF scores and the normalized DMOS on a 100-level scale with the associated CI (y-axis) with the QP value for each content (x-axis).

of qualities. Only for “Happyland” and, more moderately, “CaribbeanVacation”, we can really notice a greater gap between the VMAF and DMOS curves.

Nevertheless, we can see that there is a deviation of the DMOS curves with respect to the VMAF curves in the lowest range of qualities. The most plausible reason for that is that the perceived video quality goes into a saturation region, where users statistically barely perceive any differences. It is caused by artifacts that appear and are annoying to the user, making much more difficult for him/her to discern between such distorted contents. This saturation effect is further boosted by the characteristics of the HMD. In addition, this effect is also justified considering the computation of VMAF. The CIs associated with the VMAF score are notably higher for low qualities, decreasing the reliability of the results.

To validate these findings, we computed the Pearson’s Linear Correlation Coefficient (PLCC), the Root Mean Square Error (RMSE) and the Spearman’s Rank-Order Correlation Coefficient (SROCC) between the VMAF and DMOS values. These results are included in Table 2.3. Due to the deviation commented previously, the PLCC and the RMSE have been computed for qualities ranging from B to F and from B to E. It can be seen that the correlation between the VMAF scores and the DMOS is extremely high and is even higher for most of the sequences when the last QP is not considered. Although the overall PLCC is very similar in both cases, the RMSE clearly shows this

TABLE 2.3: Pearson correlation, RMSE and Spearman’s rank correlation between VMAF and DMOS for all contents.

| <b>Content</b>    | <b>PLCC</b>  | <b>PLCC</b><br><i>(without QF)</i> | <b>RMSE</b>  | <b>RMSE</b><br><i>(without QF)</i> | <b>SROCC</b> |
|-------------------|--------------|------------------------------------|--------------|------------------------------------|--------------|
| AbandonedBuilding | 0.995        | 0.997                              | 0.172        | 0.099                              | 1.000        |
| Alaska            | 0.989        | 0.992                              | 0.283        | 0.124                              | 1.000        |
| Beach             | 0.995        | 0.994                              | 0.211        | 0.124                              | 0.975        |
| CaribbeanVacation | 0.962        | 0.997                              | 0.349        | 0.339                              | 1.000        |
| FemaleBasket      | 0.990        | 1.000                              | 0.355        | 0.088                              | 1.000        |
| Happyland         | 0.955        | 0.981                              | 0.467        | 0.500                              | 1.000        |
| Lions             | 0.987        | 0.995                              | 0.201        | 0.222                              | 0.975        |
| Sunset            | 0.996        | 0.998                              | 0.251        | 0.275                              | 1.000        |
| Waterfall         | 0.995        | 0.986                              | 0.276        | 0.215                              | 1.000        |
| <b>Overall</b>    | <b>0.965</b> | <b>0.959</b>                       | <b>0.285</b> | <b>0.221</b>                       | <b>0.994</b> |

effect. Finally, the SROCC is either one or very close to one. Therefore, we can assure that VMAF works properly with 360VR content with homogeneous encoding, providing remarkably good results with no specific training focused on omnidirectional content.

## 2.5 Comparison of VMAF with other objective metrics

In this section, the comparison of the results obtained by VMAF with the following FR VQMs: PSNR, WS-PSNR, CPP-PSNR, SSIM, and MS-SSIM is presented. The VQMs were computed on the nine SRCs using public available software [95]. A regression analysis between the subjective DMOS and the outcome of each objective metric was conducted using a third degree polynomial without any fitting constraints and a sigmoid function. Table 2.4 shows the values of the PLCC, RMSE and the coefficient of determination  $R^2$  [99] between both fittings and the DMOS. We can see that the results for the sigmoid function are slightly worse than the ones obtained with the polynomial function.

Figure 2.7 shows the scatter plots and their corresponding curves for the polynomial fitting. The ones corresponding to the sigmoidal fitting are not included due to their great similarity to the polynomial ones. The top and second row three graphs present PSNR, WS-PSNR, and CPP-PSNR in linear and logarithmic scales, respectively. In the third row, SSIM, MS-SSIM, and VMAF are presented.

TABLE 2.4: Pearson correlation, RMSE, and coefficient of determination  $R^2$  of fitting curves and DMOS for all analysed metrics.

| Metric                     | Polinomyal fitting curve |       |       | Sigmoidal fitting curve |       |       |
|----------------------------|--------------------------|-------|-------|-------------------------|-------|-------|
|                            | PLCC                     | RMSE  | $R^2$ | PLCC                    | RMSE  | $R^2$ |
| PSNR ( <i>linear</i> )     | 0.851                    | 0.593 | 0.725 | 0.847                   | 0.708 | 0.698 |
| WS-PSNR ( <i>linear</i> )  | 0.860                    | 0.577 | 0.740 | 0.857                   | 0.684 | 0.716 |
| CPP-PSNR ( <i>linear</i> ) | 0.873                    | 0.551 | 0.763 | 0.869                   | 0.659 | 0.740 |
| PSNR ( <i>db</i> )         | 0.851                    | 0.593 | 0.725 | 0.846                   | 0.700 | 0.696 |
| WS-PSNR ( <i>db</i> )      | 0.861                    | 0.576 | 0.741 | 0.856                   | 0.683 | 0.715 |
| CPP-PSNR ( <i>db</i> )     | 0.874                    | 0.550 | 0.763 | 0.869                   | 0.657 | 0.742 |
| SSIM                       | 0.874                    | 0.550 | 0.763 | 0.866                   | 0.624 | 0.736 |
| MS-SSIM                    | 0.956                    | 0.333 | 0.914 | 0.951                   | 0.397 | 0.904 |
| VMAF                       | 0.980                    | 0.227 | 0.960 | 0.969                   | 0.304 | 0.935 |

It is clear that VMAF outperforms the rest of the metrics under evaluation, as already reported for conventional HD video [50] and still holds in 360-degree video. Besides, VMAF is the only one whose relation with DMOS is almost linear, only modified by the user perception of the lower qualities. The PLCC and the RMSE values before and after the polynomial fitting are very similar, which shows that VMAF can be used for 360-degree content without particular adaptation.

## 2.6 Conclusions

We have presented an exhaustive study on the feasibility of directly applying the original VMAF metric to assess the quality of omnidirectional contents visualized by users using an HMD. Based on the assumption that VMAF scores decrease monotonically with the QP, due to the effect of this encoding parameter in the resulting sequence, we carried out an experiment consisting of two main steps. First, we used the original implementation to obtain the VMAF score of several 360-degree sequences encoded with constant QP in the whole range of possible values to capture how it varies with the encoding parameter. Secondly, we validated the obtained VMAF scores through a subjective assessment. We done so by creating a second curve per content from a finite number of scores corresponding to several operating points, which been selected sufficiently spaced. These values were then normalized DMOS obtained in the subjective tests for the subset of input sequences encoded for the specific QP anchor points. The minimum divergence of the two curves in most cases allows us to conclude that VMAF works sufficiently correctly with this homogeneous 360-degree content, without performing any particular adjustments to prepare the metric accordingly.

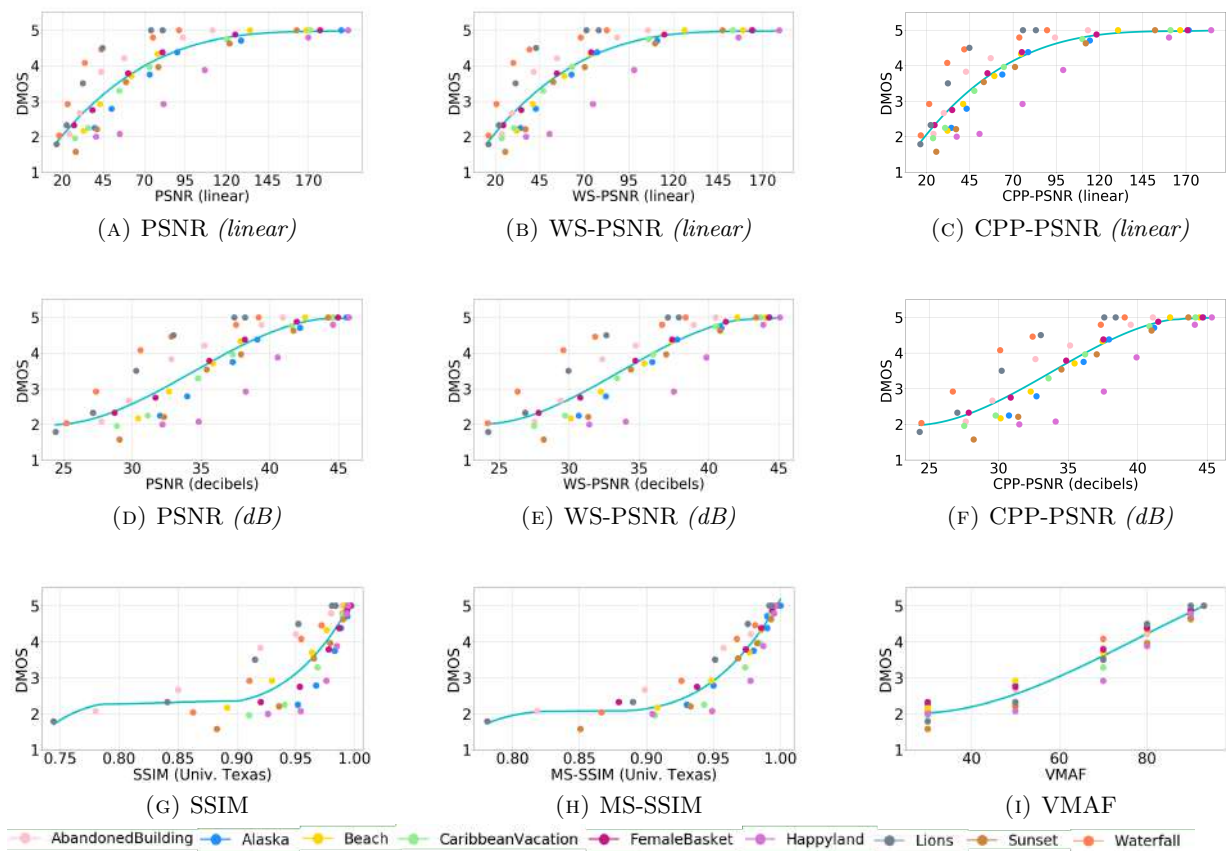


FIGURE 2.7: Mapping of DMOS ratings (y-axis) to objective scores (x-axis). Solid line represents the best fitting by a third degree polynomial curve.

In this way, in a 360-degree content visualization with an HMD scenario, it is possible to avoid the creation of a specific dataset with rich 360-degree content of acceptable quality and retraining the machine learning algorithm to obtain an omnidirectional-content-aware VMAF metric, saving computing and time resources. The suitability of this option and its associated gains in terms of resources make it an appropriate and robust choice to be incorporated throughout the encoding and transmission chain to properly and easily monitor the quality of 360-degree videos delivered to end-users.

## Chapter 3

# Subjective Video Quality and Presence Assessment

### 3.1 Introduction

Visual information, as said before, is one of the key aspects that is taken into account in the development of this thesis due to its influence on QoE. However, we are interested in evaluating video quality and other influential factors that affect the QoE. This chapter is a transition between a work focused on objective video quality metrics, Chapter 2, and subjective assessments of technical and socioemotional aspects, Chapter 4. One of the motivations for carrying out the evaluation of technical and socioemotional aspects in the same test is the influence of all of them on QoE, but also the interaction that some technical aspects could have on socioemotional ones, and vice versa. Therefore, video quality is selected as a technical parameter and presence, which relates with the sense of being in a place [100], as a socioemotional aspect.

The study is based on well-proven methodologies in the evaluation of video quality in 2D content, standardized in ITU Recommendations, and in the evaluation of presence in immersive environments.

Here, the structure of the chapter is presented. Section 3.2 presents relevant works in which the design choices presented in this chapter are based. Section 3.3 describes the main parts of an experiment design: stimuli, methodology, equipment and environment, test session, participants, and hypotheses. Section 3.4 presents the main results and finally, Section 3.5 provides general conclusions that are relevant inputs for the decisions in the following steps of the research.

### 3.2 Related work

One of the main features to take into account during the design of subjective experiments is the source content selected for the test. Despite the increase in consumption and therefore the creation of 360-degree video, high technical requirements are necessary for this kind of experiments. For

example, problems caused during video acquisition or post-processing (e.g., stitching errors) or audio artifacts can influence the quality evaluations and affect the understanding of the content narrative.

The use of short-duration videos is a common approach on audiovisual quality evaluation, which is supported by several recommendations related to subjective quality assessment, such as ITU-T Recs. P.910 and P.913 [27, 36]. As these methodologies have been highly tested in the literature with 2D video, the distribution obtained with the representation of the MOS with the associated CIs of video quality evaluations helps researchers to validate their experiments. Generally, it is more difficult for observers to appreciate the differences between very high quality content. However, in the videos encoded with intermediate qualities, the observers are able to find differences, but when the video quality is very low and annoying artifacts appear, the ratings saturate.

Most of the studies focused on video quality do not consider detailed socioemotional questionnaires [101]. Likewise, we found that there are several articles that make a great effort to analyze socioemotional features avoiding the effects of technical conditions [102, 103]. Despite the contribution of knowledge that other experiments which test together technical and socioemotional aspects, they are still missing the evaluation of high-level conditions that affect the QoE in a 360-degree environment and the differences between the selected questionnaires and scales. Then, this work is a starting point to jointly assess technical and socioemotional aspects. We mainly focused on video quality and presence. Presence has been highly in the literature with several questionnaires [104, 105], typically through questionnaires at the end of the test session using web-based platforms [105–107]. Others, evaluate presence through questionnaires in the same virtual environment [108]. This opens a new question for this research in relation to the evaluation mechanism within the virtual environment with which the participants feel more comfortable.

Based on the literature review, we proposed the evaluation of the quality of the video in the virtual environment and compared the method of collecting ratings, with handheld controller or touchpad. We also compared two of the most common consumer HMDs that offer similar technical features. Regarding presence, we used two of the most widely tested questionnaires in the literature and compared the collected results.

### 3.3 Experiment design

#### 3.3.1 Stimuli

In compliance with ITU-R Rec. BT.500-14 [25], we have chosen six clips from a total of 360-degree videos in equirectangular format as SRCs. The selected clips do not present any relevant changes

between frames nor stitching problems, and guarantee a minimum level of visual comfort to avoid any disturbances that could affect the subjects' rates. These clips have different characteristics in terms of color, texture, camera motion, and genre.

Figure 3.1 presents screenshots of the six sequences. One came from Radiotelevisión Española [109], four of them were obtained from Youtube, and other one came from a dataset produced by Nokia. For the semantic characterization, Table 3.1 describes the selected contents.

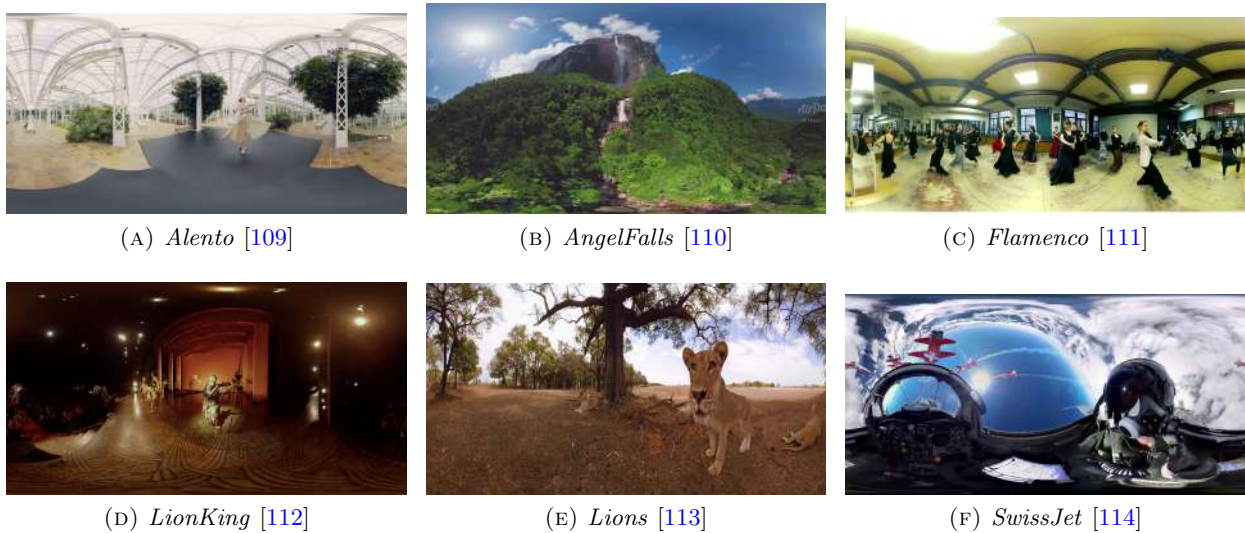


FIGURE 3.1: Video sources screenshots in equirectangular projection.

TABLE 3.1: Semantic characterization of the 360-degree videos considered in the experiment.

| Name       | Genre  | Description   |
|------------|--------|---|
| Alento     | Music  | It is characterized by the movement of a couple dance near the camera.  |
| Angelfalls | Nature | The main feature of this content relies on the motion of the camera, since it is on a drone flying over a landscape. The landscape is a jungle with a waterfall including two great challenges for the encoding process, vegetation and water movement. |
| Flamenco   | Dance  | It shows a lesson of Flamenco dance, where women are dancing around the camera.   |
| LionKing   | Music  | It presents the Lion King musical. The main challenges of this content are the illumination and the movement.   |
| Lions      | Nature | It shows a lion moving very close around the camera.  |
| SwissJet   | Sports | The camera is inside a jet so the video shows a pilot inside the cockpit. Also, other jets are flying around doing acrobatics.  |

All SRCs were objectively characterized in terms of their spatial and temporal complexity, computing SI and TI indicators, respectively, as expressed in ITU-T Rec. P.910 and P.913 [27, 36]. Figure 3.2 shows the distribution of the SI and TI values. Table 3.2 presents the original resolution and framerates of the SRCs that were maintained throughout the experiment. All six clips had a duration of 30 seconds and the audio was used to facilitate the user immersion [103].

To obtain the PVSs, SRCs were encoded with ITU-T H.265/High Efficiency Video Coding (HEVC) using fixed QPs. Specifically, 22, 27, 32, 37 and 42 QP values were applied as HRC [115]. As a

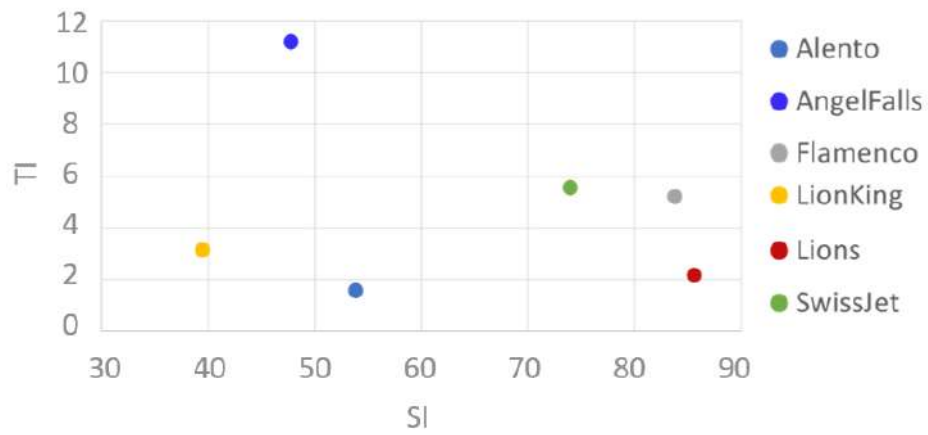


FIGURE 3.2: Spatial (x-axis) and Temporal Information (y-axis) indicators for all contents.

TABLE 3.2: Characteristics of the 360-degree videos considered in the test.

| Name       | Resolution (pixels) | Framerate (fps) |
|------------|---------------------|-----------------|
| Alento     | 3840x1920           | 25              |
| AngelFalls | 3840x2160           | 30              |
| Flamenco   | 3840x2160           | 30              |
| LionKing   | 3840x2048           | 30              |
| Lions      | 3840x1920           | 30              |
| SwissJet   | 3840x1920           | 50              |

result, we obtained five PVSs per SRC. Table 3.3 presents the results of computing several objective metrics over the whole set of PVSs.

TABLE 3.3: PSNR, WS-PSNR, CPP-PSNR, VMAF, SSIM, and MS-SSIM results of PVSs used in the work.

| Content           | Original resolution |         |          |       |       |         |      |
|-------------------|---------------------|---------|----------|-------|-------|---------|------|
|                   | PSNR                | WS-PSNR | CPP-PSNR | VMAF  | SSIM  | MS-SSIM |      |
| <b>Alento</b>     | 22                  | 45.54   | 44.53    | 44.96 | 93.79 | 1.00    | 1.00 |
|                   | 27                  | 42.98   | 41.89    | 42.17 | 90.12 | 0.99    | 0.99 |
|                   | 32                  | 40.14   | 39.00    | 39.15 | 83.39 | 0.99    | 0.99 |
|                   | 37                  | 37.14   | 35.97    | 36.04 | 72.26 | 0.98    | 0.98 |
|                   | 42                  | 34.06   | 32.88    | 32.92 | 55.32 | 0.96    | 0.96 |
| <b>Angelfalls</b> | 22                  | 43.80   | 43.18    | 43.62 | 97.59 | 1.00    | 1.00 |
|                   | 27                  | 40.32   | 39.65    | 39.94 | 92.40 | 0.99    | 1.00 |
|                   | 32                  | 36.86   | 36.18    | 36.35 | 82.61 | 0.98    | 0.99 |
|                   | 37                  | 33.65   | 33.02    | 33.11 | 66.19 | 0.97    | 0.98 |
|                   | 42                  | 30.76   | 30.26    | 30.31 | 42.95 | 0.94    | 0.95 |
| <b>Flamenco</b>   | 22                  | 45.80   | 44.76    | 45.19 | 95.90 | 1.00    | 1.00 |
|                   | 27                  | 43.11   | 42.02    | 42.32 | 92.93 | 1.00    | 1.00 |
|                   | 32                  | 40.30   | 39.19    | 39.38 | 87.32 | 1.00    | 1.00 |
|                   | 37                  | 37.37   | 36.26    | 36.38 | 77.75 | 0.99    | 1.00 |
|                   | 42                  | 34.31   | 33.22    | 33.28 | 63.35 | 0.99    | 0.99 |
| <b>LionKing</b>   | 22                  | 47.92   | 47.39    | 47.70 | 94.50 | 1.00    | 1.00 |
|                   | 27                  | 45.18   | 44.58    | 44.75 | 90.88 | 1.00    | 1.00 |
|                   | 32                  | 42.25   | 41.58    | 41.67 | 83.94 | 1.00    | 1.00 |
|                   | 37                  | 39.27   | 38.57    | 38.62 | 72.36 | 0.99    | 0.99 |
|                   | 42                  | 36.34   | 35.60    | 35.63 | 55.87 | 0.98    | 0.99 |
| <b>Lions</b>      | 22                  | 48.04   | 47.54    | 48.12 | 95.53 | 1.00    | 1.00 |
|                   | 27                  | 44.56   | 44.03    | 44.52 | 92.53 | 0.99    | 1.00 |
|                   | 32                  | 41.21   | 40.68    | 41.06 | 86.73 | 0.99    | 0.99 |
|                   | 37                  | 37.93   | 37.43    | 37.70 | 76.64 | 0.97    | 0.98 |
|                   | 42                  | 32.12   | 31.79    | 31.94 | 61.87 | 0.95    | 0.96 |
| <b>SwissJet</b>   | 22                  | 46.69   | 46.26    | 46.59 | 96.22 | 1.00    | 1.00 |
|                   | 27                  | 43.68   | 43.23    | 43.44 | 93.15 | 1.00    | 1.00 |
|                   | 32                  | 40.45   | 40.00    | 40.12 | 87.06 | 0.99    | 0.99 |
|                   | 37                  | 37.17   | 36.71    | 36.78 | 76.43 | 0.99    | 0.98 |
|                   | 42                  | 33.87   | 33.43    | 33.43 | 60.76 | 0.97    | 0.96 |

### 3.3.2 Methodology

Here, we explain in detail the methodology considered in the experiment.

**Personal information:** For each participant, we collected age, gender, and vision (corrected or normal). This was used to characterize our observers and guarantee diversity.

**Quality.** The methodology applied in this experiment was the Absolute Category Rating with Hidden Reference (ACR-HR) with a five-level scale, where the categories: “Bad”, “Poor”, “Fair”, “Good”, and “Excellent” were displayed on the HMD, as recommended in ITU-T P.910 [36]. In this sense, PVSs were randomly presented and evaluated. Different randomization of the PVSs was used for each session to reduce contextual effects.

**Presence.** Spatial presence was evaluated with the spatial presence scale of the Temple Presence Inventory (TPI) [104] and the subsampling of the Presence Questionnaire (sPQ) [105], both questionnaires rated on a seven-point Likert scale (where 1 = “Strongly disagree”, to 7 = “Strongly agree”). Specifically, presence was measured with the following questions from TPI questionnaire: *How much did it seem as if the objects and people you saw/heard had come to the place you were? (PLACE), How much did it seem as if you could reach out and touch the objects or people you saw/heard? (TOUCH), How often when an object seemed to be headed toward you did you want to move to get out of its way? (OBJECT), To what extent did you experience a sense of being there inside the environment you saw/heard? (BETHERE), To what extent did it seem that sounds came from specific different locations? (SNDLOCAL), How often did you want to or try to touch something you saw/heard? (TOUCHSMG), Did the experience seem more like looking at the events/people on a movie screen or more like looking at the events/people through a window? (WINDOW).* Likewise, from sPQ, that is composed by representative items that evaluate control, sensory, and realism factors, the following questions were used: *How natural did your interactions with the environment seem? (3), How compelling was your sense of objects moving through space? (10), How much did your experiences in the virtual environment seem consistent with your real-world experiences? (12), How completely were you able to actively survey or search the environment using vision? (14), How quickly did you adjust to the virtual environment experience? (26), How proficient in moving and interacting with the virtual environment did you feel at the end of the experience? (27), and How well could you concentrate on the assigned tasks or required activities rather than on the mechanisms used to perform those tasks or activities? (30).*

**Method of collecting ratings.** To rate video quality, observers used a developed application that allowed for visualizing contents and rating them subsequently without having to remove the HMD or interact outside the XR environment [96]. To answer questionnaires sPQ and TPI, they did it with online web application.

**Usability.** Items 3, 27, and 30 were used to evaluate usability rating video quality in the virtual environment with controllers and touchpad.

### 3.3.3 Equipment and environment

Subjective tests were carried out in two popular HMDs: Samsung Galaxy S8 with Samsung Gear VR, which includes a touchpad on its right side, and Lenovo Mirage Solo with a daydream handheld controller (Table 3.4). The observers were located in the middle of a room, being able to spin around without any limitation while seated on a swivel chair.

TABLE 3.4: Overview of the two tests depending on the order of the experimental conditions assessed in the experiment.

|               | 1st Condition       | 2nd Condition       |
|---------------|---------------------|---------------------|
| <b>Test A</b> | Samsung + touchpad  | Lenovo + controller |
| <b>Test B</b> | Lenovo + controller | Samsung + touchpad  |

### 3.3.4 Test session

Half of them took test A and the other half test B (see Table 3.4). The order of the conditions is the unique difference between them and the structure of each condition can be seen in Figure 3.3. At the beginning, subjects received the instructions for the experiment at the beginning of the test. Also, they were informed and signed a consent form that allowed us to process the information in accordance with the GDPR of the European Union. A training session to show the best and the worst qualities offered was carried out after the explanation. In this way, subjects also learned the use of the evaluation mechanism during the training session. All videos were viewed and evaluated on both devices by every subject. The duration of the whole test was around 40 minutes (20 minutes for each condition session with a break in the middle). At the end of each condition session, the presence and usability questionnaires were evaluated.

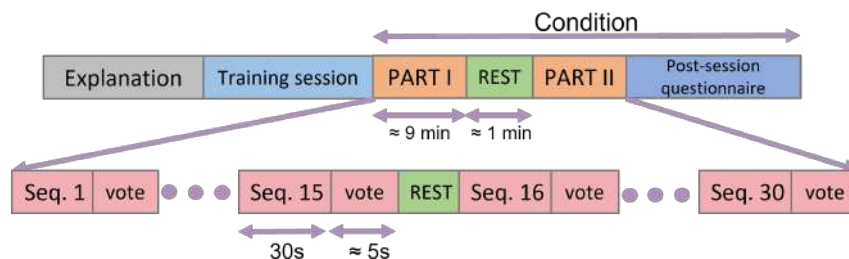


FIGURE 3.3: Test session structure.

### 3.3.5 Participants

A total of 48 observers (21 females, 27 males), with an age between 20 and 26, participated in this experiment. All of them with normal or corrected vision.

### 3.3.6 Hypotheses

The following hypotheses were investigated:

- The HMD and the order in which conditions are evaluated have influence on: sense of presence (**H1**), quality (**H2**), and perceived usability (**H3**).
- sPQ and TPI provide similar measurements (scores) for spatial presence (**H4**).

## 3.4 Experimental results

The quality evaluation was examined with the MOS and the DMOS from the observer's rates, and the associated 95% CI [116]. In relation to the sPQ and TPI analysis, Pearson & D'Agostino normality test was applied to validate the normal distribution of the collected data. Then, the 2-way Analysis of Variance (ANOVA) was applied to examine the differences between conditions [117]. Figure 3.4 presents the mean scores of TPI, sPQ, Usability, and quality (MOS). Note that MOS was rescaled to a 7-level scale, allowing the comparison between the evaluated assessments in both tests and conditions.

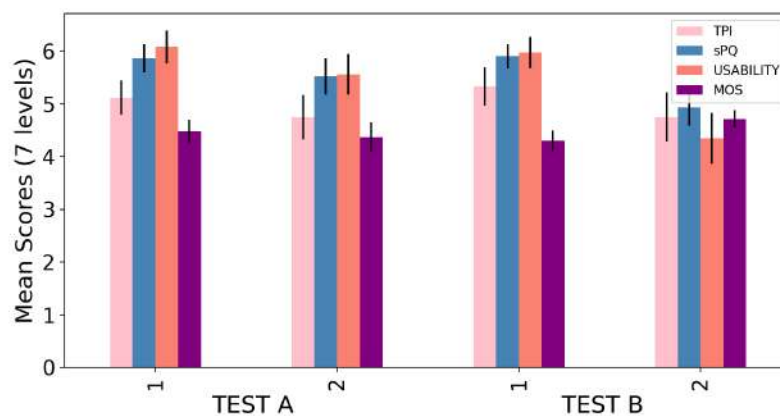


FIGURE 3.4: Average of the ratings of the participants of TPI, sPQ, Usability, and video quality (y-axis) with the associated 95% CI on a seven-level scale for each condition (x-axis).

The first hypothesis sets out how the HMD and the evaluation order affect the sense of presence evaluated with TPI and sPQ independently. The TPI and sPQ results showed that there is not

a significant effect on the test condition ( $F_{1,46} = 0.266$ ,  $p > .05$  and  $F_{1,46} = 3.896$ ,  $p > .05$  respectively), but there is a significant influence on the order of the conditions ( $F_{1,92} = 5.507$ ,  $p < .05$  and  $F_{1,92} = 17.761$ ,  $p < .05$ ). It makes sense because the sense of presence is mainly affected by the novelty of the 360-degree experience.

The second hypothesis refers to quality evaluation. Unlike the sense of presence, the quality ratings showed that there is not a significant difference between the evaluation during the first or second condition of the test ( $F_{1,92} = 1.662$ ,  $p > .05$ ). In this way, we can assure that subjects perfectly discern between the evaluation of the sense of presence and quality. Likewise, quality evaluation is highly dependent on the used HMD ( $F_{1,92} = 5.102$ ,  $p < .05$ ) where Samsung HMD obtained the best quality evaluations.

The third hypothesis is related to the usability of touchpad or controller. It was measured with the aggregation of three items of the sPQ. This aggregated measure showed a statistically significant difference between conditions (touchpad or controller), ( $F_{1,92} = 8.351$ ,  $p < .001$ ). However, it showed a stronger effect taking into account the order of the evaluation, first or second condition of the test session ( $F_{1,92} = 31.517$ ,  $p < .001$ ), as well a significant effect on the interaction of both conditions ( $F_{1,92} = 11.976$ ,  $p < .001$ ), as presented in Figure 3.4.

Finally, the fourth hypothesis formulates the comparison between the obtained results with the sPQ and TPI. For that, we computed the Linear Pearson Correlation Coefficient (LPCC). We showed that there is a relevant correlation between both questionnaires ( $r = 0.5279$ ,  $p < .05$ ) but based on the differences, as can be observed in Figure 3.4, we can not assume that sPQ and TPI measures exactly the same concept.

### 3.5 Conclusions

We conducted an experiment combining the video quality evaluation, well-known in the literature, with a characteristic of 360-degree technology: the evaluation of presence. As a result, we provided a repository [118] that contains:

- Dataset of video sources with the associated objective metrics results (PSNR, WS-PSNR, CPP-PSNR, VMAF, SSIM, MS-SSIM) and details (Spatial and Temporal Indicators [36], resolution, framerates, and brief descriptions).
- Head tracking data and video quality rates obtained from 48 participants during free-viewing experiments with two HMDs: Samsung GearVR and Lenovo Mirage Solo.

- Presence questionnaire scores, specifically TPI (Lombard et al.) and PQ (Witmer & Singer), obtained from 48 participants.
- Statistical analysis notebook.

This work was a pilot study for the followings experiments. Following the literature, we corroborated that it is difficult to evaluate socioemotional aspects in short-duration clips where there is neither narrative nor context. In addition, the fact of repeatedly visualizing the same clips in different qualities and with two devices, made the participants initially evaluate the aspects related to presence in a positive way but nevertheless, as the experiment session progressed, the sense of presence decreased notably, what we call as fatigue effect. It was a relevant motivation to explore in depth methodologies that cover the evaluation of video quality and socioemotional aspects and the interaction between them.

Some participants after the session told the researcher responsible for the experiment that the presence was highly dependent on the content. As we had not collected this information in a structured way, we considered in next experiments higher-level aspects such as acquisition perspective, camera location, and interactive elements that could influence socioemotional aspects.

The results of the usability, help us to select the handheld controller as a method of collecting ratings in future experiments to increase the comfort of the observers. As member of VQEG, the conclusions of this pilot study were a contribution to create ITU-T Rec. P.919 for the quality evaluation of VR, a new standard for subjective test methodologies for 360-degree video on HMDs [37].

Video quality evaluations obtained for each of the qualities, determined by encoding with a fixed QP, serve as a reference for future experiments. In this way, video quality ratings allow us to have a reference to compared with when new methodologies are tested.

## Chapter 4

# Subjective Video Quality and Socioemotional Aspects Assessment

### 4.1 Introduction

Considering the post-pandemic situation and the relevance of teleconferencing scenarios, VR technology can foster a change in communications. However, it is necessary to further investigate and provide standardized methodologies to consider all aspects that influence the QoE for the final boost of this technology [33, 119]. Once there is literature working on the analysis of different socioemotional and technical aspects, an important advance should be to evaluate them together in experiments closer to real scenarios and use cases, increasing ecological validity and reliability. In addition, experiments that consider aspects that have already been independently evaluated save time and resources. So, in this chapter, we not only present an experiment with a methodology designed to evaluate both technical and socioemotional features; we launch a renewed point of view: how the evaluation of technical aspects influences socioemotional features, and vice versa, accelerating immersive communications as a solution for hybrid meetings where some participants attend a meeting remotely and other participants in person.

Here, the structure of the chapter is presented. Section 4.2 presents the literature related with the work conducted in the test. Section 4.3 describes the main parts of a experiment design: research questions, experimental conditions, stimuli, methodology, equipment and environment, test session, and participants. Section 4.4 describes the main results and finally, Section 4.5 presents general conclusions that are relevant input for the decisions in the following steps of the research.

### 4.2 Related work

The studies conducted in the literature present limitations that influence the results and conclusions and should be considered for the design of the experiment. In this section, we present an overview of the works mainly related to the socioemotional features assessment.

Based on the literature and the pilot study presented in Chapter 3, video quality assessment methodologies, which were originally designed for 2D video, have been used in 360-degree video experiments and, somehow, adapted to address the new perceptual factors involved in VR [53, 120], such as simulator sickness or exploration behavior. However, there is a research line supporting that quality assessment should be done under the most realistic conditions when services and applications are addressed to end users [121, 122].

One relevant aspect to take into account when selecting 360-degree content is its characterization in terms of exploration properties [123]. Generally, contents can be classified as directed or exploratory. Directed videos can help the observer to guide attention in the scene. Although participants move freely around the scene in exploratory contents, most of them fully explore the whole scene (360-degree) in 20 seconds [124]. Nevertheless, contents of long-duration can improve the engagement and enhance the emotions of the participants [44, 125]. In addition to the duration, the genre and context of the video influence the success of the research. Specific genres of content should be considered based on the socioemotional features addressed in the experiment [124], e.g., horror stimulus to test fear [43]. Taking this into account, we examined some 360-degree datasets in the literature with different characteristics. For example, Li et al. [126] released a public database of 360-degree videos covering a wide range of arousal and valence. Also, Jun et al. [124] published a dataset containing 80 videos that were used to investigate a set of socioemotional features with a sample of 551 participants. They provided video sources with the corresponding report ratings and head movements. In addition, there are several datasets created to analyze exploration behaviors of the users when watching the content, such as the ones from Corbillon et al. [127] and Lo et al. [128] providing also head-movement data, or the one from David et al. [61] that includes both head and eye tracking data. Regarding quality evaluation, some annotated datasets have been published, mainly containing short-duration videos, such as the one from Yang et al. [129].

A great effort has been made in the analysis of socioemotional features in VR. Riva et al. [130] demonstrated the effectiveness of VR as an *affective medium*, a medium able to elicit different emotions through the interaction with its contents. Furthermore, the study demonstrated that the perceived sense of presence, related to a sense of being in a place [100], influences the emotional state. Following this research line, many studies have already confirmed the ability of VR to create more immersive environments, improving the socioemotional features. For instance, Fonseca et al. [44] demonstrated the highest emotional involvement of the participants viewing two types of narrative 360-degree contents with an HMD. MacQuarrie et al. [43] obtained a significant improvement of enjoyment of users using the HMD. In addition, VR emphasizes the phenomenon called Fear of Missing Out (FoMO) [131], defined, in the context of VR, as the apprehension that others might be having rewarding experiences from which the user with the HMD is absent. Additionally, users

can freely move around the virtual environment, selecting the most interesting area of the 360-degree scene to focus on. These factors (immersion, FoMo, user motion pattern) may influence the attention that users pay to the events and objects in the scene. Some works in the literature analyze methods for assessing attention in this kind of environment [46, 132].

Several works go one step further analyzing the use of this technology for empathy purposes and even for behaviour change purposes. Empathy is defined as the ability to view the world from another person's perspective combined with an emotional reaction to that perspective, including feelings of concern for others [41]. These studies are based on the fact that involvement created by immersive environments facilitates empathy for users and can be used for specific purposes [133]. Aitamurto et al. [102] evaluated the responsibility for resolving gender inequality visualizing a 360-degree content in which participants could choose to watch the narrative from the male or female character's perspective. Likewise, Tussyadiah et al. [134] confirm the effectiveness of VR technology in shaping consumers' attitude and behavior for tourism purposes.

Based on the constraints presented, mainly related with content, methodologies, and context, we decided the design and acquisition of the 360-degree contents taking into account the purposes and the final devices used in the experiment [43]. With this, we propose a quality evaluation on long-duration videos with a context that interests or affects the participant and with a genre selected according to the purpose of the research. In the same experiment, we assess video quality and several socioemotional features, reporting technical features, questionnaires, and sample diversity. Also, we choose an environment where the participant is isolated, facilitating the real world disassociation.

## 4.3 Experiment design

### 4.3.1 Research questions

Based on the previous analysis, we pose the following Research Questions (RQs):

- RQ1: Is it possible to evaluate video quality in videos of long-duration designed for the evaluation of socioemotional features?
- RQ2: Which technical aspects, such as the position of the camera, the type of conversation, the video quality or the acquisition perspective influence socioemotional features?
- RQ3: Which interactive elements can be provided to the remote client to improve some socioemotional aspects such as presence or attention?

To answer these RQs, we designed a subjective experiment where an immersive communication between a *provider* and a *remote client* was simulated, presented in Figure 4.1.

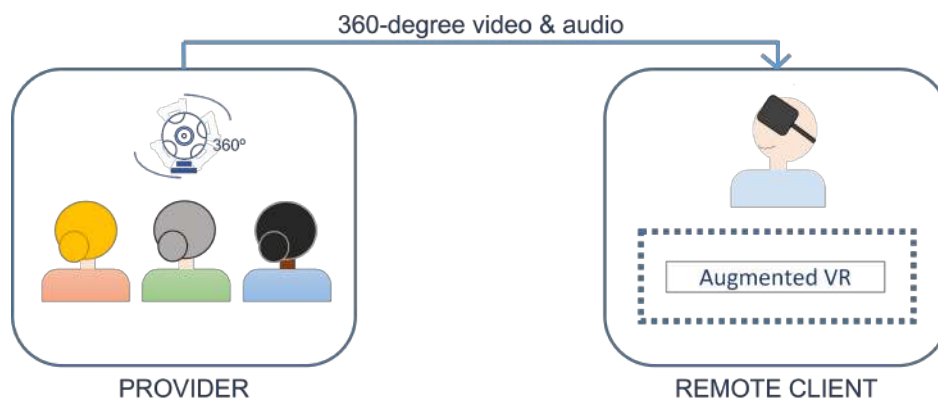


FIGURE 4.1: Simulated immersive communication environment of the experiment. 360-degree video and audio recorded on the provider side is visualized by the remote client. Participants assigned to condition C can see their hands and take notes on a physical whiteboard (Augmented VR).

At the provider side, a conversation among several people took place, and the remote client attended virtually wearing an HMD. In the subjective test, the observer took the role of the remote client and visualized pre-recorded 360-degree videos with fluctuations of quality, simulating a VR streaming communication.

The contents used in the experiment showed simulated conversations around a common topic: international experiences, e.g., working or studying abroad. The main idea behind choosing this specific context was our ability to gather a balanced sample of people who have had international experiences and with people who have not. We acquired 360-degree videos with different acquisition perspectives (actor, when the camera is located from the first person point of view or observer, when the camera is located from the third person point of view. and observer) and genre (everyday conversation, educational, and discussion). For that, student volunteers were recruited for the recordings, both exchange and national students from the university, making the conversations more realistic and fluent. Conversations were in English, making the experiment accessible to different nationalities and mother tongues and increasing the diversity of the sample.

### 4.3.2 Experimental conditions

The experiment considered three test conditions, summarized in Table 4.1, and each participant was assigned a condition. However, in all conditions, participants visualized the same video, with the same fluctuations of the quality. After each video, they were requested to rate its visual quality, as well as to evaluate the socioemotional features of interest: empathy and attitude, spatial and social presence, and attention.

TABLE 4.1: Overview of the three experimental conditions with the associated interactive element and features assessed in the experiment.

| Condition | Assessment |                | Interactive element |
|-----------|------------|----------------|---------------------|
|           | Quality    | Socioemotional | Hands               |
| A         | X          | X              |                     |
| B         |            | X              |                     |
| C         |            | X              | X                   |

Participants assigned to condition A had the additional task of periodically rating the visual quality of the video during its playback, whenever its quality changed. This is a conventional design to evaluate the subjective quality of the video sequence under different intensities of impairment. However, this focused task might have impact on the evaluation of socioemotional features compared to the baseline scenario without the task (condition B).

Finally, participants in condition C were provided with an additional interactivity element: the possibility to see their own hands and take handwritten notes about the conversation, as shown in Figure 4.2. We hypothesize that this could enhance socioemotional features such as presence and attention with respect to the other conditions.



FIGURE 4.2: Participant of condition C taking notes on a physical whiteboard. The photo was taken at the environment of the experiment.

### 4.3.3 Stimuli

The set of source videos, Student Experiences Around the World dataset (SEAW-dataset) consists of three stereoscopic contents in 4K resolution at 30 fps and a duration of approximately five minutes each were acquired and prepared specifically for the experiment. Figure 4.3 shows a screenshot of the source videos and the original ones have been made publicly available [2].



FIGURE 4.3: Video sources screenshots in equirectangular projection [2].

As it can be observed in all sequences, student volunteers were sitting around a table far enough from the camera to avoid stitching problems affecting the user’s QoE and video quality scores. In addition, the camera was placed at the position and average height of the head of a person sitting at the same table, facilitating the engaging experience [135]. Table 4.2 summarizes the genre, perspective-taking, and a brief description of the contents used in the experiment. In contents with the actor acquisition perspective, student volunteers during the recording looked at the camera, and even waved their hands to increase the immersion of the participant of the experiment visualizing the 360-degree content with the HMD.

TABLE 4.2: Semantic characterization of the 360-degree videos considered in the experiment.

| Name                        | Genre                 | Acquisition perspective | Description   |
|-----------------------------|-----------------------|-------------------------|---|
| <b>Coffee shop</b>          | Everyday conversation | Observer                | A coffee conversation between foreign and local students about cultural differences             |
| <b>International office</b> | Educational           | Actor                   | A presentation given by a professor to students about the foreign application process           |
| <b>Study in Spain</b>       | Discussion            | Actor                   | A conversation about the differences between transport and rental prices in different countries |

The SRCs were encoded with HEVC switching to a different fixed QP each 25 seconds to create one PVS per source content [25]. The QPs selected for the experiment were: 15, 22, 27, 32, 37, and 42 [115]. These QPs were randomized along the video encoding, following ITU-R Rec. BT.500-14 [25]. Based on the assumption that each video source maintains the features in terms of color, texture, composition, and light, participants rated the quality of each one of the 25-second units along the whole sequence, avoiding the repetition of the same clip [125]. Due to the duration of the contents, each QP was rated at least two times in each PVS. Finally, the original audio quality was maintained through the experiment, improving the immersion and the QoE of the observers [103].

#### 4.3.4 Methodology

Here, we explain in detail the methodology considered in the experiment. Table 4.3 summarizes the items evaluated in the three experimental conditions.

TABLE 4.3: Structure of the test session questionnaires.

| Condition | Pre-questionnaire (once) |               |                    | During each content | Post-questionnaire (for each content) |                  |                    |  | Notes |
|-----------|--------------------------|---------------|--------------------|---------------------|---------------------------------------|------------------|--------------------|--|-------|
|           | Personal information     | Empathy (IRI) | Attitude (EM1-EM4) | Quality (SSDQE)     | Quality (ACR)                         | Attention survey | Attitude (EM5-EM8) | Spatial Presence (PP1-PP5) & Social Presence (SP1-SP5) |       |
| A         | X                        | X             | X                  | X                   | X                                     | X                | X                  | X  |       |
| B         | X                        | X             | X                  |                     | X                                     | X                | X                  | X  |       |
| C         | X                        | X             | X                  |                     | X                                     | X                | X                  | X  | X     |

**Personal information.** For each participant, we collected age, gender, vision (corrected or normal), nationality, experience living in a foreign country and which one, and English level. This was used to characterize our observers and guarantee diversity.

**Empathy.** The initial empathy of each of the observers was evaluated using the Interpersonal Reactivity Index (IRI) [41]. This questionnaire is a psychometrically invariant empathy measure based on 28 statements related to the Perspective-Taking scale (PT), Fantasy Scale (FS), Empathic Concern scale (EC), and Personal Distress scale (PD). For each statement, the observer was required to indicate how well it described her/him on a five-level scale (where 1 = “Does not describe me well”, to 5 = “Describes me very well”).

**Attitude.** A survey was designed to measure the attitude towards the context of the videos, international experiences. As there was no validated questionnaire to measure the attitude of the participants towards other cultures and foreigner experiences, we decided to apply the Facet theory [136]. Facet theory consists of distinguishing the facets in which the designers of the experiment are interested. From the identified facets, the items of the questionnaire are defined and associated. In our case, we identified four characteristics that a person with a positive attitude towards foreigners and other cultures must have: interest, tolerance, respect, and social sensitivity. We established four statements related to the **interest**, **respect**, **tolerance**, and **social sensitivity** towards other cultures and traditions. These four items were evaluated before starting the session and after the visualization of each of the three videos analyzed in the experiment. In this way, we could compare the empathy and attitude evolution throughout the session. To do this, four questions (EM1-EM4) were designed for the first evaluation at the beginning of the test, and another four questions (EM5-EM8) for the evaluation after each video. The idea behind this design was to compare the ratings before the visualization of the 360-degree content and after it. Observers provided ratings on a seven-level Likert scale (where 1 = “Strongly disagree”, to 7 = “Strongly agree”) based on most works in the literature [53, 102]. Specifically, it was measured with the following questions: *I just need to know the traditions of my country of origin (EM1)*, *I think that some traditions of other cultures should not be allowed in my country (EM2)*, *I feel comfortable*

with traditions different from mine (EM3), and I am worried about the experiences of foreign people in my country (EM4), I like participating in this kind of conversation (EM5), I think that an intercultural society has a positive impact for the people (EM6), I would feel comfortable sharing traditions of other cultures (EM7), I would like to participate in a buddy program or in a project to know more about foreign experiences in my country (EM8).

**Quality.** A Single-Stimulus Discrete Quality Evaluation (SSDQE) method [38] was applied to measure the quality in observers assigned to condition A. SSDQE uses long-duration contents to evaluate quality guaranteeing the continuity of the narrative. For this, the content is divided into segments, trying to mimic realistic situations of video consumption. As represented in Fig. 4.4, impairments are inserted throughout the content used as stimuli (PVS) in alternate segments (“processed segments”) and participants rate the perceived quality during the following ones (“evaluation segments”). Note that during the evaluation segment, video playback continues and is encoded with the same quality as the previous processed segment. Specifically, they evaluated the quality on a five-level quality scale [25], where the categories: “Bad”, “Poor”, “Fair”, “Good”, and “Excellent” were displayed on the screen. Additionally, the aggregate quality was asked, following the literature, in the post-questionnaire using the Absolute Category Rating (ACR) on the same five-level scale [137, 138] with the item *Please, rate the quality of the experience.*

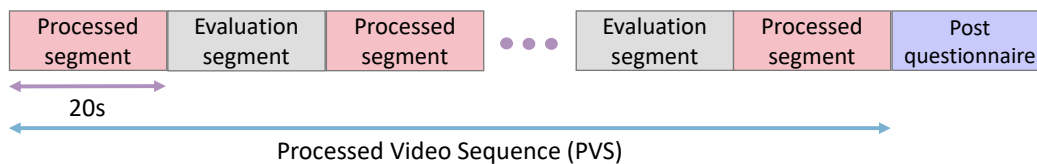


FIGURE 4.4: Structure of the test sequences used with SSDQE methodology.

**Attention.** Observer attention was assessed with three questions about the conversations taking place in the videos that had pass/fail answers [43, 132]. For each content, we designed a multiple-choice question, a short answer question, and a True/False statement, presented in Table 4.4. In this way, participants scored zero or one point for each correct answer, resulting in a maximum score of three points for the total attention score for each video.

TABLE 4.4: Attention survey: True/False statement, short answer, and multiple choice question for each of the 360-degree videos.

| Content              | True/False statement  | Short answer question  | Multiple-choice question   |
|----------------------|---|--|--|
| Coffee shop          | Students think that the second year is easier                                     | How long ago has the university education system in Spain changed? | What city are the exchange students from?                                      |
| International office | All students can apply for an internship both in research groups and in companies | The deadline to apply for double degree students                   | The deadline to apply for an internship or exchange program for the whole year |
| Study in Spain       | Norwegians spend on average less money on public transport                        | The price of the public transport card in France                   | The rent per month in Norway   |

**Presence.** Spatial and social presence experienced by the observers were evaluated with five questions obtained from the state of the art [102, 130]. The questions of social presence questionnaire were mainly related to factors that influence the involvement in the meeting, such as the feeling that people in the meeting are looking at us, talking to us or where is the group attention focused on [100]. Observers provided ratings on a seven-level Likert scale (where 1 = “Strongly disagree”, to 7 = “Strongly agree”). Specifically, spatial presence was measured with the following questions: *I felt I was present in the places shown in the video (PP1)*, *I felt surrounded by the actions in the video (PP2)*, *I felt I was sitting by the table at the place of the video (PP3)*, *I felt I could have reached out and touched the items on the table of the video (PP4)*, and *I felt that all my senses were stimulated at the same time (PP5)*. Likewise, social presence was measured with the following ones: *I felt that people were talking to me (SP1)*, *I felt that I was listening to the others in the video (SP2)*, *I felt I was present with the other people in the video (SP3)*, *I felt like the people in the video could see me (SP4)*, and *I felt I was actually interacting with other people (SP5)*.

**Notes.** Participants assigned to condition C assessed the usefulness of the notes taken during the test session. Specifically, the question *Have your annotations helped you to solve the questions?* was used to consider whether the correct answers were correct from the annotations or from the memory of the participants.

### 4.3.5 Equipment and environment

All participants visualized the contents with a Samsung Galaxy S8 and the last model of Samsung Gear VR headset endowed with head tracking. The maximum resolution that viewers could perceive with this HMD (assuming a field of view of  $85^\circ \times 100^\circ$  and a smartphone native resolution of 1440x2960 pixels), is about 680x822 pixels [19]. Monophonic audio was heard through headphones.

In all conditions, A, B, and C, the questionnaires were presented and answered using a web application. Observers who were assigned to condition A evaluated the quality of the video during the session. For this purpose, a VR application that allows users to visualize contents and answer customized questionnaires without having to take off their goggles was used [96], the same than in previous experiments. Likewise, based on the previous experiment conclusions, they used a hand-held controller as the evaluation method, avoiding any sign of discomfort [53]. Observers assigned to condition C were able to see their own hands, as well as a small whiteboard to take some notes, using an augmented virtuality approach, as shown in [139]. The local environment was captured by the smartphone camera and displayed in front of the 360-degree video. Background was removed from the camera image using chroma-keying based on red chrominance.

Regarding the local environment, the observers were seated in a swivel chair in front of a table. This chair allowed them to spin around without more limitations than the three degrees of freedom, imposed by the HMD. The table in front of them was a requirement imposed by the videos, since, as presented in Figure 4.3, the three contents simulate a meeting around a table. In this way, observers could identify the table of the videos with the real one. Additionally, participants were located in totally isolated cubicles, facilitating the immersion in the content and avoiding any external distraction that increases the sense of FoMO.

### 4.3.6 Test session

The test session structure is presented in Figure 4.5. At the beginning, participants received a brief explanation of the experiment. Also, they were informed and signed a consent form that allowed us to process the information in accordance with the GDPR of the European Union. The experiment started with the pre-questionnaires: a personal information survey, the empathy questionnaire (IRI), and the initial attitude survey. The training session consisted of a visualization of a discussion around the table with a duration of two minutes approximately. The PVS used for the training sessions was encoded with the best and worst qualities offered in the experiment (QP values of 15 and 42) every 25 seconds. Observers assigned to condition A tested the evaluation method with the handheld controller. After the training session, the assessment session started. All participants visualized the same three PVS in a randomized order following ITU-R Rec. BT.500-14 [25].

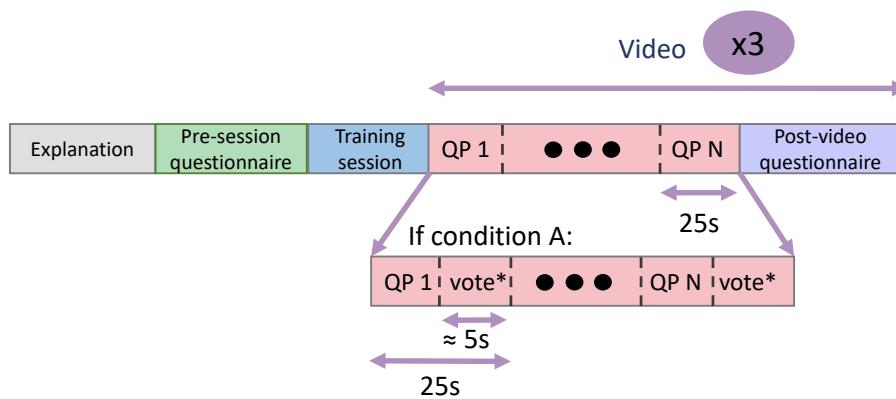


FIGURE 4.5: Test session structure.

For observers assigned to condition A, every 20 seconds the SSDQE question appeared without a time limit. After each video, all participants, regardless of the assigned condition, answered a post-questionnaire with questions about the quality and the socioemotional features. Here, participants assigned to condition C also answered the notes question.

### 4.3.7 Participants

A total of 54 observers (20 females, 34 males) took part in this experiment. There were participants in the age range between 17 and 26 years, with a Mean age (M) of 22.18 and a Standard Deviation (SD) of 1.95. All observers were checked for normal or corrected-to-normal vision. All participants were required at least an intermediate level of English to understand the conversations of the videos. They received a small financial reward for participating. In this way, we obtained a sample of participants with international experiences or nationalities from 15 countries in Europe, America, and Asia. The representation of user diversity was an additional value of the experiment, since it increased its reliability [140, 141]. Furthermore, as it can be observed in Figure 4.6, participants with international experiences and taking into account gender were distributed almost uniformly under conditions A, B, and C, guaranteeing a balanced sample.



FIGURE 4.6: Observers distribution in conditions A, B, and C taking into account the gender and international experience.

## 4.4 Experimental results

For each one of the RQs, one or more hypotheses have been laid out and investigated, to look for relevant conclusions. Besides, the methodology to analyze the results was performed according to

the nature of the data. The quality evaluation in condition A was examined with the MOS and the associated 95% CIs obtained from the scores, presented in Figure 4.7. In regard to the quality and socioemotional features, the Pearson & D’Agostino normality test was computed to validate the normal distribution of the collected data. For cases where the distribution was normal, the 2-way ANOVA was performed to examine the differences among the evaluated videos and conditions. For social and spatial presence, due to the condition of non-normality, the following transformation was implemented:  $\arcsin(\sqrt{P/7})$ , where  $P$  is the presence rating and it is divided by seven because social and spatial presence were evaluated in a seven-level scale. Once the data was transformed, it was analyzed under the normality condition. Post-hoc analyses using Bonferroni correction for multiple comparisons were applied to examine the differences among the evaluated videos and conditions. The considered level of significance was 0.05. Table 4.5 and Table 4.6 present a summary of the scores of the items evaluated in the experiment and the significance ( $F$ ,  $p$ , partial eta-squared  $\eta_p^2$ , and observed power values  $\gamma$ ) between conditions and contents, respectively. Additionally, Cronbach’s  $\alpha$  obtained for the questionnaires used in the experiment about spatial presence, social presence, and attitude are presented in Table 4.7.

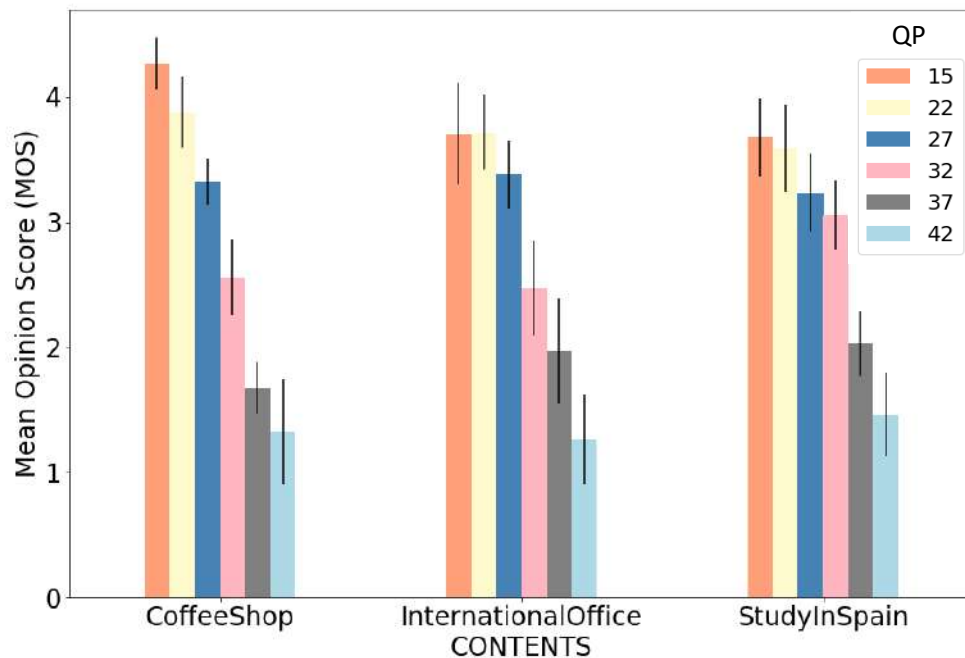


FIGURE 4.7: The mean opinion scores (y-axis) on a five-level scale obtained from 17 participants assigned to condition A who evaluated the perceived video quality in the processed segment, encoded with specific QP, every 20 seconds while visualizing each of the three contents (x-axis), following the SSDQE methodology. Error bars represent 95% CI.

### Video quality assessment

Regarding RQ1, we investigated the first hypothesis (**H1**): video quality evaluation can be adapted to long-duration videos designed for socioemotional features assessment purposes.

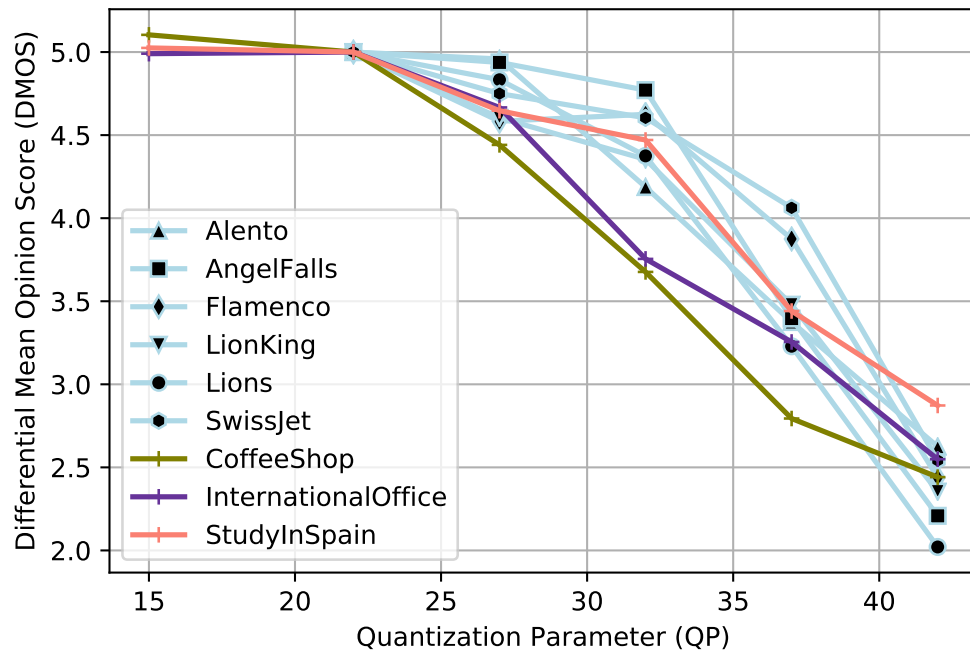


FIGURE 4.8: Comparison of DMOS (y-axis) on a five-level scale obtained from 17 participants assigned to condition A, and from participants from the pilot study. Both participants evaluated clips of short duration encoded with fixed quantization parameters (x-axis). However, participants assigned to Condition A evaluated perceived quality following SSDQE methodology and participants from the pilot study evaluated it following ACR methodology.

In this sense, we were interested in analyzing the effect of evaluating the quality of the video during the visualization of continuous sequences in which the scene features remain similar. Figure 4.7 was obtained from the scores of the 17 participants assigned to condition A. Note that the ratings of one of the observers were not collected correctly, so we remove an observer from condition A. As the evaluation of video quality on a 5-level score can be modeled by a Gaussian random process [142], we use parametric analysis for the evaluation of the scores, following the common practices for video quality data evaluation and the previous experience [117]. We have performed a ANOVA to assess the dependency of the scores on each source video and each QP value. Results show that the QP is significant ( $F_{5,596} = 186.4$ ,  $p < .001$ ,  $\eta^2 = .598$ ), while the source content is not ( $F_{2,596} = 0.85$ ,  $p > .05$ ,  $\eta^2 = .018$ ). Bonferroni-corrected pairwise t-tests show that all pairs of QPs are significantly different between them, except the two higher qualities QP values of 15 and 22, which are not. Note that due to the different duration of the videos and the randomization of the QPs, each QP was not evaluated the same number of times.

These quality scores were compared with the MOS values obtained from the study presented in Chapter 3, which was executed using a conventional ACR methodology with randomized 10-second video sequences. As the source contents were different in both experiments, we computed the Differential Mean Opinion Scores (DMOS), according to ITU-T Rec. P.910 [36]. We used QP 22 as the hidden reference, as it was the highest quality available in the pilot study, and it was shown not to be significantly different from QP 15 in our new experiment. Figure 4.8 shows that both

methodologies offer comparable results: good distribution of the ratings and a consistent decrease of the perceived quality when augmenting the QP, as expected in this type of tests [52].

TABLE 4.5: Difference in aggregate quality and socioemotional features between the three conditions.

| Questionnaire items                                | Condition A                     | Condition B                     | Condition C                     | Significance   |
|--|---------------------------------|---------------------------------|---------------------------------|--|
| <b>Aggregate quality (5-level scale)</b>           | $M = 3.537$<br>( $SD = .719$ )  | $M = 3.111$<br>( $SD = 1.022$ ) | $M = 3.167$<br>( $SD = .885$ )  | $F_{2,153} = 3.687, p < .05, \eta_p^2 = .045, \gamma = .687$ |
| <b>Spatial Presence (7-level scale)</b>            | $M = 5.463$<br>( $SD = 1.019$ ) | $M = 5.185$<br>( $SD = 1.318$ ) | $M = 5.411$<br>( $SD = .942$ )  | $F_{2,153} = .900, p > .05, \eta_p^2 = .011, \gamma = .203$  |
| <b>Social Presence (7-level scale)</b>             | $M = 5.059$<br>( $SD = 1.398$ ) | $M = 5.133$<br>( $SD = 1.287$ ) | $M = 5.144$<br>( $SD = 1.271$ ) | $F_{2,153} = .005, p > .05, \eta_p^2 = .000, \gamma = .05$   |
| <b>Attitude post-questionnaire (7-scale level)</b> | $M = 5.875$<br>( $SD = .250$ )  | $M = 6.042$<br>( $SD = .108$ )  | $M = 6.236$<br>( $SD = .191$ )  | $F_{2,153} = 4.660, p < .05, \eta_p^2 = .055, \gamma = .782$ |
| <b>Attention (3-level scale)</b>                   | $M = 1.981$<br>( $SD = .765$ )  | $M = 1.833$<br>( $SD = .885$ )  | $M = 1.685$<br>( $SD = .748$ )  | $F_{2,153} = 1.839, p > .05, \eta_p^2 = .023, \gamma = .391$ |

TABLE 4.6: Difference in aggregate quality and socioemotional features between the three contents.

| Questionnaire items                                | Coffee shop                     | Int. office                     | Study in Spain                 | Significance   |
|--|---------------------------------|---------------------------------|--------------------------------|--|
| <b>Aggregate quality (5-level scale)</b>           | $M = 3.111$<br>( $SD = .904$ )  | $M = 3.222$<br>( $SD = .883$ )  | $M = 3.481$<br>( $SD = .885$ ) | $F_{2,153} = 2.485, p > .05, \eta_p^2 = .03, \gamma = .496$  |
| <b>Spatial Presence (7-level scale)</b>            | $M = 5.326$<br>( $SD = 1.173$ ) | $M = 5.200$<br>( $SD = 1.137$ ) | $M = 5.533$<br>( $SD = .991$ ) | $F_{2,153} = 1.394, p > .05, \eta_p^2 = .017, \gamma = .297$ |
| <b>Social Presence (7-level scale)</b>             | $M = 4.748$<br>( $SD = 1.364$ ) | $M = 4.752$<br>( $SD = 1.280$ ) | $M = 5.837$<br>( $SD = .964$ ) | $F_{2,153} = 15.710, p < .01, \eta_p^2 = .169, \gamma = 1$   |
| <b>Attitude post-questionnaire (7-level scale)</b> | $M = 6.111$<br>( $SD = .253$ )  | $M = 5.866$<br>( $SD = .234$ )  | $M = 6.176$<br>( $SD = .098$ ) | $F_{2,153} = 3.271, p > .05, \eta_p^2 = .038, \gamma = .605$ |
| <b>Attention (3-level scale)</b>                   | $M = 2$<br>( $SD = .777$ )      | $M = 1.704$<br>( $SD = .743$ )  | $M = 1.796$<br>( $SD = .877$ ) | $F_{2,153} = 1.925, p > .05, \eta_p^2 = .024, \gamma = .407$ |

These results show that subjects are able to effectively assess the video quality of individual QPs, and the content does not distract them from the task. This is in line with the results already reported in the literature for conventional 2D video and similar evaluation methodologies [38, 143, 144]. Furthermore, having the subjects engaged in the content increases the ecological validity of the quality evaluation compared to traditional methods [122, 145].

The aggregate quality scores rated at the end of each video in a five-level scale were analyzed statistically to find differences between videos and conditions. Due to the normality condition, 2-way ANOVA was applied. Table 4.6 shows that there is no significant difference between videos. MOS lay somewhere in the middle between the lowest and highest scores obtained for individual QPs, which is also expected [146]. It is known that several factors, such as the amplitude, frequency,

TABLE 4.7: Cronbach’s  $\alpha$  obtained for the questionnaires used in the experiment about spatial presence, social presence, and attitude.

| Questionnaire                          | Cronbach’s $\alpha$ |
|--|---------------------|
| Spatial Presence                       | 0.857               |
| Social Presence                        | 0.865               |
| Attitude ( <i>pre-questionnaire</i> )  | -0.094              |
| Attitude ( <i>post-questionnaire</i> ) | 0.710               |

and time location of the quality switches have an effect on the formation of the overall quality opinion [121], but addressing them is outside the scope of our experiment.

However, there is a significant difference among conditions, as seen in Table 4.5. Student’s t-test with Bonferroni correction shows that this difference is significant between conditions A and B ( $p = .0307$ ). Participants assigned to condition A scored the aggregate quality higher than participants assigned to condition B and C. It means that participants that are focused on the quality evaluation throughout the sequence, change their perspective about the perceived global quality.

To the author’s knowledge, this result is new in the literature. Some authors have used similar methods to evaluate the video quality continuously during the content playback, and then a single *endpoint quality* score at the end to assess the overall quality of the sequence [144, 147]. However, none of them has also had the same sequences evaluated just at the end, as it is proposed, for instance, by ITU-T [148]. Our results show that the evaluation of quality during the sequence has a significant influence on the endpoint quality.

### Spatial and social presence

In reference to RQ2, we investigated the second hypothesis (**H2**): acquisition perspective, type of the conversation, and experimental condition have influence on: spatial and social presence. As said before, the items of the sense of spatial presence and social presence were measured on a seven-level Likert scale independently. As presented in Table 4.5 and Table 4.6, the analysis was twofold. Once the non-normality condition of the social and spatial presence ratings was corrected, ANOVA was applied to analyze differences between experimental conditions. The aggregate measure of the five spatial and social presence items respectively show that there is not a significant difference. Nevertheless, a notable result is that the perceived social and spatial presence were very high in all conditions. In this sense, we want to point out that during the design of the experiment we presumed significant differences for condition C. We consider that the absence of differences is due to the fact that there were no specific tasks that required hands-on interaction with the virtual environment.

Likewise, ANOVA was applied to examine the differences between videos. The aggregate measure of the five items of spatial presence shows that there is no significant difference, but the aggregate measure of the five items of social presence does show a difference. Student's t-test with Bonferroni correction shows that there is a significant difference between "Study in Spain" and "Coffee shop" contents ( $Z = -4.887$ ,  $p < .01$ ) and "Study in Spain" and "International office" content ( $Z = -5.023$ ,  $p < .01$ ).

Table 4.6 shows that "Study in Spain" scored higher in social presence. To better explore the difference between contents, Wilcoxon Signed-Rank test with Bonferroni correction were applied to the items of the social presence questionnaire (SP1-SP5) [102]. The analysis shows significant differences between "Study in Spain" and the other videos, "Coffee Shop" and "International office", in questions related to the perception that people in the conversation speak, look at, and interact with the participant: SP1 ( $Z = 47.5$ ,  $p < .01$  and  $Z = 108$ ,  $p < .01$ ), SP3 ( $Z = 113.5$ ,  $p = .0007$  and  $Z = 136$ ,  $p = .015$ ), SP4 ( $Z = 115.5$ ,  $p < .01$  and  $Z = 108.5$ ,  $p = .0001$ ), and SP5 ( $Z = 107$ ,  $p = .0005$  and  $Z = 86$ ,  $p < .01$ ). The reason is that in "Study in Spain" content the actors appeal to the camera more frequently, emphasizing the non-verbal side of the conversation.

### Empathy and attitude towards international experiences

Following with RQ2, we investigated the third hypothesis (**H3**): acquisition perspective, type of the conversation, and experimental condition have influence on: empathy and attitude.

Firstly, IRI ratings were examined to obtain an adequate measure of the initial empathy of the participants, avoiding any deviation that may affect the subsequent analysis of the attitude. Given the condition of normality, ANOVA test was conducted to examine the IRI scores depending on gender and international experiences. It shows that there are not significant differences ( $F_{1,50} = .76$ ,  $p > .05$  and  $F_{1,50} = 3.838$ ,  $p > .05$ , respectively). Based on the literature [149, 150], we expected significantly higher scores for females than for males. In our case, there are not significant differences but on average females scored higher empathy than males both for participants with international experiences ( $M = 3.373$ ;  $SD = .228$  and  $M = 3.316$ ;  $SD = .237$ ) and for participants without international experiences ( $M = 3.246$ ;  $SD = .302$  and  $M = 3.182$ ;  $SD = .199$ ).

Secondly, the attitude was evaluated with the questionnaire asked after the visualization of each of the three PVS. Table 4.8 summarizes the obtained results for each facet in the post-questionnaires ("Post"). Note that the data presented in the table is calculated in the original seven-level scale. The attitude was measured with the aggregation of four items of the designed survey: interest, respect, tolerance, and social sensitivity. Table 4.6 shows that there is not a statistically significant difference between contents but Table 4.5 presents a significant influence of the condition in which the content was visualized. Participants assigned to condition C achieved the highest attitude

index, followed by participants from condition B and A. After finding that the condition greatly influences on the attitude, Student's t-test with Bonferroni correction was applied to find differences between conditions. They show that the significant difference is only between A and C conditions ( $Z = -3.146$ ,  $p = .002$ ). It makes sense because participants assigned to condition A had the video quality assessment task, distracting them from the conversations taking place in the video. From this analysis, another main result is that there is an important positive impact in the three videos, and as presented, in the three conditions.

TABLE 4.8: The mean and standard deviations on a seven-level scale of the items of the attitude survey: interest, respect, tolerance, and social sensitivity in the three experimental conditions.

| Condition | Post-questionnaire |              |              |                    |
|-----------|--------------------|--------------|--------------|--------------------|
|           | Interest           | Respect      | Tolerance    | Social Sensitivity |
| A         | $M = 5.167$        | $M = 6.685$  | $M = 6.407$  | $M = 5.241$        |
|           | $(S = 1.411)$      | $(S = .571)$ | $(S = .806)$ | $(S = 1.17)$       |
| B         | $M = 5.389$        | $M = 6.574$  | $M = 6.333$  | $M = 5.870$        |
|           | $(S = 1.177)$      | $(S = .71)$  | $(S = .861)$ | $(S = 1.171)$      |
| C         | $M = 5.537$        | $M = 6.704$  | $M = 6.407$  | $M = 6.296$        |
|           | $(S = 1.343)$      | $(S = .565)$ | $(S = .913)$ | $(S = .936)$       |

### Interactive element

Finally, to answer RQ3 we investigate the fourth hypothesis (**H4**): Observers who can take notes get higher total attention scores. Participants scored one point for each correct answer, resulting on a scale from 0 to 3. Due to condition of normality, ANOVA test was applied to find differences between conditions. As presented in Table 4.5, the scores show that there is no a significant difference between participants assigned to condition A, B, and C. Among the 18 participants assigned to condition C, 10 of them reported that the notes they had taken helped them to answer the questions. However, their scores are not significantly different from the ones obtained by the other 8 participants.

## **A method to simultaneously assess video quality and socioemotional features**

Our experiment shows that the methodology used for condition A is suitable for the simultaneous evaluation of video quality and socioemotional features. As presented, SSDQE is valid to evaluate individual quality variations. Additionally, SSDQE does not affect the evaluation of presence or attention, which has two implications: on the one hand, it confirms that socioemotional features can be assessed despite having the extra task of continuous video quality evaluation; on the other, it shows that SSDQE does not reduce the observer immersion, making it a real content-immersive method.

There are, however, at least three caveats. First, using SSDQE does affect the evaluation of the overall quality of the sequence. Results obtained using this method will not be exactly the same as assessing the quality just with an endpoint evaluation. Second, using SSDQE during the video has some impact on the attitude of the observers. This means that, although the simultaneous evaluation of quality and socioemotional features is possible, it is not completely neutral, and some interaction between evaluation tasks may exist. Finally, it is worth noting that the experiment has been done with a specific type of content and visualization (360-degree videos simulating conversations on international experiences). Other types of videos or visualization setups might have different behavior.

## **4.5 Conclusions**

We have proposed and validated a methodology to jointly assess video quality and presence, empathy, attitude, and attention in immersive communications. We have simulated that users attend meetings remotely with the HMD and all meetings are focused on the international experiences context. Also, we have evaluated three conditions for the attendants. This methodology is a solution for experiments in a controlled environment but more realistic than those presented in the literature. We have proposed the use of the SSDQE method to measure the quality during the test session and the aggregate quality in a post-questionnaire using the ACR on the same five-level scale. Spatial and social presence were evaluated with an aggregate score obtained from 5 items based on the literature and adapted to our experimental environment. The initial empathy was evaluated using the IRI. The attitude was measured in pre-questionnaire and post-questionnaire designed using facet theory. Due to the reliability of the scale, we proposed to use only the post-questionnaire. The attention was addressed with three questions about the scene that have pass/fail answers.

We can conclude that video quality assessment can be adapted to conditions imposed by socioemotional feature methodologies, such as contents of longer duration where the scene background is

mainly static. This is an important contribution to the state of the art, since it shows that methodologies can be designed to simultaneously evaluate technical features and socioemotional features that go one step further. Thus, it allows this type of experiment in more realistic environments with final VR applications.

The prototype evaluated for VR communications provides high scores in terms of social and spatial presence. Significant differences in the sense of social presence have been obtained between sequences. Then, we can assure that social presence is highly influenced by the acquisition perspective, narrative, and non-verbal behaviour of the participants on the provider side, enriching the effectiveness of the conversation.

We have designed a questionnaire to evaluate attitude among participants. We have found significant differences between the experimental conditions and we can confirm that a positive impact has been achieved in all participants.

Finally, we cannot assure that the interactive element, the proper hands of the participants and a whiteboard with a whiteboard marker to take notes, significantly influences attention and spatial and social presence.

As a result of this work, we have made publicly available a SEAW-dataset of 3 video sources (stereoscopic raw format) designed and acquired specifically for the purposes of the experiment. During the recording we considered three genres and both actor and observer acquisition perspective in the same context, international experiences, working or studying in a foreign country. Additionally, the questionnaires and the associated rates obtained from a diverse and balanced sample of 54 participants were provided.



# Chapter 5

## Interactive Communication Assessment

### 5.1 Introduction

The work presented in this chapter is focused on communications, applicable to different XR technologies and use cases. The main challenge we face is to design a test to evaluate socioemotional and technical aspects on a scenario that allows interactive communication between participants, without the limitation of visualizing simulated conversations or following scripts. Additionally, the experiment is designed seeking ecological validity and reproducibility, so that it can be applied with different participants, contexts, and even communication prototypes.

With this purpose, we have conducted an experiment based on a decision-making technique used to debate and find solutions as a team. As baseline we use the in-person condition and we compare it with hybrid meeting conditions through the reference prototype presented in Figure 1.2. In this chapter, we provide preliminary results of aggregate quality, social, and spatial presence, evaluated with questionnaires. These questionnaires have previously been validated in assessments where XR communications were simulated (360-degree videos recorded for that purposes) [54], used as pilot study. We also provide preliminary results on other socioemotional aspects, evaluated with questionnaires from the literature [151], in relation to the connection with the rest of the participants, based on the need to belong to a group [30]. Since with this experiment we wanted to explore the usefulness of technology in communications, free-form feedback about the experience was collected from the participants and the main findings are provided and analyzed in this chapter.

Section 5.2 presents the works found in the literature in relation to interactive communications assessments and validated methodologies used for the evaluation of socioemotional aspects. Section 5.3 describes the main parts of the experiment design: research questions, experimental conditions, stimuli, experimental setup, methodology, equipment, test session, and participants. Section 5.4 describes the main results, which are discussed in Section 5.5. Finally, Section 5.6 presents general conclusions that are relevant input for the decisions in the following steps of the research.

## 5.2 Related work

The experiment described in this chapter is the logical next step after the assessments with simulated conversations that have been performed so far. This thesis examined the use of a particular 360-degree video immersive communication prototype, presented in Figure 1.2. The purpose of this experiment is to assess the feasibility of the prototype in a realistic setting and to identify the key features that are more critical to the satisfaction of end users. Additionally, interactive experiments allow to observe how groups of participants interact with each other using the technology, increasing the need of the evaluation of socioemotional aspects. Due to this, the literature reviewed for the design of this assessment is classified into two categories: interactive experiments and studies that apply methodologies based on questionnaires of socioemotional aspects.

Compared to experiments where participants visualize pre-recorded content, interactive communication results in a less controlled environment. How to plan the experiment such that the sessions, with different participants interacting, can be replicable is one of the main challenges tackled during the design. In the literature, there are interactive experiments intended to test specific applications developed to solve problems faced by participants with specific characteristics, such as health-care [152, 153]. There are many works focused on evaluating collaborative tasks for training and learning in immersive environments [154–156]. Most of these experiments are based on carrying out specific tasks for which communication is one more tool but not the main objective of the study [157, 158]. In some cases, the task is performed by a single individual and there is an instructor who guide the steps to solve it [159]. This role of instructor and student can also be switched, as well as the condition of local and remote [49]. Other studies are based on activities that people are used to, such as photo sharing experience or walking in immersive technologies [160, 161]. In the study conducted by Lawrence et al. [162], the authors evaluated the effectiveness of their proposed communication system by utilizing a selection of questions designed to facilitate conversations with strangers. The limitation we found is that the communication has a controlled part, which is asking the question, and an uncontrolled part, which is answering it, guiding the communication but limiting spontaneity among the participants [163].

Despite the differences between the assessments found in the literature and the one presented in this chapter, the design is based on their contributions. We propose an experiment based on communication tasks, the Six Thinking Hats technique [164], in which there is a role of moderator and a role of speaker, and in which these roles are exchanged, as well as remote and local conditions. The motivation for this selection is that the topic and timing are controlled but the flow of communication totally depends on the participants.

In relation to the questionnaires to evaluate socioemotional aspects, we decided to ask for social and spatial presence with the same items than in our previous experiments [54, 58]. The objective is to continue validating the questionnaires used, this time, in an interactive environment. Furthermore, in accordance with the literature review, an examination of socioemotional aspects related to the connection and closeness between participants was conducted [162]. Specifically, following the study by Goldstein et al. [151], we asked about the positive attitude (*Liking*), the perceived connection (*Self-other overlap*), including the Inclusion of Other in the Self Scale (IOS) scale [165], and the *perceived empathy* in relation to a reference participant. Additionally, emotions are intensively analyzed in the literature related with interactive communications since it has a great impact on conversations and the creation of affective bonds [166]. Then, *emotions* were evaluated following the Internationally Reliable Short-Form Positive and Negative Affect Schedule (I-PANAS-SF) [167].

Given the complexity of interactive experiments, methodologies based on questionnaires are usually accompanied by semi-structured interviews or free-form feedback sections [154, 168]. Semi-structured interviews usually ask about certain aspects, such as Volonte et al. [166] with the question “Did you feel very excited and willing to interact with the Virtual Humans?” in an experiment about the effect of interacting with emotional virtual humans on users’ behaviors. However, free-form feedback consists of a blank space where participants share comments about the experience, removing any bias caused by the question. Both are used in the analysis of results to try to understand phenomena that were not expected or to take into account in future experiments certain aspects that the researchers in charge had not considered. In this experiment, the free-form feedback is asked at the end of the questionnaire.

## 5.3 Experiment design

### 5.3.1 Research questions

Based on the previous analysis, we pose the following RQs:

- RQ1: Can we apply methodologies used in simulated communications assessments to evaluate the aggregate quality in interactive communications assessment?
- RQ2: Can we apply methodologies used in simulated communications assessments to evaluate presence in interactive communications assessment?
- RQ3: How to design a controlled experiment while maintaining a natural flow of the communication?

- RQ4: Can the roles (e.g., moderator or speaker) of a discussion or meeting influence on the hybrid meeting experience?
- RQ5: What socioemotional aspects are involved in interactive communications?
- RQ6: What elements of the real-time 360-degree video communication prototype are relevant for end-users?

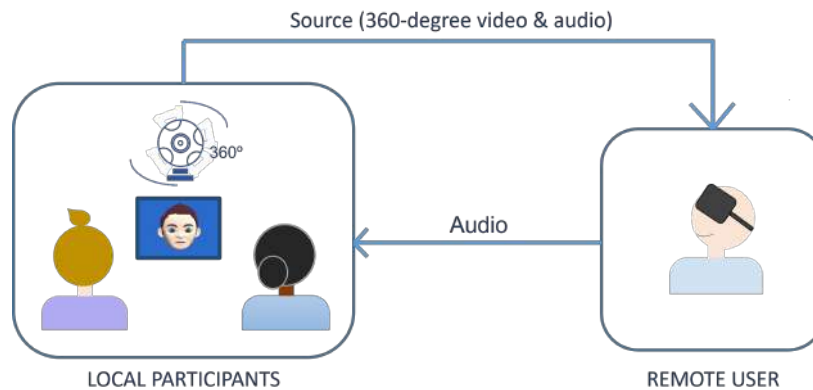


FIGURE 5.1: Prototype of 360-degree video communication considered in this experiment. The remote user visualizes the 360-degree of the local side through the HMD. Local participants visualize the synchronized head movements of the remote user in a cartoon avatar on a tablet. Additionally, the lips of the avatar move with the audio.

The prototype under evaluation, based on the reference configuration presented in Figure 1.2, is presented in Figure 5.1. Based on the finding in Chapter 4 that the possibility for remote participants to visualize their hands appears to be useful when participants have a specific tool to use them, this additional tool is not considered in this scenario. The prototype allows a real communication between local and remote participants, not pre-recorded 360-degree videos. We should take into account that the remote user visualizes the local participants through the HMD, but local participants, following the basic configuration, do not have any information about the remote user. In the previous chapter, we found significant differences in social presence depending on how actors and actresses interact and talk to the camera in the recordings. Then, to make more natural the fact that local participants speak to a camera, a tablet with a cartoon avatar is added to the local side to represent the head direction and the articulation of the mouth of the remote user. Although there are other types of more realistic avatars that can influence the experience of local participants, we decide to use this one for its simplicity since the analysis of the avatar type is not an objective in this test [169]. The cartoon avatar is modified to represent the appearance of the remote participant to enhance the sense of social presence and realism during the interaction.

### 5.3.2 Equipment

The prototype under evaluation, The Owl [170], is composed by a capture element (owllcam) that uses a Ricoh Theta V 360-degree camera, a processing and control unit based on a Raspberry Pi with a touchscreen, and a standard hands-free speaker Jabra. The video is captured and compressed in equirectangular projection from the camera and sent to remote users at 12 Mbps with 4K resolution. Network conditions are controlled so that quality changes do not occur during the session that may affect the QoE, as observed in Chapter 4. The Raspberry Pi unit also includes the communications backend, implementing the signaling for control sessions, and the data plane to send the streams to all the connected users. Audio conference uses open software (Mumble).

Remote participants use an application developed in Unity3D, in charge of decoding and rendering the 360-degree video, to attend the immersive conversation through the Oculus Quest 2 with 3 DoF. The head-movement tracking of the remote participants is captured and transmitted to the local side in real-time.

### 5.3.3 Experimental conditions

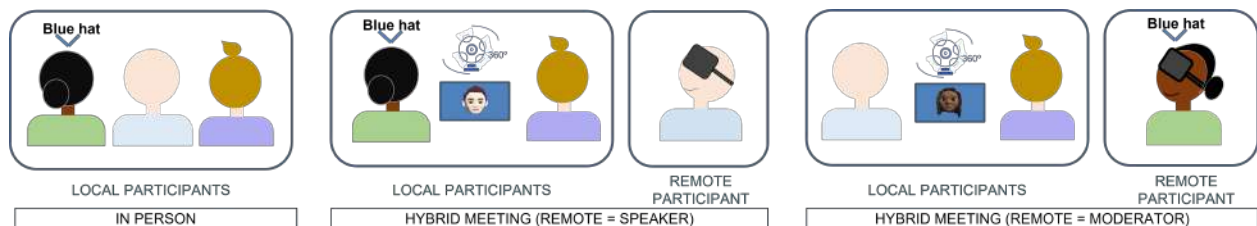


FIGURE 5.2: Interactive communication scenarios considered in the experiment based on the experimental conditions: in-person and mixed technology mediated discussion between three participants.

As presented in Figure 5.2, two experimental conditions are considered: an in-person and a hybrid meeting. In-person discussion is carried out between three participants around a table in the same local room. It is used as the baseline condition in our experiment because it is the one to which the participants are most used to [171]. Hybrid meeting is carried out between two local participants and one remote participant. Participants assigned to the remote condition attend the discussion remotely through the HMD and all of them visualize the 360-degree video with the same constant quality.

Participants sign up for the test in groups of three people and these groups are maintained in each test session. The reason behind this way of collecting participants is to ensure that the group members have a previous relationship with each other, usually university friends, trying to avoid differences in the interaction due to the different levels of relationship between the groups considered

in the experiment. Each group of three participants tests in-person and hybrid meetings but in a randomized order for each one.

### 5.3.4 Stimuli

The stimulus of this experiment is the interactive immersive communication itself. During communication, naturalness must be guaranteed. However, from the assessment perspective, the session structure must be maintained to compare between test sessions and even between different XR technologies. Then, the communication is based on a tool for everyday problem solving, the Six Thinking Hats technique [164]. It allows participants to evaluate a task from six different points of view. In this way, each perspective is deeply analyzed and explored to find a solution for the task. Each different point of view is identified with a color hat. The motivation behind this is that it allows an interactive communication while timings are guided and all participants explore the same points of view in the same order for each problem. The color of the hats are: blue, green, red, yellow, black, and white, explained in detail below.

- Blue Hat is in charge of controlling the decision-making process, gathering data, summarizing it, and drawing conclusions.
- Green Hat represents the innovation, time of thinking about innovative ideas.
- Red Hat is focused on following feelings and supporting the opinions on those feelings.
- Yellow Hat represents optimism and emphasizes the advantages and positive aspects of the ideas.
- Black Hat evaluates the dangers associated with each suggestion from a critical viewpoint.
- White Hat identifies objective data needed to make the decision in a successful way.

The blue hat, considered the moderator role, is assigned to one of the three participants, maintaining it along the whole test session. We then decide the participant with the moderator role to attend the discussion both as local and remote participant in the hybrid meeting and, obviously, as local participant in-person condition, as shown in Figure 5.2. It results on three tasks to solve for each test session. The reason of this decision is to consider the influence of the role in the experience. The other two participants discuss the ideas based on a different hat (the same hat is assigned to both participants at the same time) every two minutes.

The context of the tasks is tele-education. The participants of the experiments are students at the Universidad Politécnica de Madrid who have experienced online learning due to COVID-19. Thus,

the selected context, tele-education, is familiar for them, increasing the fluency of the conversation and discussion. The tasks are: A) Evaluation of the tele-education methods, B) Finding new ideas for tele-education, and C) Finding new ideas to conduct exams in tele-education. The tasks proposed for in-person and hybrid meetings are randomized to avoid bias in the conclusions. The blue hat opens each task by presenting it and launching some ideas to discuss and closes the task by summarizing the ideas that have been discussed, which is the final phase that leads the participants to find a solution for the assigned task. The order of the points of view/hats analyzed is selected following the recommendations of the literature [164]. For the statement A the order of the hats are: blue, white, red, yellow, black, green, and blue. For the statements B and C, the order of the hats are: blue, white, red, green, yellow, black, and blue.

### 5.3.5 Experiment setup

In this subsection, we present an overview about the experiment setup: the material that local participants need to carry out the session and the logistic of the experiment. From a high level perspective, this information is presented to help the reader understand the experiment.



FIGURE 5.3: Remote participant attending the hybrid meeting through the HMD. Local participants visualizing the remote participant at the tablet (cartoon avatar) and following the indications of the cards on the table.

**Local participants.** In the in-person meeting condition, the three participants are in the same room sitting around a round table. In this experimental condition, the researcher responsible for the study is not present in the testing room, in order to mitigate any potential inhibitions that participants may have regarding their speech and allowing them to speak freely. In the hybrid meeting condition, two local participants are in this same room, presented in Figure 5.3. Local participants are seated around the camera forming a triangle with the Owl. The capture element of the Owl is located at an average eye level height approximately, in order to create the perception for remote participants that they are seated at the same table as the local participants [135]. As can be

observed in Figure 5.3, the tablet with the avatar of the remote participant is located immediately below the camera. In this way, it is natural to look at in that direction for local participants and it can enhance the social presence perceived by remote participants.

Figure 5.4 presents the material used by local participants to guide the experiment. Cards with the colors of the hats are utilized in the specific order of colors determined by the task, provided in an additional card. Also, a summary of the point of view associated with each color of the hat is provided to facilitate participation. A timer is used to control the duration of the intervals between hat changes (2 min). If the participant with the blue hat is in the local condition, she/he is in charge of monitoring the time and introducing each of the hats. However, if she/he participates as a remote user, the timing and presentation of the hats have to be done by one of the two local participants.

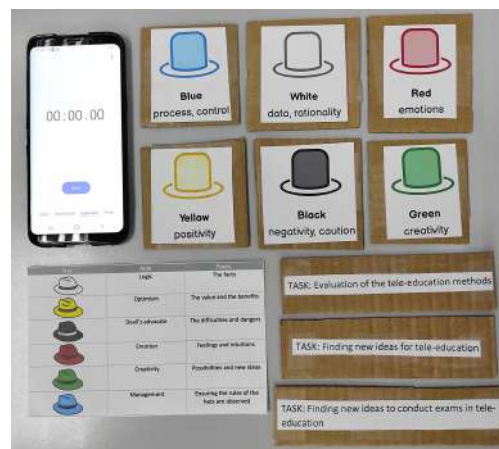


FIGURE 5.4: Material used to guide the test session: timer, cards of six colored hats, summary of the representation of the color of each hat, and tasks of the test session.

**Remote participant.** The remote participant is in another room with the responsible researcher in case they had any problems during the task. He/She is located in front of a table to identify the table of the local scene with physical one [54]. Also, he/she is able to spin around without any limitation while seated on a swivel chair.

### 5.3.6 Methodology

Here, the methodology considered in the experiment is presented. Table 5.1 shows a summary of the items evaluated in the experimental conditions.

**Personal information.** The following personal information is collected from the experiment participants to characterize them. Specifically, age, gender, vision (corrected or normal), and experience with VR (Basic: “I have used this technology less than 10 times”, Intermediate: “I have

TABLE 5.1: Overview of the questionnaire items asked depending on the participant condition and the assigned role.

| Participant Condition | Role               | Quality<br>(ACR) | Spatial Presence<br>(PP1-PP4) | Social Presence<br>(SP1-SP5) | Emotions<br>(I-PANAS-SF) | Liking, Self-other<br>overlap, Empathy |
|-----------------------|--------------------|------------------|-------------------------------|------------------------------|--------------------------|--|
| Remote                | Moderator/Blue hat | X                | X                             | X                            | X                        |  |
|                       | Speaker/All hats   | X                | X                             | X                            | X                        | X                                      |
| Local                 | Moderator/Blue hat | X                |                               | X                            | X                        |  |
|                       | Speaker/All hats   | X                |                               | X                            | X                        | X                                      |

used this technology between 10 and 30 times”, Expert: “I have used this technology more than 30 times”).

**Quality.** The aggregate quality is asked, following the literature [37], in the post-sequence questionnaire using the Absolute Category Rating (ACR) with the statement *Please, rate the quality of the experience*. It is rated on a five-grade quality scale [25], where the categories: “Bad”, “Poor”, “Fair”, “Good”, and “Excellent” are presented.

**Presence.** Spatial presence is collected from remote users. It is asked with four items adapted from the state of the art and validated in other experiments in a similar scenario, such as in Chapter 4. Specifically, spatial presence is measured with the following questions: *I felt I was present in the places shown in the video (PP1)*, *I felt surrounded by the actions in the video (PP2)*, *I felt I was sitting by the table at the place of the video (PP3)*, and *I felt I could have reached out and touched the items on the table of the video (PP4)*. The items were rated on a seven-point Likert scale (where 1 = “Strongly disagree” to 7 = “Strongly agree”).

The social presence is asked to remote and local participants in the hybrid-meetings. For remote participants, the social presence is asked relative to local participants they view through the HMD. For local participants, the social presence with respect to the remote user they view in the tablet is asked. As spatial presence, it is assessed with five items adapted from the literature and validated in other experiments in a similar scenario, such as in Chapter 4: *I felt that people were talking to me (SP1)*, *I felt that I was listening to the others in the video (SP2)*, *I felt I was present with the other people in the video (SP3)*, *I felt like the people in the video could see me (SP4)*, and *I felt I was actually interacting with other people (SP5)*.

**Emotions.** Emotions are asked with the I-PANAS-SF [167]. It is a 10-item version with 5-item PA and NA subscales rated on a 5-level scale (where 1 = “Never”, to 5 = “Always”). Specifically, *Thinking about yourself and how you normally feel, to what extent do you generally feel: upset, hostile, alert, ashamed, inspired, nervous, determined, attentive, afraid, and active*.

In this part of the questionnaire, the questions are related with the perception of the participant with the moderator role (blue hat), who is the reference in this assessment, based on [151]. For

that, these items are asked in-person and hybrid meetings but only to the participants with the speaker role.

**Liking [151].** The first item is to consider the degree of positive attitude that participants have. *How much do you like the experience?* Seven-level Likert scale (where 1 = “Not at all”, to 7 = “Very much”).

**Self-other overlap [151].** Eight questions about the similarity, bond, closeness, tie, link, close association, connection, and shared identity, each on a 7-level Likert scale (where 1 = “Not at all”, to 7 = “Very much”), are asked. Specifically, *To what extent do you feel you are similar to the participant with the blue hat?* (Similarity), *To what extent do you feel a bond with the participant with the blue hat?* (Bond), *To what extent do you feel you are close to the participant with the blue hat?* (Closeness), *To what extent do you feel you are tied to the participant with the blue hat?* (Tie), *To what extent do you feel you are linked to the participant with the blue hat?* (Link), *To what extent do you feel you are associated with the participant with the blue hat?* (Association), *To what extent do you feel you are connected with the participant with the blue hat?* (Connection), *To what extent do you feel you share identity with the participant with the blue hat?* (Shared identity). Last, participants chose a pair of circles from seven with different degrees of overlap (where 1 = “No overlap”, to 7 = “Most overlap”) to describe their relationship with the participant with the blue hat, following the IOS [165].

**Perceived empathy [151].** At the end of the task, the participant with the moderator (blue hat) role, summarizes the ideas to conclude the task. Then, the perceived empathy by the other two participants is asked with *To what extent do you think the participant with the blue hat empathized with you when summarizing about your ideas?* and rated on the same 7-level Likert scale (where 1 = “Not at all”, to 7 = “Very much”) than previous items.

### 5.3.7 Test session

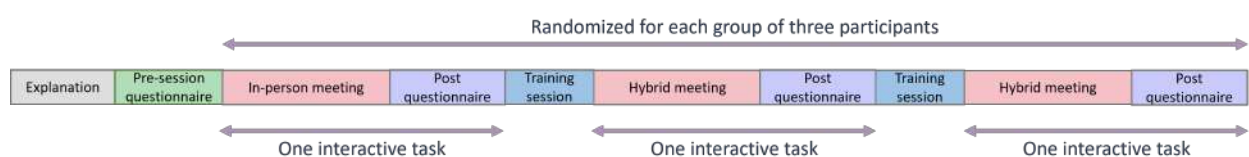


FIGURE 5.5: Test session structure.

The structure of the test session is presented in Figure 5.5. The participants receive an information sheet before the day of the assessment that explains the experiment in detail: the technique of the six hats, experimental conditions, and the tasks to be discussed. This procedure is intended to

allow participants to come to the experiment with ideas for each of the tasks, increasing the flow of communication.

At the beginning of the experiment, they sign a consent form to treat their data in accordance with the GDPR of the European Union. The assessment starts with the researcher in charge presenting to the participants an overview of the assessment, called the “Explanation” part. Then, they fill the pre-session questionnaire: a personal information survey to collect relevant data to characterize the participants. The training session consists of an informal interaction to test the hybrid condition and resolve any doubts about the system. It is carried out immediately before each hybrid condition. For each of the three phases of the test session, a task (A, B, or C) is assigned. Each task is about 15 minutes, resulting in a test session around one hour. After each experimental condition, they are requested to answer the post-session questionnaires to evaluate the socioemotional features of interest.

### 5.3.8 Participants

A total of 27 participants (10 females, 17 males) with an age range between 20 and 36 ( $M=23.6$ ,  $SD=3.8$ ) carried out this experiment. The technology from the remote side was tested by 19 participants. 89.5% of the sample of the participants reported a basic level of experience in VR and the rest reported an intermediate level.

## 5.4 Experimental results

This experiment intends to propose methodologies closer to evaluations in realistic scenarios that increase ecological validity. Given the novelty of this experiment, its complexity, and the sample of the participants, the procedure is treated as a forerunner study to propose a methodology in a real scenario that can be carried out on a massive scale. Nonetheless, it is a meaningful contribution because this prototype has not been evaluated before in a structured way. The experimental results section is divided into: quantitative and qualitative analysis. The discussion in detail of the results presented in this section is shown in Section 5.5.

### 5.4.1 Quantitative results

The analysis of the results presented here is a quantitative analysis. It aims to obtain some inputs to guide next steps of the research taking into account that, as already mentioned, the sample of participants is not enough to obtain robust conclusions. Figure 5.6 presents the mean scores

and the associated 95% CIs of the quality obtained from local and remote participants in tasks A, B, and C carried out in hybrid and in-person meetings. The MOS obtained for task C is similar for in-person and hybrid meetings. However, for tasks A and B the difference between the MOS obtained in in-person and hybrid meetings is higher.

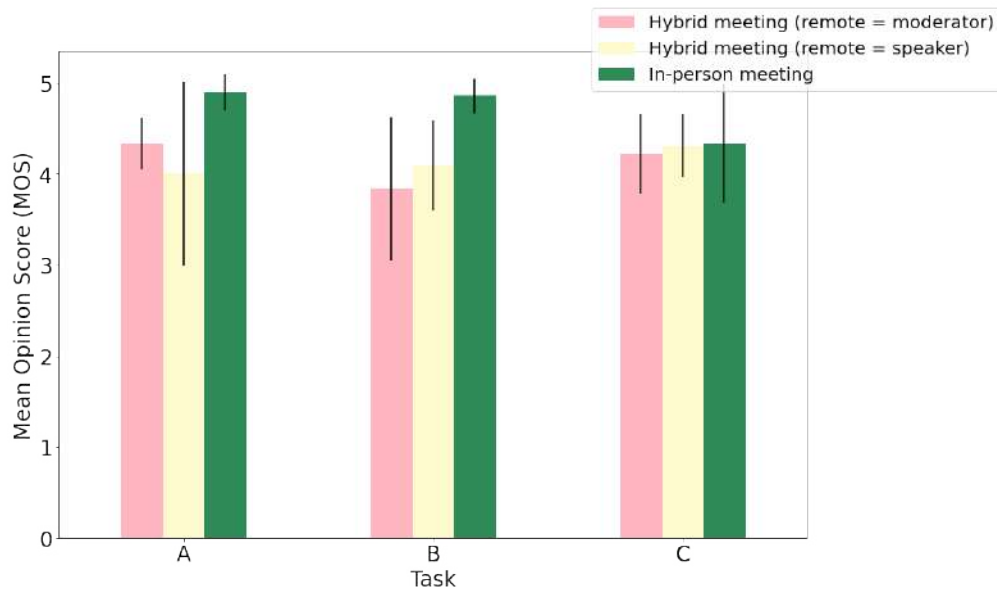


FIGURE 5.6: Mean opinion scores (y-axis) on a five-level scale obtained from 27 participants in each of the tasks (x-axis) and presented by experimental conditions.

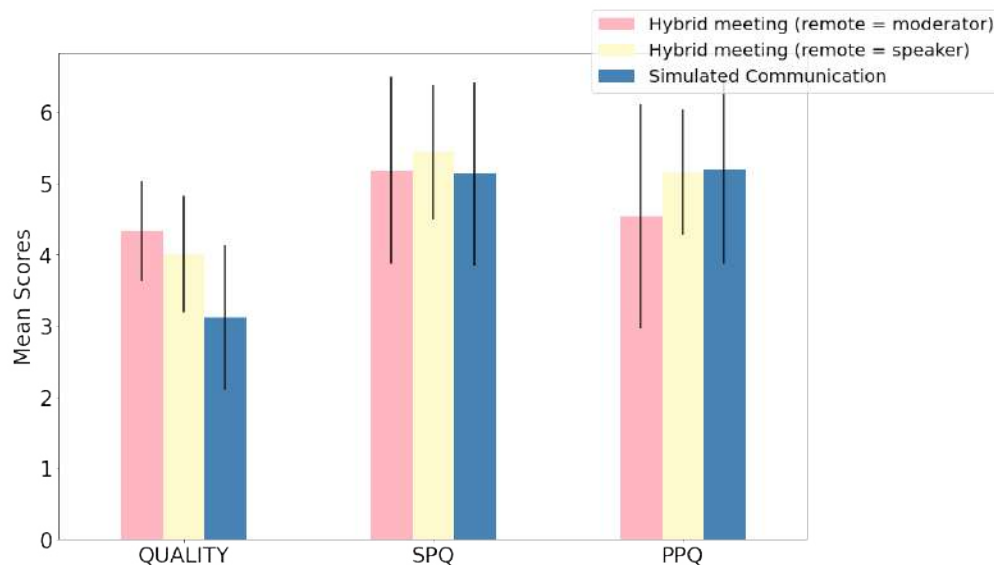


FIGURE 5.7: Mean scores (y-axis) of quality, spatial, and social presence (x-axis) collected in this experiment (Hybrid meetings) and collected in the previous test presented in Chapter 4. All ratings were provided by remote participants. Note that quality was rated on a 5-level scale, while social and spatial presence were rated on a 7-level scale. The number of participants of this experiment is not the same of the previous one.

Figure 5.7 presents the mean scores and the associated 95% CIs of the quality, Spatial Presence Questionnaire (PPQ), and Social Presence Questionnaire (SPQ) provided from the remote users in this experiment (hybrid meeting bars), and in the experiment presented in Chapter 4, in which the communication was simulated (pre-recorded videos). Note that quality is rated on a 5-level

scale, while social and spatial presence are rated on a 7-level scale. In terms of spatial and social presence, the mean scores obtained are quite similar. However, in terms of quality, the simulated communication is rated lower.

TABLE 5.2: Summary of the results obtained from remote participants in the items: aggregate quality and social and spatial presence.

| Questionnaire items                    | Hybrid meeting<br>(remote = moderator) | Hybrid meeting<br>(remote = speaker) | Simulated Communication         | Significance                                  |
|--|--|--------------------------------------|---------------------------------|---|
| Aggregate quality (5-level scale)      | $M = 4.333$<br>( $SD = .707$ )         | $M = 3.889$<br>( $SD = .782$ )       | $M = 3.111$<br>( $SD = 1.022$ ) | $F_{2,69} = 7.730, p < .001, \eta_p^2 = .183$ |
| Spatial Presence (PPQ) (7-level scale) | $M = 4.528$<br>( $SD = 1.568$ )        | $M = 5.056$<br>( $SD = .882$ )       | $M = 5.185$<br>( $SD = 1.318$ ) | $F_{2,69} = 1.027, p = .363, \eta_p^2 = .029$ |
| Social Presence (SPQ) (7-level scale)  | $M = 5.178$<br>( $SD = 1.317$ )        | $M = 5.378$<br>( $SD = .977$ )       | $M = 5.133$<br>( $SD = 1.287$ ) | $F_{2,69} = .054, p = .947, \eta_p^2 = .002$  |

Table 5.2 shows the summary of the ratings collected for aggregate quality, SPQ, and PPQ and the significance between them. Pearson & D’Agostino normality test is computed to validate the normal distribution of each questionnaire item. The level of significance considered in the analysis is 0.5. Due to the normality condition, the 2-way ANOVA is applied to examine significant differences between simulated and interactive (hybrid meeting) communications. Then, Tukey HSD test is performed to better explore the differences. It shows a significant difference between the aggregate quality collected in the hybrid meeting (remote = moderator) and the simulated communication ( $p = .002$ ). In order to fairly compare the scores collected in the two different experiments (simulated and interactive communications), spatial and social presence data is transformed as in Chapter 4. Specifically,  $\arcsin(\sqrt{P/7})$ , where  $P$  is the presence rating and it is divided by seven because social and spatial presence are evaluated on a seven-level scale. Once the non-normality is corrected, it is analyzed applying the 2-way ANOVA test without finding significant differences.

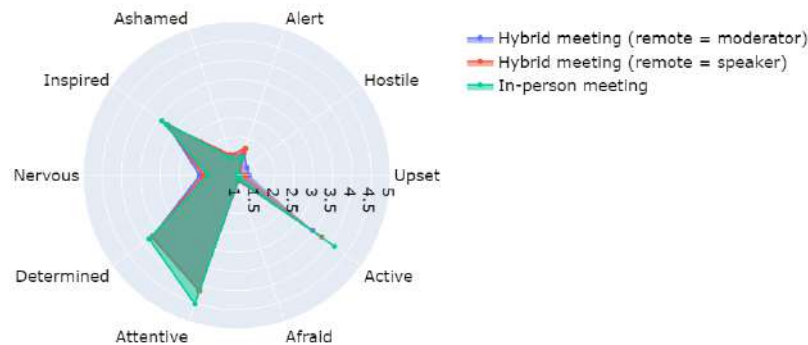


FIGURE 5.8: Reported emotions (I-PANAS-SF) divided into the experimental conditions: in-person and hybrid meetings.

Figure 5.8 presents the mean scores of the I-PANAS-SF items provided from the participants in the experimental conditions (in-person and hybrid meetings). It shows that the means of the items of the I-PANAS-SF are similar for in-person and hybrid meetings.

TABLE 5.3: Summary of the aggregate quality and social presence evaluated by local participants in the in-person and hybrid meetings. The results presented are independent of the role of the participants.

| Questionnaire items                  | In-person meeting        | Hybrid meeting<br>(remote = speaker) | Hybrid meeting<br>(remote = moderator) | Significance             |
|--------------------------------------|--------------------------|--------------------------------------|--|--------------------------|
| Aggregate Quality<br>(5-level scale) | M = 4.815<br>(SD = .395) | M = 4.250<br>(SD = .851)             | M = 4.111<br>(SD = .676)               | $X^2 = 13.054, p = .001$ |
| Social Presence (7-level scale)      | -                        | M = 5.463<br>(SD = .938)             | M = 5.078<br>(SD = 1.158)              | $X^2 = 1.412, p = .235$  |

Table 5.3 presents the ratings obtained from local participants in the hybrid or in-person meetings about the aggregate quality and social presence. Due the non-normality of the aggregate quality and the different number of collected data, Kruskal Wallis and Mann-Whitney U rank test as post-hoc analysis are conducted. We find significant differences between in person and hybrid (remote = moderator) meeting ( $U = 105, p = 0$ ) and between in person and hybrid (remote = speaker) meeting ( $U = 170, p = .027$ ). The social presence collected from local participants is evaluated in relation to the cartoon avatar presented in the tablet. In this table there is no distinction between the role of the participants during the test session. The same procedure as with aggregate quality is applied on social presence data and not significant differences are found between hybrid meetings.

TABLE 5.4: Summary of the results obtained from participants with the role of speaker in the questionnaires about: liking, self-other overlap, and perceived empathy. These items were asked in relation to the moderator role (blue hat).

| Questionnaire items                | Hybrid meeting<br>(remote = moderator) | Hybrid meeting<br>(remote = speaker) | In-person meeting         | Significance                                  |
|------------------------------------|--|--------------------------------------|---------------------------|---|
| Liking (7-level scale)             | M = 4.944<br>(SD = 1.589)              | M = 5.750<br>(SD = 1.118)            | M = 6.167<br>(SD = .786)  | $X^2 = 3.640, p = .162$                       |
| Self-Other Overlap (7-level scale) | M = 4.333<br>(SD = 1.534)              | M = 4.900<br>(SD = 1.252)            | M = 5.611<br>(SD = 1.243) | $F_{2,51} = 4.279, p = .019, \eta_p^2 = .144$ |
| Perceived Empathy (7-level scale)  | M = 5.556<br>(SD = 1.756)              | M = 5.800<br>(SD = 1.281)            | M = 6.333<br>(SD = .907)  | $X^2 = 1.592, p = .45$                        |

Table 5.4 shows the summary of the items evaluated by participants that discussed the ideas of each task (speaker role) in relation to the participant with the moderator role (blue hat). Due to the non-normality of the liking and perceived empathy data, Friedmann test was conducted. It shows that there are not significant differences between conditions in terms of liking and perceived empathy. Due to the normality of the self-other overlap data, ANOVA and Student paired-samples t-test with Bonferroni correction are applied to find significant differences between in-person and hybrid meetings. In fact, we find significant differences between in-person and hybrid (remote = moderator) meetings ( $t_2 = -2.945, p = .027$ ).

## 5.4.2 Qualitative results

The analysis of the results presented here is a qualitative analysis. Analyzing the feedback obtained in the free-form feedback gives us insights about what is important to users and which aspects of the prototype work better, orienting the following steps of the research. The analysis has been divided based on the topic of each comment. Starting with the technical comments up to those related to the socio-emotional aspects, they are listed below.

### **Video**

In general, participants had a good experience in the meeting but some of them found that the video quality was poor to feel it as a real experience and, therefore, the sense of being there. Also, they found difficulty with the superposition of real objects in the virtual environment (e.g., handheld controllers), which hindered complete immersion.

### **Audio**

Participants pointed out that most of the time they were able to hear and be heard well, increasing the fluency of the hybrid meeting. However, at some points, the sound was distorted, distracting them for the conversation. Additionally, they noticed a small delay in audio and other remote participants were interrupted by local participants. In relation to interruptions, the communication was less fluent and effective with the remote participant, specifically if she/he had the moderator role. Specifically, it was difficult to introduce the remote one in the conversation, making him/her start talking. They noticed similar interruptions as on web-platforms video calls.

### **Avatar**

Some participants focused their comments about using an avatar in the hybrid meeting, finding it distracting and not portraying the presence of the participant attending the meeting through the HMD, specifically due to the lack of nonverbal signals, such as hand gestures. They suggested that the experience would have been better if the avatar was more similar to the participant and had more realistic movements and expressions. Overall, users felt that the avatar was not as effective as being able to see the participant in person or through a live video feed. However, some participants pointed out that the hybrid meeting with an avatar looking at them when they spoke was better than just a voice call.

### **Comfort**

Some participants experienced some dizziness, even a feeling of being overwhelmed, although they got used to it throughout the task. They noticed that glasses were heavy and not comfortable at all.

### **Additional tools**

Participants found limitations of not being able to take notes with pen and paper, affecting their experience. Specifically, one participant commented: *this technology may be suitable for discussion or follow-up meetings, but not for note-taking meetings (from the remote side).*

### **Six thinking hats technique and tasks**

Participants found entertaining to debate current topics, such as tele-education, but also found that the tasks were repetitive and monotonous, thus limiting the flow of conversation as there were not many new ideas to discuss. However, they enjoyed the six hats technique to categorize different viewpoints and address them in an interesting way.

### **Questionnaires**

Some participants found that the questions during the meeting were quite similar, tedious to understand for someone not familiar with psychological surveys.

### **Differences depending on the role of the remote participant**

Some participants highlighted some feelings about the role of the remote users in the hybrid meeting. They believed that the remote participant's role in the meeting is decisive for good communication and that a proactive role benefits the communication and reduces the possible disconnection related to technology. Generally, participants accomplished a better experience with the moderator in person and with a colleague remotely. One reason given for that was that the moderator did not know when to intervene, so they perceived the experience of interacting with the moderator remotely as peculiar and not particularly engaging. From the point of view of a remote moderator, some participants were comfortable with it but felt that it was less natural and they lost control comparing to moderating in person. One of them concluded: *Even though I nod a lot and it can be seen through my cartoon avatar, I prefer to be in person.*

Although it was a minority, there were contrary opinions that considered that the remote participant with the role of moderator was the best option because in that way the remote participant was not ignored during the conversation.

### **Comments from remote participants**

Remote participants shared their thoughts about feeling present but not getting the same attention from local participants. They considered it was an interesting and interactive way of using XR, but it was hard to remember what happened in the conversation. They found it to be an improvement over regular calls but still need further improvement.

### **Comments from local participants**

Some of the local participants indicated that they felt the remote participant was present, enhancing

the fluency of the conversation. However, they were not able to offer him/her the same attention as their local colleague. This appreciation was noticed from both sides, local and remote participants.

## 5.5 Discussion

This section discuss in detail the quantitative and qualitative analysis presented before. Specifically, it is structured following the research questions posed during the design of the experiment and their associated hypotheses investigated.

Regarding RQ1, we explore the hypothesis (**H1**): aggregate quality ratings can be affected comparing assessments based on simulated or interactive communications. In terms of aggregate quality, the analysis shows that interactive experiments improve it, which seems reasonable. Figure 5.7 shows that the average of the aggregate quality decreases in the simulated communication. As presented in Table 5.2, there is a significant difference on the aggregate quality between conditions. Specifically, the hybrid condition in which the remote participant is assigned to the moderator role and a simulated communication. In this table the data is collected from the evaluations of the remote participants. This difference is discussed, also relying on the qualitative analysis in RQ4. However, this analysis supports the need to perform interactive experiments and increase the ecological validity. Otherwise, aspects whose evaluations may be different in a real environment would be analyzed, obtaining confusing results and conclusions.

Regarding RQ2, we explore the hypothesis (**H2**): spatial and social presence ratings can be affected comparing assessments based on simulated or interactive communications. According to the results of Figure 5.7, it does not seem that there are differences between the social and spatial presence. It is corroborated with the statistical analysis presented in Table 5.2, which shows that there are not significant differences between interactive and simulated communications in terms of spatial and social presence. It should be remembered that in the simulated communication, during the recording of the videos, the actors/actresses interacted with the camera, pretending that there was a remote user. As a positive findings, we highlight that there is consistency on the questionnaires used to evaluate social and spatial presence in similar experiments and the mean scores are quite high.

Regarding RQ3, we explore the hypothesis (**H3**): if the collected data on H1 and H2 is consistent with the one obtained in the previous experiment, the one presented in Chapter 4, the interactive assessment allows the natural flow of a conversation and the assessment of socioemotional aspects. Following the analysis of H1 and H2, we can assure that there is a consistency in the collected data in terms of aggregate quality and spatial and social presence. Then, a technique as six thinking

hats can work for interactive assessments. Additionally, the free-form feedback gives us optimistic insights because this assessment works in terms of QoE. Despite the problems pointed out by some participants, the experience was positive for most of them.

Regarding RQ4, we explore the hypothesis (**H4**): the different role and the interaction as local and remote participant can affect the experience.

Table 5.2 has already been analyzed in terms of aggregate quality and presence evaluated by remote participants. Table 5.3 summarizes the aggregate quality and social presence between in-person and hybrid meetings from the point of view of the local participants. As expected, there are significant differences in terms of aggregate quality between in person and hybrid meetings in which the average of the ratings are quite similar. In average, these ratings are slightly higher than the obtained from remote participants. In terms of social presence, there are not significant differences between conditions from local or remote perspective. The social presence rated from remote participants (table 5.2) is similar from the rated by local participants related to the social presence perceived through the cartoon avatar on the tablet (table 5.3).

Regarding the role assigned during the experiment: moderator or speaker, it was initially chosen randomly. However, during the experiments and by talking with the participants, we observed that certain people might not feel comfortable in their role. The obtained feedback may have influenced by this so, we propose to randomize the choice of the role of moderator by the users and by the researcher in charge of the experiment to avoid its influence on the analysis.

Table 5.4 shows that the liking, self/other overlap, and perceived empathy ratings in the hybrid meeting condition is higher than in in-person meeting from the point of view of the participants with the speaker role. On average, the results of the hybrid meeting in which the remote participant is an speaker differ from those found in the in-person meeting. However, a higher decrease in the feeling of perceived closeness with the remote moderator is observed, along with an increased dispersion in the results, on hybrid meetings in which the remote participant is the moderator. These findings may be attributed to participants' feedback indicating insecurity and a lack of clarity on how to effectively act as a remote moderator, as well as the absence of tools to facilitate note-taking and summarizing the ideas.

We assume that, since it is an interactive experiment, there are highly evaluated socioemotional features in the literature that need to be assessed. Then, regarding RQ5, we explore the hypothesis (**H5**): the interactivity of the experiment can involve other socioemotional aspects, such as emotions.

Figure 5.8 does not present relevant differences between the experimental conditions in terms of emotions, which makes sense since the tasks were discussion tasks in the context of tele-education

in assertive mode. However, there are some categories of emotions that are more stimulated than others. Specifically, inspired, active, determined, and attentive.

Regarding RQ6, we explore the hypothesis (**H6**): the elements highlighted in the free-form feedback are elements that participants consider relevant for their experience.

Analyzing the comments, we structured the topic in categories that refers to the most relevant elements noticed by participants. Specifically: presence, the quality of the video and audio, the avatar, comfort, additional tools for the technology, and the six thinking hats technique. Presence has already been discussed in RQ2, but the rest are discussed below.

### **Video and audio**

Evaluation of the video has revealed that the quality offered by current 360-degree cameras and display devices is still lower than the expected one, particularly compared with the provided by traditional screens. Furthermore, audio has emerged as a crucial aspect in interactive environments. This finding is a strong motivation to continue conducting experiments in interactive environments, as audio had not previously been a significant topic of feedback from participants. Additionally, interruptions in tele-meetings remain a persistent problem that has yet to be effectively addressed.

### **Avatar**

Despite comments from some participants about how unrealistic the avatar is, the ratings of the social presence of the local participants relative to the remote participant are not so negative, presented in Table 5.3. On the other hand, in this experiment the cartoon avatar is a solution to improve interaction but it was not a goal to evaluate it in detail. This result is in line with the literature that states that more realistic avatars improve interaction [160], but it is still an open question that needs to be addressed.

According to feedback obtained from the participants, it has been suggested that an intermediate approach between utilizing a cartoon avatar and a photorealistic avatar would be to incorporate the inclusion of hands, thereby enabling the avatar to gesture in a more natural manner. Despite these considerations, the integration of a cartoon avatar into the system is an improvement in comparison to not having feedback from the remote participant.

### **Comfort**

There were few participants who reported dizziness or discomfort. Although it is a problem, the devices are continuously improving.

### **Additional tools**

The limitation of not being able to take notes is something that could be solved by proposing a solution similar to the experiment in Chapter 4 in which the participants took notes on a whiteboard.

This result validates what we observed in the experiment in Chapter 4 in which the visualization of hands only affected the experience if there was a specific task to be done with them. Otherwise, it is observed that the participants with the role of speaker have not commented anything in this regard.

### **Six thinking hats technique and tasks**

Although the topic of tele-education could change between tasks, it was not done to avoid topics with which the participants were more familiar than others. However, the tasks could be modified for next experiments.

The six hats technique has demonstrated effectiveness in facilitating communication and preserving the natural flow of the conversation. Given its success, the use of other techniques from the field of psychology to increase the ecological validity of the studies can be considered.

### **Questionnaires**

The feedback received regarding the questionnaires in this study aligns with previous observations. It motivates the argument that the questionnaires must be clear, direct, and concise. It is the responsibility of the researcher to carefully consider the design of the questionnaires in order to minimize potential misunderstandings.

Furthermore, in experiments of such complexity, questionnaires do not provide information about aspects that are not specifically asked. Even those that are asked cannot be guaranteed that the participants have understood the question. However, valuable insights can be obtained from participant free-form feedback, which can provide a better understanding of the strengths and weaknesses of the experiment, as well as perspectives that may not have been considered during the design phase.

## **5.6 Conclusions**

This work presents an experiment design for interactive immersive communications that is replicated while maintaining the naturalness of the communication. Due to the complexity of the experiment and the number of participants that have been considered, the results can be used as a pilot studies for future interactive experiments.

The results obtained in aggregate quality, spatial, and social presence have been compared with those obtained in controlled experiments where communication was simulated. In terms of aggregate quality, the interactivity of the session, as expected, increases the mean scores. However, the mean scores of social and spatial presence remain similar. This is a relevant contribution since it

shows the consistency of the methodologies applied. Therefore, these preliminary results validate the proposed methodology to jointly assess technical and socioemotional aspects in interactive communications. This also makes possible to analyze the influence that some technical aspects may have on socioemotional, and vice versa.

As we already knew, the quality of the video for the participants is still a big limitation, even for end users who are not experts in VR. In addition, this study has revealed the importance of the audio quality and delay, as well as interruptions in the conversation, something that we miss in previous experiments due to the simulation of the communication, serving as a motivation to further investigate interactive experiments.

In relation to the cartoon avatar, although it is a solution if compared with just a voice call, the local participants missed more gestures and non-verbal expressions, for example with the hands, and realism in the representation.

Different trends have been observed depending on the role of the remote participant. These differences have been motivated by the own character of the participants that made them feel more comfortable with their role, by the lack of additional tools such as taking notes or being able to see the material to guide the meeting from the remote side or the facility to interrupt and intervene from the remote side. This leads us to consider exploring for what type of meetings and from what roles and condition (local or remote), this technology can be more effective for hybrid meetings.

Since this experiment is based on a highly validated method, the Six Thinking Hats, other contribution is the usefulness of including known tools that aim to communicate into interactive XR environments. So, it is recommended to explore them instead of following scripts or simulate conversations to create a controlled environment.

As a general conclusion, this experiment has shown the valuable inputs obtained, in this case, from the free-form feedback that highly helps to understand the experience of the end-users and support the quantitative analysis. Additionally, other aspects obtained from this feedback will be considered to improve the experiment design.



# Chapter 6

## Tele-education Use Case

### 6.1 Introduction

The starting point to validate the prototype and the conclusions obtained from the proposed methodologies is a tele-education scenario. Before COVID-19, this technology was already a solution for those cases in which students must remain isolated or for those in rural areas who can not access school for weeks [57]. The pandemic has caused millions of children and adolescents to suffer confinement situations that have kept them away from their daily life at school. The most common modality of a virtual lesson has been based on the use of conventional cameras to acquire and transmit a small portion of the classroom, usually focused on the teacher and screen. This entails that remote students may lose a lot of information about the classroom and their classmates. As a consequence, there are students who have lost their performance in academic terms, but also students who have psychological disorders, motivated mainly by social distancing together with a feeling of fear and a deficit of social interaction.

Given this problem and the literature that demonstrates the benefits that 360-degree video communications can provide to end users, especially in teenagers where the acceptance of the technology is very high, this chapter focuses on tele-education [172, 173]. It presents a study of 360-degree video communications through a configuration of the prototype specially adapted for this use case. Although immersive communications can significantly increase the user experience offering new interaction and exploration possibilities [33], they still present issues related to the fact that the remote user can only visualize the viewport, likely missing relevant information taking place elsewhere. Thus, event occurrence or object presence notifications can help overcome these issues, improving the offered QoE. This prototype is capable of detecting the events of interest, through deep learning techniques, and notify them to the students that are using the virtual environment.

### 6.2 Prototype

Figure 6.1 presents the prototype proposed for the use case of tele-education and evaluated in this work. It allows a 360-degree communication between the classroom and the remote client, the

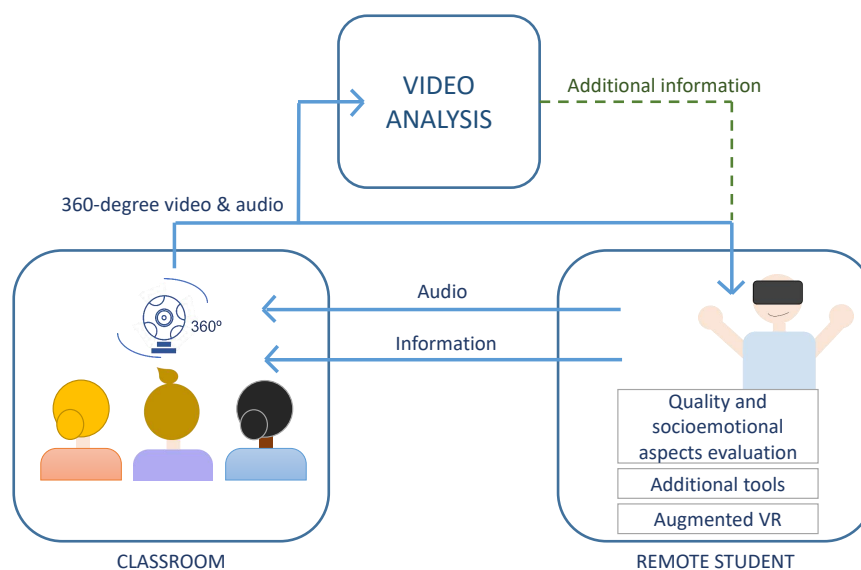


FIGURE 6.1: Architecture of the tele-education scenario.

student, who attends the lesson through an HMD [4]. Additionally, a video analysis module is added to detect events of interest in the classroom scene and notify them to the remote students, improving the QoE. In this work, two events of interest are considered: 1) students raising their hands to ask a question, and 2) changes in the slide presented in the screen. The video analysis module relies on Detectron2 [174] to detect and locate people (an example is presented in Figure 6.3) and screens. For each detected person, we first classify the position as standing or seated by means of a first Convolutional Neural Network (CNN). If seated, a second CNN is applied to classify if the hand of the person is raised. Both CNNs follow a similar architecture to the VGG [175] network with just three convolutional layers. Finally, to detect screen changes, the intensity changes within the area of the detected screen are monitored along time. However, the study of this module is out of scope in this work where the main interest is focused on the evaluation of the usability and acceptance of the developed prototype by remote students.

Notifications were displayed in the center position as it was the most noticeable location [176]. Once the notification appeared, there were three options: A) if participants attended the notification, it disappeared automatically when they looked at the interest area, B) if students chose to ignore the notification, they could disable it with the handheld controllers, and C) if the notification was not disabled and participants ignored it, it was removed by timeout. Two examples of notifications in the immersive environment are presented in Figure 6.2.

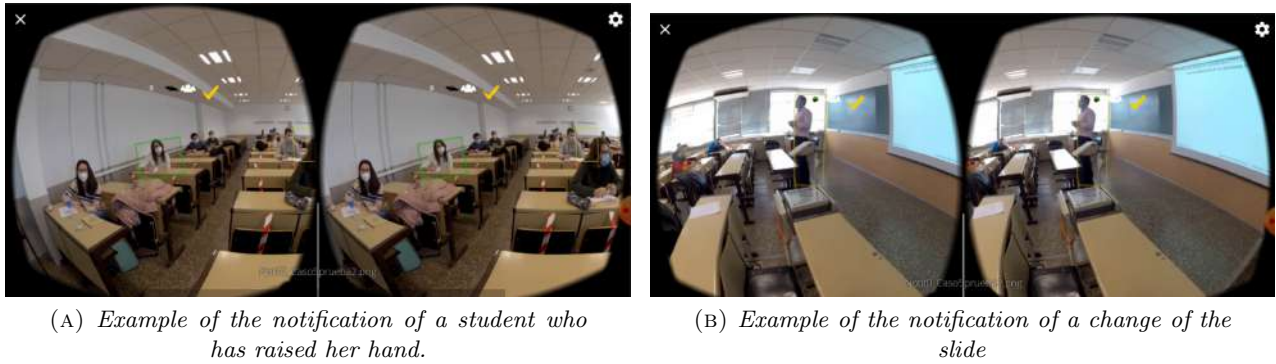


FIGURE 6.2: Example of accepted notifications of events of interest considered in tele-education scenario. The bounding box frames the student who has been detected raising her hand. At the top appears, the number of notifications waiting to be accepted, an arrow that indicates the direction in which the next event of interest has occurred, and a check that indicates that the notification has been accepted. This is also indicated by the color of the bounding box, yellow if user has not centered the viewport on the event, or green when user has already looked at and therefore accepted the notification.

### 6.3 EVENT-CLASS database

Prior to the notification of events or objects in the virtual environment, these elements need to be detected. To that end, deep learning techniques can be applied. Although most of the object or event detection algorithms required in the considered scenarios have already been successfully used in conventional video processing, the analysis of how deep learning techniques work on 360-degree remains highly unexplored. In this sense, distortions caused by the projections used to represent the spherical content should be considered. For example, in the commonly used equirectangular projection objects get increasingly distorted as they move away from the equator [19]. Although some works in the literature have applied corrections or transformations in the scene to preserve the similarity with conventional contents [177, 178], it would be more appropriate to have directly annotated training datasets of omnidirectional videos with acceptable video quality and the characteristics required for the use case [179].

In this work, we present a dataset of 360-degree videos acquired in the context of education. Given the lack of available datasets, the main goal is to contribute to the community with rich 360-degree videos taking into account: a) acquisition perspective (actor or observer), b) scene (different classrooms from the university), c) lighting (natural and/or artificial), d) camera location, e) density population, f) several 360-degree cameras, g) resolution, h) framerate, and i) bitrate. Additionally, in our dataset, the height of the camera approximately simulates the eye level of the remote student sitting at his/her desk [135], since it provides a more comfortable experience [180]. We also presumed that using this camera height may facilitate person detection with deep learning techniques trained with conventional videos, given that people in the scene appear around the equator, which is the least distorted area of the equirectangular projection.

### 6.3.1 Specifications

To record the 360-degree videos for the dataset, we used four cameras: RICOH Theta V, Vuze VR, Samsung Gear 360, and Insta360. The idea of using cameras with different specifications is to enrich the variety of sequences in the database, including scenarios with low-cost devices, which can support the extension of the technology to reach more schools and students. The main technical characteristics of the sequences captured with the different cameras are described in Table 6.1 (and in the folder of the dataset<sup>1</sup>). The raw material acquired with the cameras was stitched using the software from the corresponding manufacturers to obtain equirectangular videos.

TABLE 6.1: Technical properties of the stitched sequences recorded with each camera.

| Camera           | Resolution ( <i>pixels</i> ) | Framerate ( <i>fps</i> ) | Codec     | Bitrate ( <i>Mbps</i> ) |
|------------------|------------------------------|--------------------------|-----------|-------------------------|
| RICOH Theta V    | 3840x1920                    | 30                       | H.264/AVC | 56                      |
| Vuze VR          | 3840x2160                    | 30 & 60                  | H.264/AVC | 57                      |
| Samsung Gear 360 | 2560x1280                    | 30                       | H.264/AVC | 22 & 44                 |
| Insta360         | 3840x1920                    | 30 & 60                  | H.264/AVC | 62 & 126                |

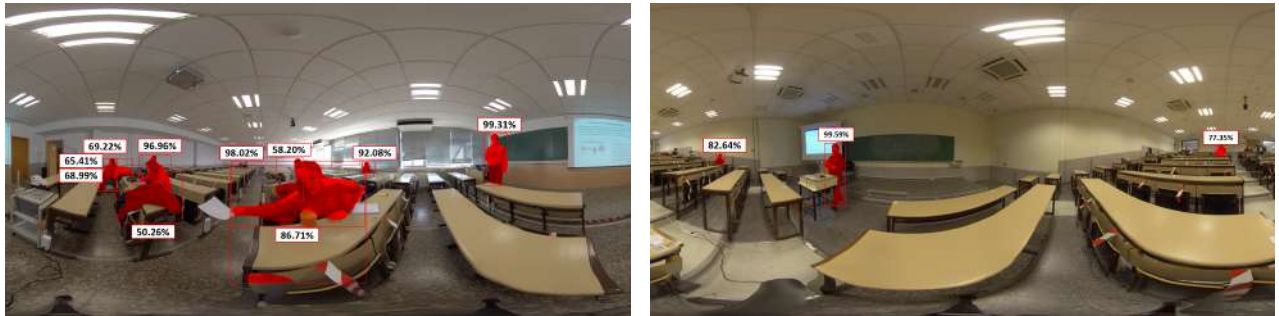
The dataset already includes more than 50 videos (more than 9 hours of recordings in total). The videos have been recorded at the Universidad Politécnica de Madrid considering different classrooms, subjects, students, and professors. To increase the variability of the scenarios, we consider a conventional classroom, a laboratory, where students are seated at their desks in front of the computers, and an auditorium type classroom. Different lighting conditions are also covered, with artificial lights in all environments and some of them also with natural light.

The recordings have been acquired during real lectures, increasing the spontaneity of the participants of the contents. All participants have been informed and signed a consent form that allows us to make publicly available the EVENT-CLASS dataset in accordance with the GDPR of the European Union.

The camera is located at a height of 150 cm, providing a comfortable experience for both standing and sitting viewers [180]. Also, we consider the distance to the students and the professor to avoid stitching artifacts in the videos. In relation to the acquisition perspective, we contemplate both, the actor’s perspective, when the camera is located on the desk, and the observer’s one, when the camera is located at the corridor, front-rows, center position, one side of the classroom, etc.

Additionally, audio has also been recorded, since it enriches the dataset and it would be necessary if some content of our dataset is used in subjective assessments. The recordings were done with the microphones of the cameras and their default configurations, and the languages are Spanish and English.

The dataset is publicly available for research purposes [181] containing the videos and their characteristics (e.g., resolution, framerate, bitrate, semantic descriptions, spatial and temporal information [123, 182], etc.). More materials will be added soon: more videos, ground-truth annotations for training and testing machine-learning algorithms for person detection, etc.



(A) *Conventional classroom.*

(B) *Auditorium classroom.*

FIGURE 6.3: Examples of people detection obtained in equirectangular frames.

## 6.4 Evaluation of the performance of an immersive system for tele-education

### 6.4.1 Related work

The interest in education and training through VR has received a great boost in several disciplines from economics, political science or literature, to biology, art, or medicine [159, 183, 184]. In fact, Stanford students and professors have decided to move the lessons to VR to learn about the technology and abandon the web-based platforms, such as Zoom or Teams. The future of classes is the metaverse, which is an ideal environment for dynamics to work and discuss in small groups [185]. Other studies have already analyzed the learning curve when the subject is learning through VR platforms, obtaining noteworthy results [186], as well as how the engagement of remote users of creative collaboration scenarios can be improved through immersive environments [187].

Several factors influence the experience of the users of immersive technologies and have been studied in related works, such as immersion [188], spatial awareness [43], and presence [100]. In our work, we mainly focused on presence, which is one of the most relevant socioemotional factors studied in immersive experiences and we assume that it influences the social distancing problems detected in the use case of tele-education. However, the main problem found in the literature, and corroborated with the work presented in Chapter 3, is that this type of questionnaires evaluate factors that are not useful in all type of applications and, sometimes, a large number of items are used. In the case of 360-degree video communications, where an interaction is provided or simulated, presence can be divided

into social and spatial presence. They can be influenced by the acquisition perspective: actor or observer. Some adaptations of the highly tested questionnaires have been already validated in the literature to consider this phenomenon [102]. This technology entails another factor that makes the acceptance curve decrease, which is related to the fact that HMDs may not be comfortable and cause dizziness or discomfort. Specific questionnaires such as the Simulator Sickness Questionnaire (SSQ) [189] or the Vertigo scale [190] have been already validated to measure it.

In addition, although we did not find in the literature works related to the detection of events that occur outside the viewport of the participant with the HMD, there have been some works related to how to present notifications to the user. Some of these notifications can be: a message received or an interaction in a social network [191]. Rzaev et al. [176] analyzed the effect of this type of notifications taking into account the placement (the considered environment offered six DoF), task, and environment. They proposed a methodology based on subjective questionnaires and objective measures, such as missed notifications or response time. Likewise, several studies in the literature explored different formats of notifications. George et al. [192] compared notifications with text, spotlight or ambient light in two scenarios. They conclude that the notifications with text are the ones that obtained the shortest reaction times, followed by the spotlight and the ambient light. In addition, the spotlight is the only one with which notifications are lost. Taking this literature into account, we based the presented methodology from studies which explore notifications that come from the smartphone placed in the HMD used to attend the immersive session.

#### 6.4.2 Research questions

In this section, the Research Questions (RQs) investigated in the experiment are presented:

- **RQ1:** What overall quality does the developed tele-education prototype based on 360-degree videos offer?
- **RQ2:** Could the platform (immersive or non-immersive, such as Zoom or Teams) influence the experience of the remote students?
- **RQ3:** Could the acquisition perspective and factors, such as camera location, influence the experience of the remote students?
- **RQ4:** Could the camera, non-professional or professional, influence the experience of the remote students?
- **RQ5:** Could this prototype offer a suitable solution for all types of classrooms?
- **RQ6:** Could the notifications influence the quality of experience of the remote students?

To answer these questions, we have considered several conditions in the test to validate the proposed prototype in the tele-education scenario. Specifically, to respond RQ1, we evaluated the overall quality perceived by users at the end of the visualization of each sequence. Hypothesis H1 was that the overall quality perceived by the participants was higher in the immersive than in the non-immersive ones, since the source contents were the same and despite the limitations of the HMD, other socioemotional aspects could increase the overall quality of the experience in immersive systems.

To answer RQ2, we considered three experimental conditions: visualizing the 360-degree scene with and without notifications, and one non-immersive simulating the type of platform that most students were using when they could not attend lessons in person, such as Teams or Zoom. So, the hypothesis H2 was that QoE of the participants was higher in VR with notifications than in VR, and in both cases higher than in the screen condition.

RQ3, RQ4, and RQ5 are questions related with the physical conditions of the environment while the acquisition and transmission of the scene. Then, we have looked for a 360-degree videos dataset that considered several video specifications in terms of acquisition perspective, cameras, and types of classrooms. To answer RQ3, we hypothesised H3 that actor perspective and centered location provided higher QoE than observer perspective and aside location of the camera. To answer RQ4, we chose three cameras, one professional (Insta360) and two non-professional (RICOH Theta V and Samsung Gear 360). Hypothesis H4 was that there were perceivable visual differences between the professional and the two non-professional cameras in the QoE of the participants. RQ5 led to hypothesis H5, which assumed that the system offered a solution, without significant differences, for the three types of classrooms considered: conventional, laboratory, and auditorium.

Finally, the RQ6 was proposed to analyze the usefulness of the notifications developed in the prototype, expecting as H6 a good acceptance of them.

### 6.4.3 Experimental conditions

One of the main goals of this test is to compare the proposed prototype and the technologies currently used in tele-education. In order to control the environment of the experiment, the 360-degree video communication was simulated. Participants visualized the same 360-degree videos, independent of the experimental conditions, recorded in regular lessons at the Universidad Politécnica de Madrid [193].

Three experimental conditions were considered in this test. Each participant evaluated two conditions that were randomized and counterbalanced to avoid order and learning effects. Conditions explained in detail are listed below.

- **Screen condition.** Participants assigned to this condition visualized a viewport of the 360-degree scene, centered on the professor and the slides, in a screen. This condition was the most similar with the web-based platforms that are currently used in most cases of tele-education.
- **VR condition.** Participants assigned to this condition visualized the 360-degree scene, without any additional tool.
- **VR + notifications condition.** The video analysis module was applied on the sequences detecting the interest event and creating the notifications. Then, participants assigned to this condition visualized the 360-degree scene and notifications appeared when an event occurred, as presented in Figure 6.2.

#### 6.4.4 Stimuli

Table 6.2 presents the main specifications of the SRCs. An example of each of the SRC is presented in Figure 6.4. The motivation behind this selection was to cover a wide range of properties: two videos acquired in each type of classroom (which also determines the lighting condition), three cameras considered in the acquisition compared in pairs using the professional camera as reference to avoid missing information due to the quality of the video, two acquisition perspectives, and two camera locations.

TABLE 6.2: Main specifications of the video sources considered in the study in terms of the scene (classroom type, acquisition perspective and camera location) and camera specifications (camera, resolution and framerate, and bitrate).

| Content | Classroom type | Camera           | Acquisition perspective | Camera location | Resolution ( <i>pixels</i> ) & framerate ( <i>fps</i> ) | Bitrate ( <i>Mbps</i> ) |
|---------|----------------|------------------|-------------------------|-----------------|---|-------------------------|
| Case2   | Laboratory     | Insta360         | Observer                | Centered        | 3840x1920/30  | 4.47                    |
| Case4   | Conventional   | Insta360         | Actor                   | Centered        | 3840x1920/30  | 7.81                    |
| Case5   | Conventional   | Insta360         | Actor                   | Aside           | 3840x1920/30  | 3.89                    |
| Case7   | Laboratory     | RICOH Theta V    | Observer                | Centered        | 3840x1920/30  | 8.38                    |
| Case15  | Auditorium     | Insta360         | Actor                   | Aside           | 3840x1920/30  | 4.22                    |
| Case16  | Auditorium     | Samsung Gear 360 | Actor                   | Aside           | 2560x1280/30  | 1.39                    |

The SRCs were maintained in the PVSs, videos presented to the participants, used in the test. It is based on the fact that the only process applied to the SRCs is the video analysis module to detect the events of interest, send them to the mobile application, and finally, record the scene with the notifications that appear in the virtual environment, resulting in the PVSs. The PVSs were randomized during the test session avoiding the consecutive presentation of videos recorded in the same classroom. The number of notifications events and people in the classroom are summarized in Table 6.3.



**Personal information:** For each observer, we collected age, gender, vision (corrected or normal), experience with VR (Basic: I have used this technology less than 10 times, Intermediate: I have used this technology between 10 and 30 times, Expert: I have used this technology more than 30 times), and experience attending online lessons, tele-education experience (Basic: I have attended online lessons less than 10 times, Intermediate: I have attended online lessons between 10 and 30 times, Expert: I have attended online lessons more than 30 times). We use this information to characterize the participants and guarantee the diversity in the sample.

**Quality.** The aggregate quality was asked, following the literature [37], in the post-sequence questionnaire using the Absolute Category Rating (ACR) on a five-grade quality scale [25], where the categories: “Bad”, “Poor”, “Fair”, “Good”, and “Excellent” were presented.

**Presence.** Spatial and social presence experienced by the observers were evaluated with five and four questions, respectively, obtained from the state of the art [102] and tested in our previous experiments in the same environment [54]. Observers provided ratings on a seven-level Likert scale (where 1 = “Strongly disagree”, to 7 = “Strongly agree”). Specifically, spatial presence was measured with the following questions: *I felt I was present in the places shown in the video (PP1)*, *I felt surrounded by the actions in the video (PP2)*, *I felt I was sitting by the table at the place of the video (PP3)*, and *I felt I could have reached out and touched the items on the table of the video (PP4)*. Likewise, social presence was measured with the following ones: *I felt that people were talking to me (SP1)*, *I felt that I was listening to the others in the video (SP2)*, *I felt I was present with the other people in the video (SP3)*, *I felt like the people in the video could see me (SP4)*, and *I felt I was actually interacting with other people (SP5)*.

**Simulator Sickness.** To evaluate the discomfort of the participants while visualizing the sequences, we chose one of the most tested questionnaires: SSQ [189]. It measures 16 symptoms divided into three categories: oculomotor, nausea, and disorientation. They were rated on a four-level scale (where 0 = “None”, 1 = “Slight”, 2 = “Moderate”, and 3 = “Severe”). In addition, a single-question was asked: “Are you feeling any sickness or discomfort now?” [190]. It was rated on a five-level scale (where 0 = “no problem” to 5 = “unbearable”).

**Usability.** System Usability Scale (SUS) considers general questions, applicable to different technologies, that assess how useful and functional the tested technology is [194]. SUS was evaluated with 10 items rated on a seven-point Likert scale (where 1 = “Strongly disagree”, to 7 = “Strongly agree”). *I think that I would like to use this system frequently (SU1)*, *I found the system unnecessarily complex (SU2)*, *I thought the system was easy to use (SU3)*, *I think that I would need the support of a technical person to be able to use this system (SU4)*, *I found the various functions*

*in this system were well integrated (SU5), I thought there was too much inconsistency in this system (SU6), I would imagine that most people would learn to use this system very quickly (SU7), I found the system very cumbersome to use (SU8), I felt very confident using the system (SU9), I needed to learn a lot of things before I could get going with this system (SU10).*

**Notifications.** Based on the literature, to evaluate the notifications, a specific questionnaire was adapted to the tested prototype [176, 191]. Then, participants rated seven items in a seven-point Likert scale. *I would be comfortable using this notification mechanism in VR (NQ1), These notifications are annoying when using the HMD (NQ2), This system provides to me the notifications I would like to receive (NQ3), With this notification system I have the feeling that I no longer lose information (NQ4), How easy or difficult is it to notice the notification? (where 1 = “very easy”, to 7 = “very hard”) (NQ5), Once you notice the notification, how easy or difficult is it to understand what it stands for? (where 1 = “very easy”, to 7 = “very hard”) (NQ6), How much of a hindrance was the notification to the overall VR experience? (where 1 = “not a hindrance at all”, to 7 = “totally an hindrance”) (NQ7).*

#### 6.4.6 Equipment and environment

The visualization of the contents in immersive conditions was carried out using a Samsung Gear VR with a Samsung Galaxy S8. These devices were selected because they provide the user with three degrees of freedom and acceptable quality, which are considered enough to attend an online lesson. In addition, the price makes them accessible devices for a greater number of students and therefore, a more realistic environment for the experiment. Due to the COVID-19 restrictions, HMDs were covered with a mask to avoid any infection between participants. In addition, different HMDs were alternated between consecutive participants and disinfected before each use. In the non-immersive condition, screen condition, the content was displayed in a 20” screen. Participants in all conditions heard the monophonic audio of the source contents through headphones.

The environment selected for the experiment was a quiet room, spacious to be able to maintain safety distances and ventilated. Swivel chairs were used to allow movement of participants.

#### 6.4.7 Test session

Figure 6.5 presents the structure of the session followed in the experiment. First, a brief explanation of the experiment and conditions was presented to the participants. In addition, in this part of the session participants filled in the consent form, necessary to attend the test and process their data in accordance with the GDPR of the European Union. They filled in the pre-session questionnaire

and once they completed it, they were ready to start the training session. The training session was used to show the participants a short video of similar characteristics in order to understand the dynamics of the experiment. Also, participants assigned to the VR+notifications condition, had the opportunity to see how the notifications are presented and how they could ignore or remove them in the virtual environment. Once the training session was over and their doubts were resolved, they visualized a sequence and at the end they answered the post-sequence questionnaire. This process was repeated with each of the six sequences evaluated with each condition. After the visualization of the last one, they filled in the post-condition questionnaire and enjoyed a break of 15 minutes. After this break, assuring that they were ready to continue with the test, each participant was assigned to other condition and repeated the structure presented in Figure 6.5 from the training session. The duration of the test session was approximately of 50 minutes.

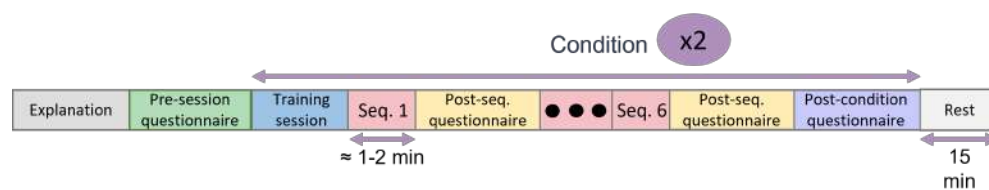


FIGURE 6.5: Diagram of the test session structure. Each participant was randomly assigned to two different conditions, so each participant carried out this part of the test twice. Then, the duration of the test session was around 50 minutes.

#### 6.4.8 Participants

A total of 42 observers (50% female, 50% male) participated in the experiment. Three participants were eliminated because their scores were not collected correctly, resulting in a total of 39 observers (48.7% female, 51.3% males). There were participants in the age range between 18 and 29 years, with an M age of 22 and a SD of 2.631. Each observer tested two experimental conditions so, each condition was evaluated by 26 participants and each video was rated 78 times. They received a small financial reward for participating.

The gender was used to uniformly distribute participants under VR, VR+notifications, and Screen conditions, guaranteeing a balanced sample. All observers were checked for normal or corrected-to-normal vision. Following the literature, Snellen chart and Ishihara test were carried out before the tests [25]. Additionally, the level of experience in VR use and tele-education were collected and this information is presented in Figure 6.6. As can be observed, only 16.7% of the sample have used this technology more than 10 times. On the contrary, only 2.4% of were not very familiar attending online lessons. This result was expected because most of the participants were students from the university and due to COVID-19 they have attended online lessons.

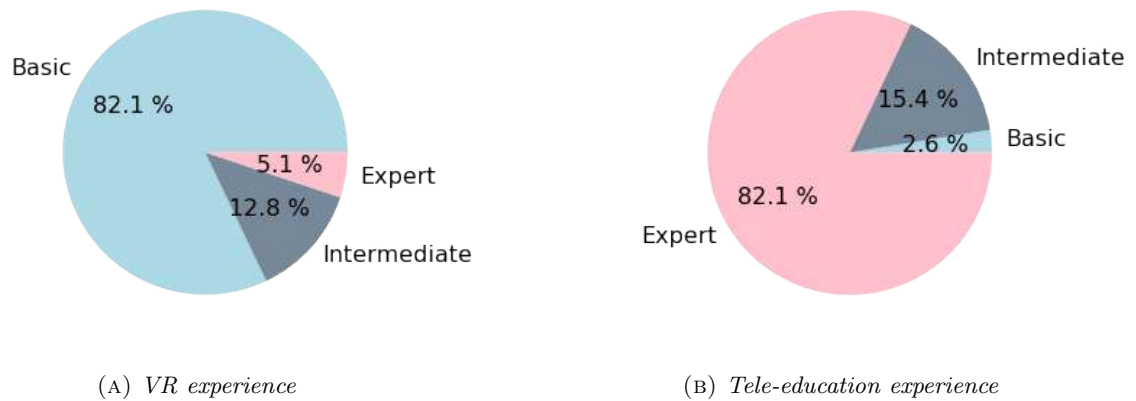


FIGURE 6.6: Characterization of the 39 participants of the study in terms of their experience in VR and attending online lessons.

TABLE 6.5: Difference in aggregate quality, spatial and social presence, and usability between the three conditions.

| Questionnaire items               | Screen                          | VR                              | VR+notifications                | Significance                                    |
|-----------------------------------|---------------------------------|---------------------------------|---------------------------------|---|
| Aggregate quality (5-level scale) | $M = 3.038$<br>( $SD = 1.185$ ) | $M = 3.359$<br>( $SD = 1.009$ ) | $M = 3.417$<br>( $SD = 1.053$ ) | $F_{2,450} = 8.274, p = .0004, \eta_p^2 = .023$ |
| Spatial Presence (7-level scale)  | $M = 2.686$<br>( $SD = 1.203$ ) | $M = 5.011$<br>( $SD = 1.284$ ) | $M = 5.285$<br>( $SD = 1.163$ ) | $\chi^2 = 164.839, p = 0, df = 2$               |
| Social Presence (7-level scale)   | $M = 2.678$<br>( $SD = 1.315$ ) | $M = 4.89$<br>( $SD = 1.301$ )  | $M = 5.119$<br>( $SD = 1.194$ ) | $\chi^2 = 133.52, p = 0, df = 2$                |
| System Usability (7-scale level)  | $M = 4.688$<br>( $SD = 1.095$ ) | $M = 4.562$<br>( $SD = .99$ )   | $M = 4.696$<br>( $SD = .896$ )  | $\chi^2 = 1, p = .39, df = 2$                   |

#### 6.4.9 Experimental results

The main results obtained from the analysis are presented in this section. The factors that are analyzed are those that are asked through questionnaires: quality, spatial and social presence, simulator sickness, system usability, and notifications. We also indicate the research questions and hypotheses addressed in each analysis.

The data has been analyzed depending on its nature. For normal distribution of the data, the ANOVA was applied to examine the significant differences. Likewise, for cases where the distribution was not normal, non-parametric tests were compute. Specifically, the Friedman test was applied due to the dependence between the samples. As a post-hoc analysis, multiple comparisons with Bonferroni correction were applied to examine the differences among the evaluated videos and conditions. In all analysis, the considered level of significance was 0.05. Table 6.5 summarizes the main results of the items evaluated in the experiment and the significance ( $F$ ,  $p$ , partial eta-squared  $\eta_p^2$  between conditions).

**Quality.** The aggregate quality, evaluated at the end of each sequence in all conditions, was examined with the MOS and the associated 95% CIs, presented in Figure 6.7. The main goal was to analyze significant differences between conditions and qualities, determined by the specifications of the cameras used for the recordings. To better explore the significant differences between the contents, a parametric analysis was applied, following the recommendations in the literature [117]. It is based on the fact that the evaluation of video quality on a 5-level score can be modeled by a Gaussian random process [142]. ANOVA results show significant differences between contents ( $F_{2,450} = 48.401$ ,  $p = 0$ ,  $\eta^2 = .335$ ). There are significant differences between all the videos and the Case16, recorded with the Samsung Gear 360, which offers a much lower quality than the rest of the tested cameras. In addition, there are significant differences between Case7 ( $M = 3.179$ ;  $SD = .922$ ) and Case2 ( $M = 3.827$ ;  $SD = .755$ ;  $p = .001$ ) and Case 5 ( $M = 3.627$ ;  $SD = .969$ ;  $p = .027$ ). This result means that there is a noticeable difference between attend a laboratory lesson through a professional (Insta360) and non-professional camera (RICOH Theta V). These result relates to our RQ4 and verifies our hypothesis H4 that non-professional cameras may provide a lower QoE. However, the differences between other types of classrooms (RQ5/H5) or acquisition perspectives (RQ3/H3) should be addressed in depth in future experiments. In addition, the results obtained are the expected results in subjective quality assessments following the literature so, the quality was not influenced by the specific use case and context of tele-education.

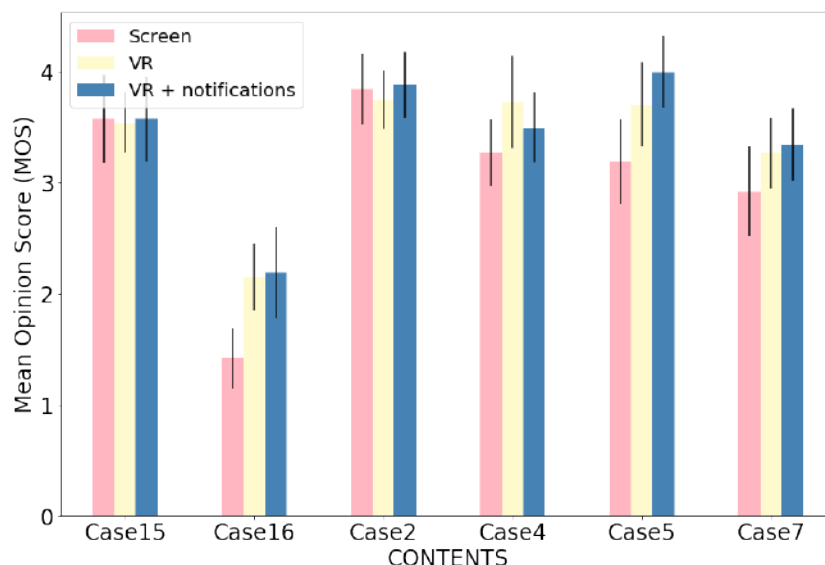


FIGURE 6.7: The mean opinion scores (y-axis) on a five-level scale obtained from 39 participants, taking into account that each one rated two conditions, who evaluated the perceived video quality at the end of each sequence (x-axis). Error bars represent 95% CI.

As presented in Table 6.5, there are significant differences between the experimental conditions. A parametric post-hoc analysis, Tukey HSD, was applied and significant differences were found

between non-immersive and immersive conditions, VR ( $p = .007$ ) and VR+notifications ( $p = .003$ ). This result shows that the immersive systems can provide a higher overall quality in tele-education than current technologies, as hypothesized (H1) for our RQ1, which can improve the experience of attending a class remotely, answering our RQ2.

**Spatial and Social Presence.** Spatial and Social presence were analyzed aggregating the rates obtained in the four and five items, respectively (PPQ and SPQ). Figure 6.8 presents the distribution of the aggregated evaluations of social and spatial presence for each of the analyzed contents and by conditions. From the figure we can obtain some insights such as that the three conditions offers a similar, and very positive for immersive conditions, sense of social and spatial presence in different types of class, responding to RQ5. Therefore, next study could focus on one type of classroom to analyze different perspectives of acquisition and location of the camera to obtain more reliable conclusions that respond to our RQ3.

Due to non-normal condition of the data, Friedman test was applied. As presented in Table 6.5, significant differences were found between experimental conditions. Wilcoxon signed-rank test with Bonferroni correction was applied to better analyze these differences. Significant differences were found in spatial presence between non-immersive condition and immersive experimental conditions, VR ( $p = 0$ ) and VR+notifications ( $p = 0$ ). Likewise, significant differences were found in social presence between non-immersive condition and immersive experimental conditions, VR ( $p = 0$ ) and VR+notifications ( $p = 0$ ). As expected, this result motivates the benefits of 360-degree video communication to attend a lesson remotely, confirming our hypothesis H2 for the RQ2.

**Simulator Sickness.** As previously presented, the SSQ, highly tested in the literature, was used to evaluate the simulator sickness. First, Figure 6.9a presents the histogram distribution of the Total Score (TS) taking into account the two parts of the test session. As can be observed, the TS values obtained during the session are low and therefore we can guarantee that all the participants could attend the lesson remotely with this technology without symptoms or discomfort. This result supports the appropriateness of the immersive solutions for tele-education (RQ2/H2), which provide a high quality to the end users. Second, Figure 6.9b presents the global scores for each evaluated factor, the TS, and the Vertigo Scale in the three points of measurements. The evolution presented in this figure is ascending but the maximum values are in the expected range for both scales, SSQ and Vertigo Scale [37].

**System Usability.** Observers evaluated the usability of the system at the end of each condition, as presented in Table 6.4. Then, we analyze the usability of the experimental conditions aggregating the 10 items of the SUS and directly computing the average of the collected scores. Note that some items on the scale were scored in reverse fashion, so the ratings were modified to calculate

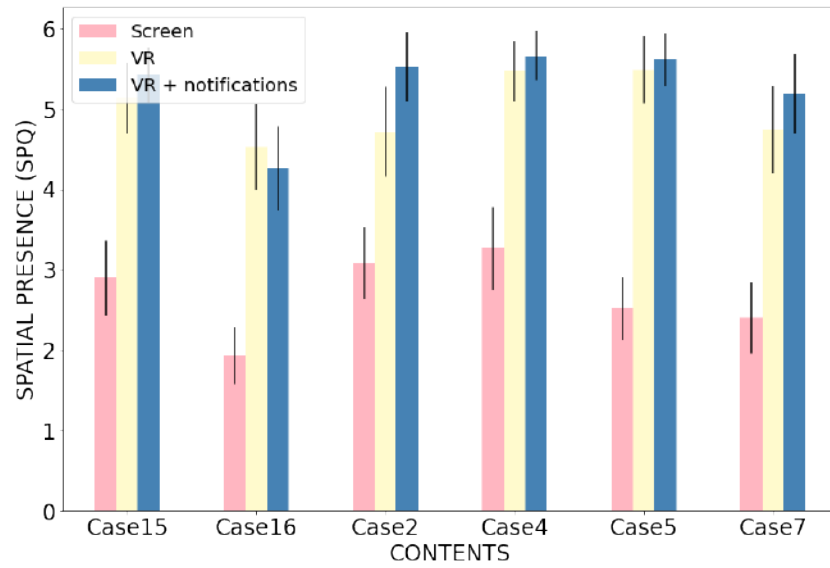
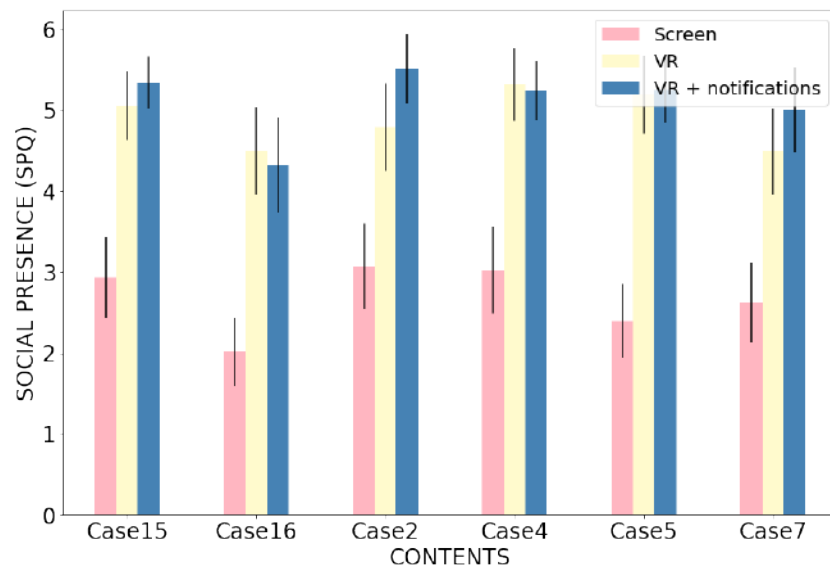
(A) *Spatial Presence*(B) *Social Presence*

FIGURE 6.8: The aggregate Spatial Presence and Social Presence (y-axis) on a seven-level scale obtained from 39 participants after the visualization of each sequence (x-axis) in three conditions. Error bars represent 95% CI.

the aggregate measure of usability. Friedman test was applied to explore differences between experimental conditions, but there are not significant differences between them. This result, related to RQ2/H2 and RQ5/H5, could be explained based on the fact that the session was simulated and participants had not the possibility to interact with their colleagues and professor. However, we can assure that the system is useful and could be accepted in different type of classrooms. In addition, the non-immersive condition is the one that the participants are currently using at the

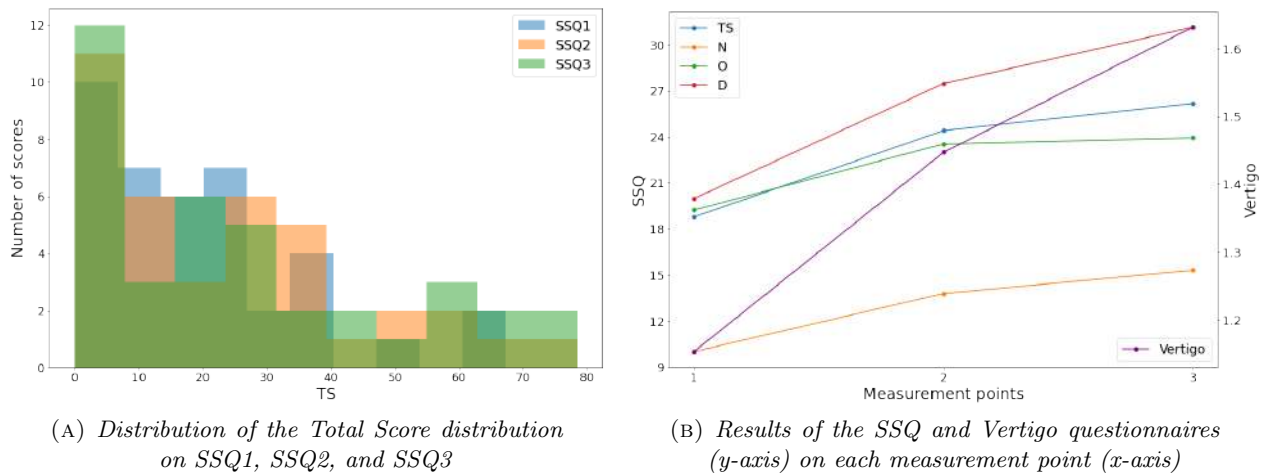


FIGURE 6.9: Simulator Sickness results obtained from the evaluations of the 39 participants at different time points during the test session.

university, so they are more familiar with it. Non-immersive condition, whose usability has been shown during this pandemic time, can be improved with immersive communications, as seen with the quality scores.

TABLE 6.6: Notifications scores obtained from the evaluations of the 28 participants assigned to VR+notifications condition at the end of the test session

| Question | Average ( $M$ ) | Standard deviation ( $SD$ ) |
|----------|-----------------|-----------------------------|
| NQ1      | 5.429           | 1.116                       |
| NQ2      | 5.321           | 1.44                        |
| NQ3      | 5.393           | 1.047                       |
| NQ4      | 5.643           | 1.288                       |
| NQ5      | 6.5             | 0.779                       |
| NQ6      | 5.893           | 0.9                         |
| NQ7      | 6.107           | 1.080                       |

**Notifications.** Table 6.6 presents the evaluations obtained for notifications. As can be observed, the scores are very high, which means that the notifications have a positive impact and they are useful for attending a lesson. This confirms our hypothesis H6 for the RQ6, showing that notifications, such as those covered in this study, can be useful for the users of immersive tele-education systems. Note that NQ2 and NQ7 were scored in reverse fashion, so the ratings were modified to calculate the final ratings.

#### 6.4.10 Conclusions

This work presents a user study carried out to validate a remote communication system for tele-education based on 360-degree video. This system streams in real time a class, detects events such as changes of slides and raised hands, and notifies them to remote participants. This way, students that are attending remotely the class may not miss any event happening in the physical classroom, as it may happen when using conventional tele-education streaming services that only show a small portion of the scene (e.g., the teacher and the slides).

The results from the study, in which a balanced group of 39 participants was involved, showed that using immersive communications in tele-education scenarios significantly improve in social and spatial presence perceived by remote students in the case of the use of tele-education. Additionally, events of interest and notifications in the virtual environment of their detection are highly valued and it can be stated that the acceptance of notifications is high. In this way, remote students can explore and be aware of what is happening in the physical classroom without missing important events.

Other conclusions obtained from the experiment are:

- As expected, there are significant differences in the overall quality. This means that despite the limitations of the 360-degree cameras and the HMD used during the test session, the quality is better than the obtained in the on-screen version such as Teams or Zoom. This result is a motivation to continue researching in this type of environment that offers benefits and feelings that are not determined purely by technical aspects.
- The results obtained from the simulator sickness questionnaires show that the use of this technology does not cause severe symptoms or discomfort that can hinder its use in real cases (considering sessions of 50 minutes as tested in the current experiment).
- Although the results for usability do not show significant differences between the three tested conditions, this can be caused by the questionnaire used in the tests, which may not be appropriate for the performed experiment, where no interactions were really involved, since the participants watched recorded videos and not online classes.

Pre-recorded videos have been used in the experiment that, despite having been acquired in real lessons, do not allow remote interaction with the classroom by the remote student. As future work, this tele-education scenario with interaction should be considered. Also, testing more professional and non-professional cameras in different environments would help to validate the prototype for the use case. Finally, other relevant factors related to immersive experiences should be covered

in future studies, such as immersion and awareness. This work can be a primary study for these future works in which the proposed methodology could be used.

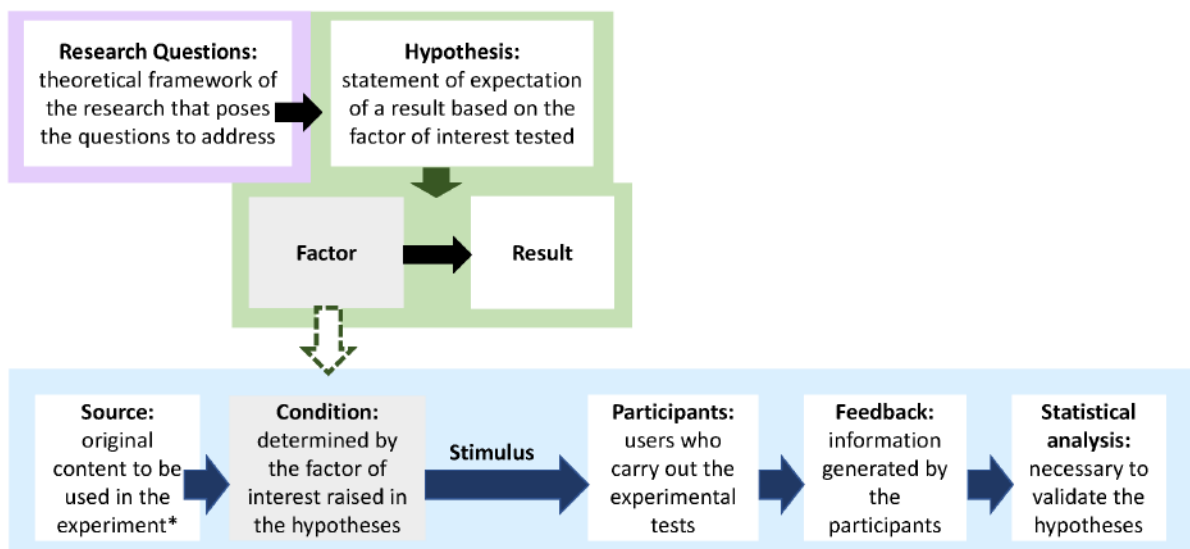


# Chapter 7

## Lessons Learned for the QoE Assessment of Immersive Communications

### 7.1 Introduction

The works presented previously have allowed us to understand some of the challenges presented by the assessment of the QoE in XR technologies. Specifically, the possibility to understand the two points of view that found in the research community, telecommunication and HCI research. Thanks to this understanding, we propose a common framework to merge them that summarizes and analyzes each of the stages required to design an experiment. The idea of conducting this analysis and presenting the findings of this research in this way is to share what we have learned in guidelines, which can be useful for developers, researchers, or service providers from different areas and levels of expertise.



\*It can be pre-recorded content or the real-time streamed scene during the XR communication

FIGURE 7.1: Theoretical framework of the stages necessary to design an experiment.

The framework is presented in Figure 7.1. At the top half of the diagram, we have the conceptualization of the experiment. Following HCI research principles: an original research question to generate hypotheses, typically about how the variation of a factor (independent variable) may affect the result of the experiment (dependent variable). At the bottom half of the diagram, the framework details the process to characterize this factor-result relation. It is based on ITU-T methodology (e.g., ITU-T Recs. P.913, P.919): selection of source content for the experiment, preparation of the test stimuli based on the source and the technical conditions evaluated in the experiment, collection of feedback from participants, and statistical analysis of the results. Specifically, the analysis is focused on: source, condition, participants, and feedback.

Our analysis is based on the reference configuration of the immersive telepresence scenario described in Figure 1.2. Nevertheless, the guidelines provided could be easily applicable to other formats of XR communication.

## 7.2 Source

The source content used in the evaluation, whether pre-recorded or streamed in real-time, greatly influences the socioemotional aspects and QoE perceived by the participants [32]. Then, a correct characterization of the source goes through homogenizing the origin of the stimulus evaluated by the participants, removing noise and increasing the reliability of the comparisons between conditions, results, and contributions. Here we present the most important aspects to take into account.

Initially, tests focused on the evaluation of parameters of the system in XR were based on practices designed for traditional content. In terms of video quality evaluation, source content was selected to cover a wide range of characteristics (e.g., color or texture). These methodologies can work for the evaluation of parameters of the encoding and transmission chain in XR technologies with some adaptations. However, if we also want to assess socioemotional aspects, we must take into account higher level aspects, such as the acquisition perspective or the narrative of the source [54]. If only questionnaires are used for the evaluation, the duration is chosen taking into account the number of conditions to be evaluated. Following ITU-T Rec. P.919, it is important to find a balance between the test duration and the limits of cognitive load and fatigue [37], avoiding discomfort in the participants. Such is the relevance of cognitive load that there are subjective and objective measures and even applications that measure it are being developed [195]. If biosensors are used in the experiment, then the duration of the interactions should consider the minimum time that a sensor needs to capture a biosignal to obtain results [196].

Once the context and scene of the XR environment have been selected, the processing of this information must be characterized objectively and semantically.

Regarding the **objective characterization**, spatial and temporal complexity of the content is considered following ITU-T Recs. P.910 and P.913 to compute the SI and TI indicators. Video quality offered during the test session must be evaluated to know the influence on socioemotional aspects and in what situations it can be critical. To measure this, objective metrics can be used. As summarized in Chapter 2, there are several options to apply objective metrics: on 360-degree video planar representation or on the viewport if the objective metric has been created for 2D content, or directly on 360-degree content if the objective metric has been adapted to the peculiarities of this type of content. The issue with options a) and b) is the same as that with 2D content in that there is sometimes a poor correlation between test results and human perception. In fact, our work presented in Chapter 2 has shown that objective quality metrics designed for 2D contents can be directly applied on planar representations. We also concluded that VMAF, the most relevant objective metric for 2D content, on the planar domain outperforms the results of PSNR, WS-PSNR, CPP-PSNR, SSIM, and MS-SSIM [52]. Although more research is required with other datasets and comparing with other new metrics, it can be used to provide a good characterization of the visual quality. An additional problem with option b) is that modelling eye movement and content saliency is still a challenge [197]. Once SI, TI, and an objective quality metric have been computed, their values must be reported with specifications such as resolution, framerate, bitrate, or encoder. A common phenomenon that occurs in acquisition of 360-degree videos is stitching. It is strongly recommended to avoid it, but sometimes it is not possible, so the camera should be carefully placed to minimize the interference between the stitching and the area of interest or post-process the scene [65].

For the **semantic characterization** it is necessary to report the **spatial features**, referring to the camera location and the observed scene (space or objects). In XR communications the location of the camera is the point from which users view the scene. Then, this information is important to prepare the participant's environment. Moreover, it is necessary to consider whether the remote user with the HMD is standing or sitting to set the height of the camera at approximately the average eye level [54]. Also, if remote users visualize the scene with a table in front of them, having a physical table when performing the experiment will increase the association between the table in the virtual environment with the real one. Likewise, the **temporal features**, referring to sudden scene changes that can occur throughout the session, can explain different phenomena of the results (e.g., camera motion which can increase the sickness of the participants).

The fact that users with HMD only visualize the viewport makes the acquisition perspective and the placement of the camera more relevant. The acquisition perspective can be: first-person/actor's

perspective or third-person/observer's perspective. At first it was believed that the freedom to choose the viewport could cause a negative effect, the FoMO, on users [198]. This could be a consequence of the fact that users can decide where to look at and miss events that may occur in the rest of the sphere. However, it has been shown that FOMO does not affect presence. Even the control over one's viewing experience can cause Joy of Missing Out (JoMO), which refers to the positive feelings that arise from the freedom to choose among mutually exclusive options [198]. We have found that a first person acquisition perspective and local people interacting with the camera makes remote users feel like local participants are looking at them, waving, smiling at them, increasing social presence [54] and, probably, JoMO.

This opens another discussion, if we use this XR environment for communication, we should encourage local users to interact more with remote users. In our scenario, local participants speak to a 360-degree camera, which is not a natural interaction. Thus, the representation of the remote participant in the local environment (e.g., traditional video, avatar, etc.) and the information that is sent from the remote participant should be addressed by the effect it induces on the interaction [199].

Additionally, the type of conversation also play a role in the design of the experiment. A discussion can evoke higher level of social presence than everyday or educational conversations. This result is in line with the capacity for persuasion and debate that 360-degree content offers and its possible use to change attitudes (e.g., environmental awareness or gender equality) [200].

The premise is that XR allows a better transmission of the non-verbal part of a conversation than previous technologies as, according to the literature, it makes a conversation more effective [30]. So, addressing the aspects that influence the transmission of this non-verbal part of communication will improve the experience. These aspects range from system factors (e.g., the representation of the remote user in the local environment) to contextual or human factors (e.g., the relationship between the participants).

### 7.3 Conditions

After selecting the sources, conditions are applied to generate the stimuli that evaluate participants. Each condition is a variation of the *influencing factor* considered in the hypothesis and each factor an *independent variable* of the experimental design.

From the design perspective, conditions are classified into two categories: **continuous** and **categorical**, with different implications for the assessment methodology.

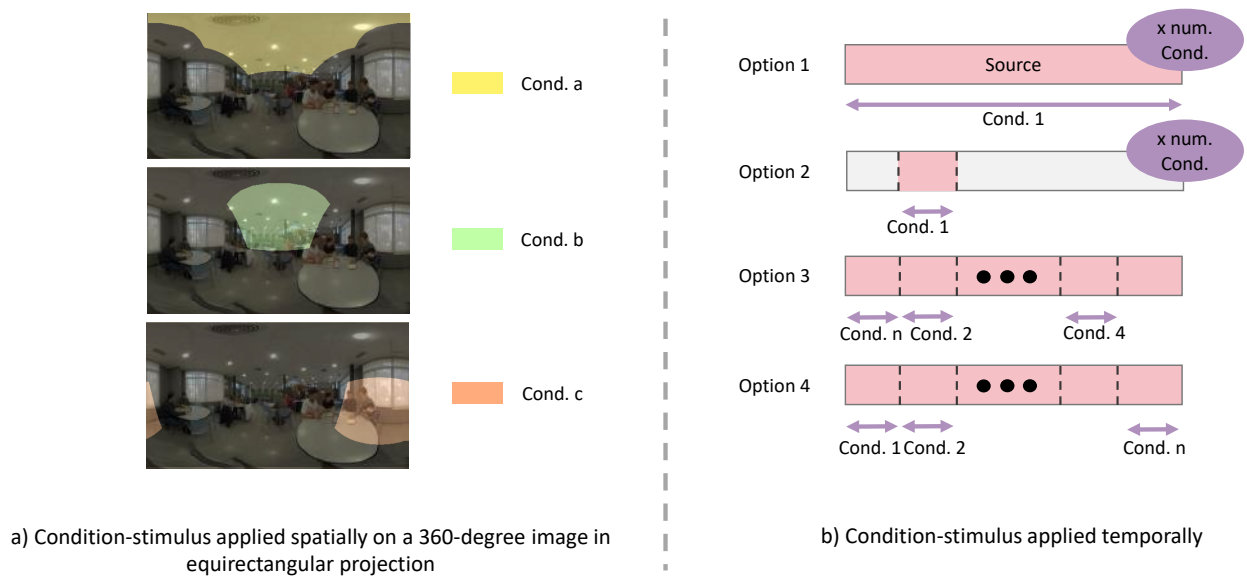


FIGURE 7.2: Example of the condition-stimulus applied spatially on a 360-degree scene and the condition-stimulus applied temporally.

Typical continuous variables are system factors related to communication or processing restrictions, such as bitrate or delay. On the other hand, categorical variables are normally associated to user experience design decisions (e.g., interaction method), system-wide restrictions (e.g., HMD), or context factors (e.g., genre of source contents) that cannot be changed without interrupting the session.

Here, we discuss the implications of the type of condition using two examples: video compression as continuous variable and interaction method as categorical one. The analysis, however, can be generalized to other conditions of the same kind.

### 7.3.1 Continuous conditions

Continuous Conditions can be applied spatially (e.g., encode each frame of the source content with different qualities) and/or temporally (e.g., encode source content with quality fluctuations). Then, the condition is either the encoding used for each area of the frame or the encoding used for each segment of the source content. The result of the application of this condition on the source is the *stimulus* rated by participants.

The spatial application of the condition is an intrinsic feature of the nature of the immersive environment. So, it is necessary to analyze the influencing factors in different areas of the scene. Figure 7.2, in the left side, presents an example of three different encodings inside the viewport (represented with different color) depending on where user is looking. 360-degree frames are presented in the most common planar representation, equirectangular projection. This type of solution

is proposed to offer better quality within the viewport than in the rest of the frame [37]. This solution requires the adaptation of the encoding scheme throughout the session to the behaviors of the users with low latency. The greatest interest appeared with the tiles from the H.265/HEVC standard [21] that divide the frame into independently encodable and decodable regions [197]. Despite their potential, in a real-time 360-degree video transmission the following limitations are tackled: 1) Difficulty in adapting the encoding scheme to user behavior with low latency, 2) Decoding tiles in real-time involves the use of high-capacity devices that often have a cost that is not accessible to the majority of the population, and 3) the relationship of bitrate savings with the computational and resources load that it entails is not worth it [201].

The options for applying the conditions temporarily are summarized in the right side of Figure 7.2. The direct option is **option 1** that applies the condition on all the source content. To make comparisons, different conditions must be applied to the same content to be visualized and evaluated by participants. Thus, testing multiple conditions requires that contents must be of short duration and/or not many to not increase the duration of the experiment, making it unfeasible. This limitation led to **option 2**, extracting a short clip from that source to apply the condition. In this way, a battery of short duration clips with different applied conditions is generated. Once the process is carried out with several source contents, the order of the visualization is randomized for each participant in each test session. Applying these options, each short clip is visualized and immediately evaluated. The process is repeated for each clip several times, depending on the number of conditions. This scenario is unrealistic and viewing the same content repeatedly causes users to lose attention due to fatigue [122], negatively affecting QoE.

Janowski et al. [125] propose **option 3**, an adaptation which avoids the repetition of the same clip. It chooses several short duration clips from the same source sequence and assumes similar characterization. In this way, a different clip is chosen for each of the conditions to be applied. With this, it was observed that the feeling of fatigue was decreased and the quality of the video could be evaluated while the scene remained mostly static. This assumption is applicable to a communication environment where the camera is supposed to be placed in the same location during a meeting without significant changes in the scene. However, since these clips are randomized for the visualization and the selection do not have to be sequential, this option is not yet suitable for a real-time transmission and interactive communication.

Finally, different conditions on the source content can be applied randomly with a fixed frequency and duration, **option 4**. Multiple stimuli are generated by one source to which different conditions are randomly applied over time. In the case of video compression evaluation, the stimulus is the source with quality fluctuations, simulating unstable network conditions. The same assumption as in the previous option can be made: the characterization of the scene is similar across the source

and, therefore, video quality can be assessed during the viewing of the content. Although this option has been less tested than the previous ones in the XR paradigm, it has many advantages. The fatigue in participants could be removed and this stimulus allows the evaluation of the studied system factors and how they affect QoE in XR real-time communication [122].

### **7.3.2 Categorical conditions**

Categorical Conditions evaluate system design decisions that cannot be changed immediately because it would disrupt the experience.

Categorical conditions are usually combined with continuous conditions. To compare categorical conditions, they should be randomized, and the continuous conditions analyzed in every categorical condition should be the same. Pilot studies with pre-selected conditions help to direct the research towards the most interesting ones, looking for significant or not significant differences between conditions.

As an example, XR technology is powerful in making remote users feel like they are moving into a virtual world. However, one of the great challenges it presents is related to the disruption with the real world. When users want to interact with the real or virtual environment, realistic techniques are needed that do not decrease the sense of presence.

Several works propose the use of augmented information to facilitate the interaction “with handheld controllers or hands. We explored the visualization of the hands of the user (categorical condition) applying color-segmentation techniques in a teleconferencing scenario [54]. We assume that the visualization of the hands to take notes in the physical world would influence the presence. To explore the impact of this categorical condition, we compared it with the condition in which participants neither visualize their hands nor take notes. In this case, continuous conditions were the same three contents that participants visualized in the categorical conditions. However, we did not find influence on presence. Even it did not influence the answers to the questions about the conversations. It means that XR technology as a communication platform is very powerful on its own. Additional tools given to the user enhance the experience depending on the specific tasks and use cases. However, they can add noise if the goal is to evaluate the technology in people’s daily lives as a communication platform. Therefore, categorical conditions are highly dependent on the use case.

The main works in the literature that have explored categorical conditions in communication use cases can be grouped into three groups:

- System specifications: 360-degree camera placement positions, HMDs, tablet, or other 2D devices, etc.
- Interaction with the virtual environment: real hands, controllers, hand gestures, cursor pointer, head pointer, etc.
- Self and others representations: realistic, full-body avatar, avatar without body, only hands, etc.

## 7.4 Feedback

The methodologies used to collect the participants' evaluations of the subjective assessments are described here. Although there are other forms (e.g., biosensors), this analysis focuses on questionnaires.

### 7.4.1 Questionnaires

Like the conditions section, it is focused from the point of view of analyzing a purely systemic factor, video compression, and a socioemotional factor, presence. However, the conclusions are applicable to other factors of the system, human, and context.

Questionnaires can be used at the beginning, during, or at the end of the session depending on the structure of the test and the factor evaluated. There are standardized questionnaires for the most analyzed aspects in interactive XR environments (e.g., presence). Nevertheless, there is a lack of methodologies that allow the joint analysis of system, context, and human factors, posing several challenges to solve in our experiments.

The **overall quality** of experience is usually the first question asked at the end of the stimulus. It is evaluated on a 5-level scale and the results are aggregated through a MOS. This is a recommended question as it provides an overall experience score. Of course, it is an indicator that needs to be supported by proper evaluation of other aspects. If differences are found between the tested conditions, a deeper analysis and understanding of the origin of these differences is conducted.

Regarding video compression, the most relevant methodologies are briefly described below.

- **Absolute Category Ratings (ACR)** consists of evaluating short-duration clips encoded with different qualities. This is the most proven methodology in 2D and 360-degree content. It is used to analyze the factors that affect the visual dimension of communication. It is

applicable with conditions such as those of options 1, 2, and 3, presented in Figure 7.2, and the disadvantages and advantages have already been discussed.

- **Single-Stimulus Discrete Quality Evaluation (SSDQE)** consists of presenting long-duration contents encoded with quality fluctuations (option 4). Before each quality change, the quality is rated. The question appears and the video continues to play in the background without interrupting the narrative. It is suitable for the simultaneous assessment of video quality and socioemotional features such as presence, attitude or attention [54].
- **Single-Stimulus Continuous Quality Evaluation (SSCQE)** consists of presenting long-duration contents encoded with quality fluctuations. The difference with the previous one is that participants are continuously rating the perceived quality with, for example, a slider. Although this methodology is applicable to the same conditions as the previous one, it is more demanding for the participants. The continuous task can influence, for example, the attention to the conversation.

The XR environment is so wide and there is such a diversity of experimental conditions that all elements of the questionnaires can not fit into the tested environment. Generally, many factors are chosen to be evaluated and, therefore, the use of long questionnaires for each of the aspects is not feasible. An alternative is to subsample these questionnaires. The problem is that these questionnaire subsamplings are not always validated, which can lead to unreliable conclusions. In relation to the evaluation of presence, we have experienced these challenges and compared different presence questionnaires obtaining results that correlate but present differences, since the questionnaires focus the evaluation of presence on different dimensions [54].

Based on this analysis, we decided to apply a questionnaire based on five items to assess spatial presence, defined as the sense of being there, and social presence, defined as the sense that people is talking to you [102]. The main motivation was that the items used are brief, clear, and perfectly adapted to the teleconference environment under evaluation [54].

Standardized or specific questionnaires of the experiment can be supplemented by semi-structured interviews or free-form feedback. With them, researchers can have a closer opinion from the participants that can be relevant to explain some unexpected phenomena during the analysis of the data. At the origin of subjective experiments, this type of interview was widely used, but the processing and drawing of conclusions was laborious. Currently, semi-structured interviews and free-form feedback are used again thanks to transcription and machine learning algorithms. The selection between a semi-structured interview or a free-form feedback method depends on the research objectives. A semi-structured interview is useful when the goal is to obtain specific information, while

a free-form feedback approach is preferred when the objective is to gather unbiased information and provide participants with the freedom to express their opinions without constraints.

Our assumption in this regard is that the success of interactive XR environments rests on the fact that the essential dimensions of the technology (e.g. video quality, latency, presence, accessibility) are solved. It is necessary to find a way to assess these aspects during XR interaction avoiding abrupt interruptions. Also, it is important to stop using questionnaires that are difficult to understand. Therefore, great effort should be made in A) designing short item questionnaires using clear and concise language, B) applicable in interactive sessions, and C) evaluating basic dimensions of the technology.

#### 7.4.2 Method of collecting ratings

The way to answer questionnaires in XR has not yet been standardized. In interactive XR experiences, it becomes even more important because it can influence on the data collected from participants. As Alexandrovsky et al. [202] conclude, there is no right or wrong solution, but researchers must consider that asking questionnaires in the XR environment may place additional mental demands on the participants.

Based on experience, questionnaires used during the communication (e.g., system aspects assessment: video quality or latency) must be filled with applications that allow users to score them within the XR environment (without removing their HMDs). It is motivated because it is typically a single item evaluated at different points throughout the test session. Also, we have found that asking a question in the XR environment every 25 seconds does not influence socioemotional aspects, such as social or spatial presence [54]. For this, it is recommended to use open-source tools that can be customized with the items that researchers are interested in [96].

Additionally, the evaluation method must be taken into account (e.g., oral, handheld controller or touchpad). It can influence users' behavior, limiting their exploration of the scene and other aspects such as attention or engagement. Gutiérrez et al. [37] concluded that rating orally could increase comfort and, therefore, the exploration of the scene. However, in more realistic scenarios where there is interaction, the use of the handheld controller is recommended as it is more natural than the touchpad located on the side of the HMD [54]. Based on the responses of the semi-structured interviews, participants concluded that the touchpad generated stress in them because pressing other buttons could interrupt the session.

Answering questionnaires in the physical reality disrupts the immersive experience [202]. Focusing on presence, Schwind et al. [108] reviewed validated questionnaires and concluded that there are no

differences between evaluating presence with a questionnaire in the physical reality or in the XR environment. This contribution is important since the use of long questionnaires in XR environments cause discomfort to users, especially in those with less experience using the technology.

If questionnaires outside the XR environment are used, forms are often evaluated through web-based applications. It implies considering the policies of the country to process and store the data.

## 7.5 Participants

The acceptance of a technology in society depends on the expectation of performance and the expectation of effort [42]. This expectation depends on the population sample that is using the technology. Most of the literature focuses on methodologies and experiments on specific XR applications and participants. This fact makes the reproducibility of the experiments difficult and, therefore, limits the generalization of the results. Other works report unbalanced samples of participants or do not even report gender, age or other differences. This fact can lead to a different level of acceptance of the technology, creating a gap in the use of one of the communication technologies of the future. Realistic environments that allow testing XR communications with a diverse part of the population are required. This generalization in the sample also has an implication in the type of tasks and questionnaires used, since accessibility must be addressed.

Researchers must characterize in detail the participants who perform the tests, taking into account their basic personal information as well as any additional information that may affect the experience. For example, if the context of the conversations is international experiences, it is interesting to consider the experience of the participants working or studying abroad [54]. Likewise, the assumptions based on works from the literature must also provide a correct characterization of the sample.

## 7.6 Conclusions

Here we present the best practices and guidelines in XR interactive experiments which stem from our own research and from a comprehensive review of the literature. These guidelines are provided to:

- tackle the assessment of the QoE to understand the influence of system restrictions (e.g., network bandwidth), while taking into account socioemotional factors.
- help researchers, developers, or service providers from different research areas.

- contribute to build more realistic, transparent, and inclusive evaluation environments, increasing the reliability of the conclusions and the acceptance of the technology.

**Source.** Source contents can be pre-recorded or streamed in real-time. Conditions are applied on the sources to obtain the stimuli rated by participants in the test. Therefore, a correct and detailed characterization, both objectively and semantically, of the source content is necessary to compare conditions and respond to the hypothesis.

- Select the **duration** of the source considering the number of conditions to test, to find a balance between test duration and cognitive load of the participants during the test.
- Compute and report SI, TI, and objective quality metric score and provide an **objective characterization** of the visual quality that highly influences QoE. Also, report **technical specifications**, such as resolution, framerate, bitrate, or encoder.
- Place the **camera**, which is the point of view of the user, at approximately the eye level of the participants to increase the naturalness of the experience. **Associate** objects in the **virtual** environment with the **real** one to increase the realism.
- Report abrupt **scene changes** during the session and consider their influence during the analysis if unexpected results appear.
- Do not forget that the **acquisition perspective (actor or observer)** highly influences social presence. The camera can be located at actor/first-person point of view or observer/third-person point of view.
- Take into account that the representation from remote users in the local environment influences the **interaction**. The fact of speaking to a 360-degree camera is not natural, so this aspect is a key piece during the design because it can influence on the effectiveness of the communication.
- Consider that the **type of conversation** can evoke different emotions and levels of non-verbal transmission.

**Condition.** Conditions express variations of the influencing factor considered in the hypothesis. They can be **continuous** (communication restrictions) or **categorical** (system decisions) conditions and both types can be combined in the same test.

- Apply **continuous conditions spatially** to analyze influencing factors around the XR scene.

- Choose the best option of **continuous conditions temporally** based on: number of **stimuli**, **cognitive load** for participants, and **social interactivity** during the test.
- Pre-select conditions to carry out **pilot studies** to focus the research on the most interesting ones.
- Combine categorical and continuous conditions in the same experiment paying attention to the **randomization** to assure reliable comparisons.
- Choose categorical conditions (e.g., video, interaction with the virtual environment, and self and others representations) according to the **tasks** and **use case** of the test.

**Feedback.** These guidelines are based on questionnaires asked at the beginning, during, or after the stimulus/test.

- Make the **overall quality** of experience the first question after the stimulus.
- Choose **brief, clear, and well-fitting** items to increase the accessibility.
- Include **semi-structured interviews** or **free-form feedback** to explain phenomena during the data analysis.
- Do not interrupt the communication, but ask **short items** in the **virtual environment** with the **handheld controller** to make it more natural.
- Use web platforms for questionnaires outside the XR environment and consider the **policy** of each country for **data treatment**.

**Participants.** They are users who test the technology.

- **Characterize participants** considering their basic personal information and additional information that may influence the results.
- Consider the **pre-experiment relationship** between the participants. It can influence the way they interact.
- Base the decisions on previous works that report **diverse participant samples**.
- Generalize the participants sample and tasks, guaranteeing maximum **accessibility** during the evaluation of XR as a communication technology.



# Chapter 8

## Conclusions and Future Work

### 8.1 Conclusions

This thesis presents an evaluation of immersive communications from the technical aspects to the socioemotional aspects. The research has followed an incremental line of research in which different conditions of the prototype and the experimental scenario have been modified from a reference configuration to understand the challenges of XR technologies in terms of QoE. The understood lessons can be applied to other XR technologies.

One of the goals of this thesis was to evaluate the quality of 360-degree video as a significant factor that impacts the QoE. Therefore, the use of the VMAF metric, one of the most robust objective video quality metrics, was validated for this specific type of content. Since 360-degree video quality still does not meet the expectations of end users, it is crucial to accurately measure it in order to improve encoding schemes. Regarding the subjective evaluation, the use of a methodology, SSDQE, has been validated. One of the main advantages of this methodology is that it can be used with long-duration content in which there is a narrative and context, unlike traditional methodologies, allowing for more realistic assessment scenarios that increase ecological validity. Furthermore, we have validated that this methodology allows the simultaneous evaluation of socioemotional and technical aspects. Therefore, it can be applied to evaluate other socioemotional and technical aspects and understand the interaction between them from the perspective of the end-users.

The immersive communication system has been explored mainly from the point of view of the remote user. First, we present the conclusions of the low-level decisions of the system. The interaction with the virtual environment has been compared between handheld controllers and with the HMD touchpad, resulting that the handheld controllers are more natural for the participants. Likewise, the possibility of visualizing their hands has been analyzed, not finding an influence on the presence or quality. However, the use case must be taken into account since in situations in which users have to perform a specific task with their hands (e.g., take notes) it will be essential that they can use them in a natural way. Second, we present the conclusions of the high-level decisions of the system. It is important to think about the position in which the camera is placed, as well as the height at eye level to increase the sense of presence. It has been observed that the ease of interruptions during

immersive communications, as occurs with other video call systems, is strongly influenced by the personality of the remote participant. However, the feedback provided by the local participants regarding the remote user is crucial for the QoE of the remote participant as it allows the local participants to interact with the capture element as if it were one of their peers in the conversation. It should be noted that in order to carry out the assessments, 360-degree videos have been captured and made publicly available in which the perspectives of the actor or first person or observer or third person have been considered, as well as the height of the camera, lighting, or positioning of the camera.

In this research, the use case of tele-education has been selected to analyze technology as a solution for students who cannot attend lessons in person, causing them problems of social distancing. For this, a video analysis module for event detection has been added to the immersive communication system. This module analyzes the 360-degree scene and notifies events of interest in the virtual environment of the remote users to help them follow the lesson remotely. The study of the notifications as an additional tool has shown a high acceptance by students. Both for the training of the video analysis module and for the subjective assessment, a database of 360-degree videos of real lessons from the Universidad Politécnica de Madrid has been generated and made public, in which people and events of interest have been annotated.

In relation to the use of questionnaires in research, the main disadvantages is the lack of control over participants' understanding and that, in general, they are not fully adapted to the specific conditions of the assessment. These limitations using questionnaires, that have been acknowledged in the literature and corroborated in this research, lead to an increased use of semi-structured interviews and free-form feedback to obtain more useful findings from the end user's perspective, also guiding future research steps.

As a practical contribution to the research community, we have summarized our current understanding of immersive communications and sharing best practices for designing experiments. These guidelines are intended to serve as a useful tool for developers and researchers interested in evaluating immersive communication systems and overcoming technical limitations. In addition, all guidelines are discussed from the point of view of performing experiments in more realistic scenarios to increase ecological validity and reproducibility. The analysis is based on a common framework for people of the community with different backgrounds and expertise level. Specifically, the elements addressed are: source, condition, feedback, and participants. We consider that having access to this information in a transparent and easily accessible manner could be beneficial for researchers in the early stages of their research as well as a powerful tool for people recently interested in subjective experimentation.

## 8.2 Future work

A natural step that we would like to explore is the generation of a battery of evaluation items of socioemotional and technical aspects obtained from methodologies already used in previous experiments. The motivation is to evaluate the aspects during the communication and not in post-questionnaires at the end of the communication where there may be influence of the memory effect or misunderstandings. From this, we would like to propose interactive communication sessions in which the questions appear while technical problems, such as compression or transmission, are introduced. The objective would be to go one step further in subjective experimentation, having an interactive communication, a battery of questions that appear during the session, and an experiment design that can be applied on a larger scale. At the end, the free-form feedback or semi-structured interview from participants could be collected. Based on this, we could explore the application of this type of assessment with several XR technologies, participant samples, and use cases. The goal of the research would be to directly compare the effectiveness of different XR technologies as a communication platform. This could include to understand how participants experience and interact from local and remote conditions and how this condition affects their ability to communicate. Overall, the aim of the research could be to promote the use of XR technologies as an effective technology for communications and to increase the ecological validity of the assessments.

Other future line of research could be to assess the accessibility of XR technologies, as well as identifying potential additional tools that may be necessary in order to increase accessibility for those who are currently excluded. As we explained at the beginning of this thesis, acceptance depends on the expectation of performance and effort. So far, we have focused our research on the potential of this technology compared to current communication technologies, but the remaining challenge is to investigate how to reduce this effort expectation for everyone.

The 360-degree video communications paradigm is in continuous development and evolution, leading to the emergence of new applications and use cases that present novel challenges to be addressed. This thesis has also been guided by this (r)evolution of technology and research community, requiring the acquisition of knowledge from multidisciplinary areas. Nevertheless, there is still a significant amount to be learned. Thus, a potential direction for future research is the further development of a comprehensive framework for evaluating and comparing different XR technologies to make them a solution to current social distancing issues.



# Appendix A

## Scientific Contributions

### JOURNALS

- (2023) M. Orduna, P. Pérez, J. Gutiérrez, N. García, “Best Practices for eXtended Reality Communications Assessment”. *IEEE Multimedia*, submitted.
- (2022) M. Orduna, P. Pérez, J. Gutiérrez, N. García, “Methodology to Assess Quality, Presence, Empathy, Attitude, and Attention in 360-degree Videos for Immersive Communications”. *IEEE Trans. on Affective Computing*.
- (2022) J. Gutiérrez, P. Pérez, M. Orduna, et al., “Subjective Evaluation of Visual Quality and Simulator Sickness of Short 360° Videos: ITU-T Rec. P.919”, *IEEE Trans. on Multimedia*.
- (2020) L. Muñoz, C. Díaz, M. Orduna, J.I. Ronda, P. Pérez, I. Benito, N. García. “Methodology for fine-grained monitoring of the quality perceived by users on 360VR contents”. *Digital Signal Processing*.
- (2020) M. Orduna, C. Díaz, L. Muñoz, P. Pérez, I. Benito, N. García. “Video Multimethod Assessment Fusion (VMAF) on 360VR contents”. *IEEE Trans. on Consumer Electronics*.

### CONFERENCES

- (2023) M. Orduna, S. Serino, P. Pérez, G. Riva, N. García, “QoE Assessment of Interactive Immersive Communications”, submitted.
- (2022) D. González, M. Orduna, C. Cortés, M. J. López Morales. “The Owl: An Accessible Immersive Telepresence System for the Future of Human Communication”, *IEEE Globecom*, Rio de Janeiro, Brasil.  
*First prize at IEEE Communications Society Student Competition.*
- (2022) M. Orduna, “Quality, Presence, Empathy, Attitude, and Attention in 360-degree Videos for Immersive Communications”. *ACM CHI*, New Orleans, LA, USA.  
*ACM CHI Doctoral Consortium Award.*

- (2022) H. Elmimouni, J.P. Hansen, S.C. Herring, J. Marcin, M. Orduna, P. Pérez, I. Rae, J.C. Read, J. Rode, S. Sabanovic, V. Ahumada, “Emerging Telepresence Technologies in Hybrid Learning Environments”. ACM CHI, New Orleans (LA), USA.
- (2022) C. Cortés, M. Orduna, P. Pérez, N. García. “Natural Collaborative interfaces for XR immersive learning”, ACM IMX, Aveiro, Portugal.
- (2022) M. Orduna, J. Gutiérrez, A. Sánchez, J. Cabrera, C. Díaz, P. Pérez, N. García, “Evaluation of the Performance of an Immersive System for Tele-education”, ACM IMX, Aveiro, Portugal.
- (2021) R. Kachach, M. Orduna, J. Rodríguez, P. Pérez, A. Villegas, J. Cabrera, N. Garcia, “Immersive Telepresence in Remote Education”, ACM MMSys-MMVE, Istanbul, Turkey.
- (2021) M. Orduna, J. Gutiérrez, C. Manzano, D. Ruiz, J. Cabrera, C. Díaz, P. Pérez, N. García, “EVENT-CLASS: Dataset of events in the classroom”, QoMEX, Montreal, Canada.
- (2020) M. Orduna, P. Pérez, C. Díaz, N. García, “Evaluating the Influence of the HMD, Usability, and Fatigue in 360VR Video Quality Assessments”, IEEE VR, Atlanta (GA), USA. *IEEE VR Diversity Award.*
- (2020) M. Orduna. “Quality, Presence, and Emotions in Virtual Reality Communications”, IEEE VR, Atlanta (GA), USA.

## CONTRIBUTIONS TO STANDARDS

- (2022) M. Orduna, P. Pérez, J. Gutiérrez, N. García, “Comparing ACR, SSDQE, and SSCQE in long duration 360-degree videos”, VQEG contribution.
- (2021) M. Orduna, J. Gutiérrez, P. Pérez, N. García, “Methodology to Assess Quality, Presence, Empathy, Attitude, and Attention in Social VR: International Experiences Use Case”, VQEG contribution.
- (2020) P. Pérez, J. Gutiérrez, A. Singla, I. Viola, F. Battisti, D. Juszka, M. Orduna, Z. Chen, Y. Hu, “Draft baseline for ITU-T P.360-VR Subjective Test Methodologies for 360-Degree Video on HMD”, ITU-T Study Group 12 meeting contribution.
- (2020) F. Adeyemi-Ejeye, F. Battisti, K. Brunnström, M. Carli, P. César, Z. Chen, N. Cieplińska, C. Cortés, C. Díaz, S. Fremerey, N. García, J. Gutiérrez, O. Hamsis, J. Hedlund, F. Hofmeyer, Y. Hu, L. Janowski, D. Juszka, P. Lambert, M. Leszczuk, P. Mazumdar, M. Orduna, P. Pérez, A. Raake, A. Singla, G. Van Wallendael, I. Viola, “IMG Test Phase 1 - Short Sequences: Results and Outcomes”. VQEG contribution.

- (2020) F. Adeyemi-Ejeye, F. Battisti, K. Brunnström, M. Carli, P. César, Z. Chen, N. Cieplińska, C. Cortés, C. Díaz, S. Fremerey, N. García, J. Gutiérrez, J. Hedlund, F. Hofmeyer, Y. Hu, L. Janowski, D. Juszka, P. Lambert, M. Leszczuk, P. Mazumdar, M. Orduna, P. Pérez, A. Raake, A. Singla, G. Van Wallendael, I. Viola, “VQEG Test Plan for Quality Assessment of 360-degree Video. Phase 1: Short sequences”. MPEG 131th meeting contribution.
- (2019) M. Orduna, C. Cortés, P. Pérez, N. García, “IMG Work plan: Pre-test discussion - UPM tests”. VQEG contribution.
- (2019) M. Orduna, C. Díaz, L. Muñoz, P. Pérez, N. García, “Quality Metrics for Immersive 360VR Content”. VQEG contribution.
- (2018) M. Orduna, C. Díaz, L. Muñoz, P. Pérez, I. Benito, N. García, “Video Multimethod Assessment Fusion (VMAF) on 360VR”. VQEG contribution.

#### **PUBLIC AVAILABLE DATA**

- Supplemental material of Evaluating the Influence of the HMD, Usability, and Fatigue in 360VR Video Quality Assessments: <https://www.gti.ssr.upm.es/data/360VR>
- EVENT-CLASS database: <https://www.gti.ssr.upm.es/data/event-class>
- Student Experiences Around the World-dataset (SEAW-dataset): <https://www.gti.ssr.upm.es/data/seaw-dataset>



# Appendix B

## Outreach Activity

### Press articles

- (2022) M. Orduna, P. Pérez, J. Gutiérrez, N. García. “¿Cómo influye el análisis de la interacción en el éxito del metaverso?”. Madri+d, find it [here](#).
- (2021) M. Orduna, P. Pérez, N. García. “Educative telepresence based on distributed reality: a solution for tele-education”. BIT, find it [here](#)

### Audiovisual media

- (2022) M. Orduna Cortillas. “Quality, Presence, and Emotions in Virtual Reality Communications”. Doctoral Symposium Universidad Politécnica de Madrid - My thesis in a nutshell, find it [here](#).  
*Finalist award.*
- (2022) M. Orduna. #somoscientificxs - Initiative to promote science between High School students.
- (2021) M. Orduna Cortillas. “Comunicaciones inmersivas”. Cuéntame11F - Iniciativa 11 de Febrero, find it [here](#).  
*Second prize.*
- (2021) M. Orduna Cortillas. “Quality, Presence, and Emotions in Virtual Reality Communications”. Doctoral Symposium Universidad Politécnica de Madrid - My thesis in a nutshell, find it [here](#).  
*Finalist award.*
- (2021) M. Orduna. Your thesis in a twitter thread: #hilotesis. RedDivulga-CRUE, find it [here](#).

## Talks

- **(2022, 2023)** M. Orduna. Selected researcher at #Escaparates11F organized by 11defebrero and MadeinZGZ. Initiative to promote women in science in Zaragoza.
- **(2022)** M. Orduna and E. Fitchmann. “From the lab to the market”. Digital Summit organized by DigitalES, Spanish Association for Digitalization, find it [here](#).
- **(2021,2022)** M. Orduna. “Primeros pasos en investigación. Realidad Virtual y Comunicaciones inmersivas” . ICT Professional Environment, Universidad Alcalá de Henares (UAH).
- **(2021)** M. Orduna Cortillas. “Entrevista a Marta Orduna Cortillas”. #voz11F - Podcast de ciencia y divulgación en femenino, find it [here](#).

# Bibliography

- [1] Richard Skarbez, Missie Smith, and Mary C Whitton. Revisiting Milgram and Kishino’s Reality-Virtuality Continuum. *Frontiers in Virtual Reality*, 2:647997, March 2021.
- [2] Grupo de Tratamiento de Imágenes. SEAW – DATASET. <https://www.gti.ssr.upm.es/data/seaw-dataset>, Accessed: 2023-02-18.
- [3] Leslie Neely, Amarie Carnett, John Quarles, Hannah MacNaul, Se-Woong Park, Sakiko Oyama, Guenevere Qian Chen, Kevin Desai, and Peyman Najafirad. The Case for Integrated Advanced Technology in Applied Behavior Analysis. *Advances in Neurodevelopmental Disorders*, pages 1–11, December 2022.
- [4] Redouane Kachach, Marta Orduna, Jesús Rodríguez, Pablo Pérez, Álvaro Villegas, Julián Cabrera, and Narciso García. Immersive Telepresence in Remote Education. In *ACM International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE)*, page 21–24, Istanbul, Turkey, 2021.
- [5] Zhenbo Li, Jun Yue, and David Antonio Gómez Jáuregui. A New Virtual Reality Environment used for E-learning. In *IEEE International Symposium on IT in Medicine & Education (ITME)*, volume 1, pages 445–449, Jinan, China, 2009.
- [6] Teresa Monahan, Gavin McArdle, and Michela Bertolotto. Virtual Reality for Collaborative E-learning. *Computers & Education*, 50(4):1339–1353, May 2008.
- [7] Daniel Roth, Kevin Yu, Frieder Pankratz, Gleb Gorbachev, Andreas Keller, Marc Lazarovici, Dirk Wilhelm, Simon Weidert, Nassir Navab, and Ulrich Eck. Real-time Mixed Reality Teleconsultation for Intensive Care Units in Pandemic Situations. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 693–694, Lisbon, Portugal, 2021.
- [8] Pablo Pérez, Elena Vallejo, Marta Revuelta, María Victoria Redondo Vega, Esther Guervós Sánchez, and Jaime Ruiz. Immersive Music Therapy for Elderly Patients. In *ACM International Conference on Interactive Media Experiences (IMX)*, page 47–52, Aveiro, JB, Portugal, 2022.
- [9] Cedric Di Loreto, Jean Remy Chardonnet, Julien Ryard, and Alain Rousseau. WoaH: A Virtual Reality Work-At-Height Simulator. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 281–288, Tuebingen/Reutlingen, Germany, 2018.
- [10] Mustufa Haider Abidi, Abdulrahman Al-Ahmari, Ali Ahmad, Wadea Ameen, and Hisham Alkhalefah. Assessment of Virtual Reality-based Manufacturing Assembly Training System. *International Journal of Advanced Manufacturing Technology*, 105(9):3743–3759, December 2019.
- [11] Mehdi Hafsia, Eric Monacelli, and Hugo Martin. Virtual Reality Simulator for Construction Workers. In *ACM International Conference Proceeding Series (VRIC)*, Laval, France, 2018.

- [12] Hadj Sassi Mohamed Saifeddine, Battisti Federica, and Carli Marco. Simulation-based Virtual Reality Training for Firefighters. *Electronic Imaging*, 34:1–5, January 2022.
- [13] Ignacio Reimat, Yanni Mei, Evangelos Alexiou, Jack Jansen, Jie Li, Shishir Subramanyam, Irene Viola, Johan Oomen, and Pablo Cesar. Mediascape XR: A Cultural Heritage Experience in Social VR. In *ACM International Conference on Multimedia (MM)*, pages 6955–6957, 2022.
- [14] Pietro Cipresso, Irene Alice Chicchi Giglioli, Mariano Alcañiz Raya, and Giuseppe Riva. The Past, Present, and Future of Virtual and Augmented Reality Research: a Network and Cluster Analysis of the Literature. *Frontiers in Psychology*, 9:2086, November 2018.
- [15] Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal Communication in Human Interaction*. Cengage Learning, 2013.
- [16] Divine Maloney, Guo Freeman, and Donghee Yvette Wohn. "Talking without a Voice": Understanding Non-verbal Communication in Social Virtual Reality. *ACM on Human-Computer Interaction*, 4(CSCW2):1–25, October 2020.
- [17] Thammathip Piumsomboon, Gun A Lee, Andrew Irlitti, Barrett Ens, Bruce H Thomas, and Mark Billingham. On the Shoulder of the Giant: A Multi-Scale Mixed Reality Collaboration with 360 Video Sharing and Tangible Interaction. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–17, Glasgow, Scotland Uk, 2019.
- [18] Simon Gunkel, Hans Stokking, Martin Prins, Omar Niamut, Ernestasia Siahaan, and Pablo Cesar. Experiencing Virtual Reality Together: Social VR Use Case Study. In *ACM International Conference on Interactive Experiences for TV and Online Video (TVX)*, pages 233–238, SEOUL, Republic of Korea, 2018.
- [19] Lara Muñoz, César Díaz, Marta Orduna, José Ignacio Ronda, Pablo Pérez, Ignacio Benito, and Narciso García. Methodology for Fine-Grained Monitoring of the Quality Perceived by Users on 360VR Contents. *Digital Signal Processing*, 100:102706, May 2020.
- [20] Miska M Hannuksela, Ye-Kui Wang, and Ari Hourunranta. An Overview of the OMAF Standard for 360 Video. In *IEEE Data compression conference (DCC)*, pages 418–427, Snowbird, UT, USA, 2019.
- [21] ITU-T Recommendation H.265. H.265: High Efficiency Video Coding. August 2021.
- [22] ITU-T Recommendation H.264. H.264: Advanced Video Coding for Generic Audiovisual Services. October 2021.
- [23] Carlos Cortés, Pablo Pérez, Jesús Gutiérrez, and Narciso García. Influence of Video Delay on Quality, Presence, and Sickness in Viewport Adaptive Immersive Streaming. In *International Conference on Quality of Multimedia Experience (QoMEX)*, 2020.
- [24] Shyamprasad Chikkerur, Vijay Sundaram, Martin Reisslein, and Lina J Karam. Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison. *IEEE Transactions on Broadcasting*, 57(2):165–182, June 2011.

- [25] ITU-R Recommendation BT.500-14. Methodology for the Subjective Assessment of the Quality of Television Pictures. October 2019.
- [26] ITU-T Recommendation P.800. Methods for Subjective Determination of Transmission Quality. November 2019.
- [27] ITU-T Recommendation 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. June 2021.
- [28] ITU-T Recommendation P.919. Subjective Test Methodologies for 360° Video on Head-Mounted Displays. October 2020.
- [29] Margaret H Pinson, Lucjan Janowski, Romuald Pepion, Quan Huynh-Thu, Christian Schmidmer, Phillip Corriveau, Audrey Younkin, Patrick Le Callet, Marcus Barkowsky, and William Ingram. The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):640–651, August 2012.
- [30] Janto Skowronek, Alexander Raake, Gunilla Berndtsson, Olli S Rummukainen, Paolino Usai, Simon NB Gunkel, Mathias Johanson, Emanuel AP Habets, Ludovic Malfait, David Lindero, and et al. Quality of Experience in Telemeetings and Videoconferencing: A Comprehensive Survey. *IEEE Access*, 10:63885 – 63931, May 2022.
- [31] ITU-T Recommendation P.10/G.100. Vocabulary for Performance, Quality of Service and Quality of Experience. October 2020.
- [32] Patrick Le Callet, Sebastian Moller, and Andrew Perkis. Qualinet White Paper on Definitions of Quality of Experience. *Proc. Output 5th Qualinet Meeting*, March 2013.
- [33] Andrew Perkis, Christian Timmerer, Sabina Barakovic, Jasmina Barakovic Husic, Soren Bech, Sebastian Bosse, Jean Botev, Kjell Brunnstrom, Luis Cruz, Katrien De Moor, and et al. QUALINET White Paper on Definitions of Immersive Media Experience (IMEx), 2020. arXiv:2007.07032.
- [34] ITU-T Recommendation P.1320. QoE Assessment of Extended Reality (XR) Meetings. Jul 2022.
- [35] VQEG is Co-Chaired by: Margaret Pinson, NTIA/ITS and Kjell Brunnstrom, RISE Research Institute of Sweden AB. Video Quality Experts Group (VQEG). <https://vqeg.org/vqeg-home/>, 2022. Accessed: 2023-02-18.
- [36] Recommendation ITU-T P.910. Subjective Video Quality Assessment Methods for Multimedia Applications. April 2008.
- [37] Jesus Gutierrez, Pablo Perez, Marta Orduna, Ashutosh Singla, Carlos Cortes, Pramit Mazumdar, Irene Viola, Kjell Brunnstrom, Federica Battisti, Natalia Cieplinska, Dawid Juszka, Lucjan Janowski, Mikoaj Igor Leszczuk, Anthony Adeyemi-Ejeye, Yaosi Hu, Zhengzhong Chen, Glenn Van Wallendael, Peter Lambert, Cesar Diaz, John Hedlund, Omar Hamsis, Stephan

- Fremerey, Frank Hofmeyer, Alexander Raake, Pablo Cesar, Marco Carli, and Narciso García. Subjective Evaluation of Visual Quality and Simulator Sickness of Short 360 Videos: ITU-T Rec. P.919. *IEEE Transactions on Multimedia*, 24:3087–3100, June 2022.
- [38] Jesús Gutiérrez, Pablo Pérez, Fernando Jaureguizar, Julián Cabrera, and Narciso García. Validation of a Novel Approach to Subjective Quality Evaluation of Conventional and 3D Broadcasted Video Services. In *International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 230–235, Melbourne, VIC, Australia, 2012.
- [39] Mikko Salminen, Simo Järvelä, Antti Ruonala, Ville Harjunen, Giulio Jacucci, Juho Hamari, and Niklas Ravaja. Evoking Physiological Synchrony and Empathy Using Social VR with Biofeedback. *IEEE Transactions on Affective Computing*, 13(2):746–755, April 2022.
- [40] Giuseppe Riva and Fabrizia Mantovani. From the Body to the Tools and Back: a General Framework for Presence in Mediated Interactions. *Interacting with Computers*, 24(4):203–210, July 2012.
- [41] Mark H. Davis. Measuring Individual Differences in Empathy: Evidence for a Multidimensional Approach. *Journal of Personality and Social Psychology*, 44(1):113–126, July 1983.
- [42] Richard J Holden and Ben-Tzion Karsh. The Technology Acceptance Model: its Past and its Future in Health Care. *Journal of biomedical informatics*, 43(1):159–172, February 2010.
- [43] Andrew MacQuarrie and Anthony Steed. Cinematic Virtual Reality: Evaluating the Effect of Display Type on the Viewing Experience for Panoramic Video. In *IEEE Virtual Reality*, pages 45–54, Los Angeles, CA, USA, 2017.
- [44] Diana Fonseca and Martin Kraus. A Comparison of Head-Mounted and Hand-Held Displays for 360 Videos with Focus on Attitude and Behavior Change. In *International Academic Mindtrek Conference*, pages 287–296, Tampere, Finland, 2016.
- [45] Silvia Serino and Claudia Repetto. New Trends in Episodic Memory Assessment: Immersive 360 Ecological Videos. *Frontiers in psychology*, 9:1878, October 2018.
- [46] Lingwei Tong, Sungchul Jung, and Robert W Lindeman. Action Units: Directing User Attention in 360-degree Video based VR. In *ACM Symposium on Virtual Reality Software and Technology*, pages 1–2, 2019.
- [47] Natalia Cieplińska, Lucjan Janowski, Katrien De Moor, and Michał Wierzchoń. Long-Term Video QoE Assessment Studies: A Systematic Review. *IEEE Access*, 10:133883–133897, December 2022.
- [48] John F Kihlstrom. Ecological Validity and “Ecological Validity”. *Perspectives on Psychological Science*, 16(2):466–471, February 2021.
- [49] Kevin Yu, Gleb Gorbachev, Ulrich Eck, Frieder Pankratz, Nassir Navab, and Daniel Roth. Avatars for Teleconsultation: Effects of Avatar Embodiment Techniques on User Perception in 3D Asymmetric Telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 27(11):4129–4139, August 2021.

- [50] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Krishna Moorthy, and M. Manohara. Toward A Practical Perceptual Video Quality Metric. <https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652>, Accessed: 2023-02-18.
- [51] Zhi Li, Kyle Swanson, Christos Bampis, Lukáš Krasula, and Anne Aaron. Toward a Better Quality Metric for the Video Community. <https://netflixtechblog.com/toward-a-better-quality-metric-for-the-video-community-7ed94e752a30>, Accessed: 2023-02-18.
- [52] Marta Orduna, César Díaz, Lara Muñoz, Pablo Pérez, Ignacio Benito, and Narciso García. Video Multimethod Assessment Fusion (VMAF) on 360VR Contents. *IEEE Transactions on Consumer Electronics*, 66(1):22–31, February 2020.
- [53] Marta Orduna, Pablo Pérez, César Díaz, and Narciso García. Evaluating the Influence of the HMD, Usability, and Fatigue in 360VR Video Quality Assessments. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 683–684, Atlanta, GA, USA, 2020.
- [54] Marta Orduna, Pablo Pérez, Jesús Gutiérrez, and Narciso García. Methodology to Assess Quality, Presence, Empathy, Attitude, and Attention in 360-degree Videos for Immersive Communications. *IEEE Transactions on Affective Computing*, February 2022. Early Access.
- [55] Marta Orduna, Silvia Serino, Pablo Pérez, Giuseppe Riva, and Narciso García. QoE Assessment of Interactive Immersive Communications. 2023. submitted.
- [56] Marta Orduna, Jesús Gutiérrez, Carlos Manzano, David Ruiz, Julián Cabrera, Pablo Pérez, and Narciso García. EVENT-CLASS: Dataset of Events in the Classroom. In *International Conference on Quality of Multimedia Experience (QoMEX)*, pages 81–84, Montreal, QC, Canada, 2021.
- [57] Houda Elmimouni, John Paulin Paulin Hansen, Susan Herring, James Marcin, Marta Orduna, Pablo Pérez, Irene Rae, Janet Read, Jennifer Rode, Selma Sabanovic, and Verónica Ahumada. Emerging Telepresence Technologies in Hybrid Learning Environments. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, New Orleans, LA, USA, 2022.
- [58] Marta Orduna, Jesús Gutiérrez, Alejandro Sánchez, Julián Cabrera, César Díaz, Pablo Pérez, and Narciso García. Evaluation of the Performance of an Immersive System for Tele-Education. In *ACM International Conference on Interactive Media Experiences (IMX)*, page 209–220, Aveiro, JB, Portugal, 2022.
- [59] Marta Orduna, Pablo Pérez, Jesús Gutiérrez, and Narciso García. Best Practices for eXtended Reality Communications Assessment. *IEEE MultiMedia*, December 2022. submitted.
- [60] Peter Willemsen, Mark B Colton, Sarah H Creem-Regehr, and William B Thompson. The Effects of Head-mounted Display Mechanical Properties and Field of View on Distance Judgments in Virtual Environments. *ACM Transactions on Applied Perception (TAP)*, 6(2): 8:1–8:14, March 2009.

- [61] Erwan J David, Jesús Gutiérrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. A Dataset of Head and Eye Movements for 360 Videos. In *ACM Multimedia Systems Conference (MMSys)*, pages 432–437, Amsterdam, Netherlands, 2018.
- [62] Prमित Mazumdar and Federica Battisti. A Content-based Approach for Saliency Estimation in 360 Images. In *IEEE International Conference on Image Processing (ICIP)*, pages 3197–3201, Taipei, Taiwan, 2019.
- [63] Jeroen Van der Hooft, Maria Torres Vega, Stefano Petrangeli, Tim Wauters, and Filip De Turck. Tile-based Adaptive Streaming for Virtual Reality Video. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(4):1–24, December 2019.
- [64] Demóstenes Z Rodríguez, Renata L Rosa, Eduardo A Costa, Julia Abrahão, and Graca Bressan. Video Quality Assessment in Video Streaming Services Considering User Preference for Video Content. *IEEE Transactions on Consumer Electronics*, 60(3):436–444, August 2014.
- [65] Roberto G de A Azevedo, Neil Birkbeck, Francesca De Simone, Ivan Janatra, Balu Adsumilli, and Pascal Frossard. Visual Distortions in 360° Videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2524–2537, August 2020.
- [66] Sohee Park, Arani Bhattacharya, Zhibo Yang, Samir R. Das, and Dimitris Samaras. Mosaic: Advancing User Quality of Experience in 360-Degree Video Streaming With Machine Learning. *IEEE Transactions on Network and Service Management*, 18(1):1000–1015, January 2021.
- [67] Mai Xu, Lai Jiang, Chen Li, Zulin Wang, and Xiaoming Tao. Viewport-based CNN: A Multi-task Approach for Assessing 360 Video Quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2198–2215, October 2020.
- [68] H. T. T. Tran, C. T. Pham, N. Pham Ngoc, A. T. Pham, and TC. Thang. A Study on Quality Metrics for 360 Video Communications. *IEICE Transactions on Information and Systems*, 101(1):28–36, January 2018.
- [69] Matt Yu, Haricharan Lakshman, and Bernd Girod. A Framework to Evaluate Omnidirectional Video Coding Schemes. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 31–36, Fukuoka, Japan, 2015.
- [70] Yule Sun, Ang Lu, and Lu Yu. Weighted-to-Spherically-Uniform Quality Evaluation for Omnidirectional Video. *IEEE Signal Processing Letters*, 24(9):1408–1412, September 2017.
- [71] Vladyslav Zakharchenko, Kwang Pyo C., and Jeong H. P. Hoon. Quality Metric for Spherical Panoramic Video. In *Optics and Photonics for Information Processing X*, volume 9970, pages 57–65, 2016.
- [72] Ramin Ghaznavi. 360-Degree Panoramic Video Coding. *Master Thesis - Faculty of Computing and Electrical Engineering, Tampere University of Technology*, 2016.

- [73] Deepti Pappusetty, Hari Kalva, and Howard S. Hock. Pupil Response to Quality and Content Transitions in Videos. *IEEE Transactions on Consumer Electronics*, 63(4):410–418, November 2017.
- [74] Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale Structural Similarity for Image Quality Assessment. In *IEEE Asilomar Conference on Signals, Systems & Computers*, volume 2, pages 1398–1402, Pacific Grove, CA, USA, 2003.
- [75] Zhi Li, Christos George Bampis, A. A. Novak, K. Swanson, Anush Krishna Moorthy, and J. De Cock. VMAF: The Journey Continues. <https://medium.com/netflix-techblog/vmaf-the-journey-continues-44b51ee9ed12>, Accessed: 2023-02-18.
- [76] Reza Rassool. VMAF Reproducibility: Validating a Perceptual Practical Video Quality Metric. In *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–2, Cagliari, Italy, 2017.
- [77] Nabajeet Barman, Steven Schmidt, Saman Zadtootaghaj, Maria G. Martini, and Sebastian Möller. An Evaluation of Video Quality Assessment Metrics for Passive Gaming Video Streaming. In *ACM Proceedings of the Packet Video Workshop (PV)*, pages 7–12, Amsterdam, Netherlands, 2018.
- [78] C. Lee, S. Woo, S. Baek, J. Han, J. Chae, and J. Rim. Comparison of Objective Quality Models for Adaptive Bit-Streaming Services. In *IEEE Conference on Information, Intelligence, Systems & Applications (IISA)*, pages 1–4, Larnaca, Cyprus, 2017.
- [79] Christos George Bampis and Alan C. Bovik. Learning to Predict Streaming Video QoE: Distortions, Rebuffering and Memory. *arXiv preprint arXiv:1703.00633*, 2017.
- [80] Christos G. Bampis, Zhi Li, and Alan C. Bovik. Enhancing Temporal Quality Measurements in a Globally Deployed Streaming Video Quality Predictor. In *IEEE International Conference on Image Process. (ICIP)*, pages 614–618, Athens, Greece, 2018.
- [81] Yashas Rai, Jesús Gutiérrez, and Patrick Le Callet. A Dataset of Head and Eye Movements for 360 Degree Images. In *ACM on Multimedia Systems Conference (MMSys)*, pages 205–210, Taipei, Taiwan, 2017.
- [82] Thomas Maugey, Olivier Le Meur, and Zhi Liu. Saliency-based Navigation in Omnidirectional Image. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, Luton, UK, 2017.
- [83] Kashyap Kammachi Sreedhar, Alireza Aminlou, Miska M. Hannuksela, and Moncef Gabbouj. Viewport-Adaptive Encoding and Streaming of 360-Degree Video for Virtual Reality Applications. In *IEEE International Symposium on Multimedia (ISM)*, pages 583–586, San Jose, CA, USA, 2016.
- [84] El-Ganainy Tarek and Hefeeda Mohamed. Streaming Virtual Reality Content. *arXiv preprint arXiv:1612.08350*, December 2016.

- [85] Hanli Wang and Sam Kwong. Rate-Distortion Optimization of Rate Control for H.264 with Adaptive Initial Quantization Parameter Determination. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(1):140–144, January 2008.
- [86] John D. McCarthy, M. Angela Sasse, and Dimitrios Miras. Sharp or Smooth?: Comparing the Effects of Quantization vs. Frame Rate for Streamed Video. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 535–542, Vienna, Austria, 2004.
- [87] Haiqiang Wang, Weihao Gan, Sudeng Hu, Joe Yuchieh Lin, Lina Jin, Longguang Song, Ping Wang, Ioannis Katsavounidis, Anne Aaron, and C-C Jay Kuo. MCL-JCV: a JND-based H.264/AVC Video Quality Assessment Dataset. In *IEEE International Conference on Image Processing (ICIP)*, pages 1509–1513, Phoenix, AZ, USA, 2016.
- [88] Virtual Human Interaction Lab. Stanford University. A Public Database of 360 Videos with Corresponding Ratings of Arousal and Valence. <http://vhil.stanford.edu/360-video-database/>, Accessed: 2023-02-18.
- [89] Chenglei Wu, Zhihao Tan, Zhi Wang, and Shiqiang Yang. A Dataset for Exploring User Behaviors in VR Spherical Video Streaming. In *ACM on Multimedia Systems Conference (MMSys)*, pages 193–198, Taipei, Taiwan, 2017.
- [90] Ashutosh Singla, Stephan Fremerey, Werner Robitza, Pierre Lebreton, and Alexander Raake. Comparison of Subjective Quality Evaluation for HEVC Encoded Omnidirectional Videos at Different Bit-rates for UHD and FHD Resolution. In *Thematic Workshops of ACM Multimedia (MM)*, pages 511–519, Mountain View, California, USA, 2017.
- [91] Ashutosh Singla, Werner Robitza, and Alexander Raake. Comparison of Subjective Quality Evaluation Methods for Omnidirectional Videos with DSIS and Modified ACR. *Electronic Imaging*, (14):1–6, January 2018.
- [92] Bo Zhang, Junzhe Zhao, Shu Yang, Yang Zhang, Jing Wang, and Zesong Fei. Subjective and Objective Quality Assessment of Panoramic Videos in Virtual Reality Environments. In *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 163–168, Hong Kong, China, 2017.
- [93] Ashutosh Singla, Stephan Fremerey, Werner Robitza, and Alexander Raake. Measuring and Comparing QoE and Simulator Sickness of Omnidirectional Videos in Different Head Mounted Displays. In *IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, Erfurt, Germany, 2017.
- [94] Hak Gu Kim, Heoun-Taek Lim, Sangmin Lee, and Yong Man Ro. VRSA Net: VR Sickness Assessment Considering Exceptional Motion for 360 VR Video. In *IEEE Transactions on Image Processing*, volume 28, pages 1646–1660, 2018.
- [95] Netflix. VMAF - Video Multi-Method Assessment Fusion. <https://github.com/Netflix/vmaf>, Accessed: 2023-02-18.

- [96] Carlos Cortés, Pablo Pérez, and Narciso García. Unity3D-based App for 360VR Subjective Quality Assessment with Customizable Questionnaires. In *IEEE International Conference on Consumer Electronics*, pages 281–282, Berlin, Germany, 2019.
- [97] European Union. General Data Protection Regulation 2016/679, 2016, Accessed: 2023-02-18. URL <http://data.europa.eu/eli/reg/2016/679/oj>.
- [98] Jan Ozer. Finding the Just Noticeable Difference with Netflix VMAF. <https://www.linkedin.com/pulse/finding-just-noticeable-difference-netflix-vmf-jan-oz/>, Accessed: 2023-02-18.
- [99] A Colin Cameron and Frank AG Windmeijer. An R-squared Measure of Goodness of Fit for Some Common Nonlinear Regression Models. *Elsevier Journal of Econometrics*, 77(2): 329–342, April 1997.
- [100] Mel Slater and Sylvia Wilbur. A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 6(6):603–616, December 1997.
- [101] J. Thompson, A. K. Karembai, and P. Seeling. Immersive Image QoE in Mobile Consumer Virtual Reality Settings. In *IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pages 1–4, Las Vegas, NV, USA, 2019.
- [102] Tanja Aitamurto, Shuo Zhou, Sukolsak Sakshuwong, Jorge Saldivar, Yasamin Sadeghi, and Amy Tran. Sense of Presence, Attitude Change, Perspective-Taking and Usability in First-Person Split-Sphere 360 Video. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–12, Montreal QC, Canada, 2018.
- [103] Audrey Tse, Charlene Jennett, Joanne Moore, Zillah Watson, Jacob Rigby, and Anna L. Cox. Was I There?: Impact of Platform and Headphones on 360 Video Immersion. In *Conference Extended Abstracts on Human Factors in Computing Systems (CHI)*, pages 2967–2974, Denver, Colorado, USA, 2017.
- [104] Matthew Lombard, Theresa B. Ditton, and Lisa Weinstein. Measuring Presence: the Temple Presence Inventory. In *Annual International Workshop on Presence*, pages 1–15, 2009.
- [105] Bob G. Witmer and Michael J. Singer. Measuring Presence in Virtual Environments: A Presence Questionnaire. 7(3):225–240, June 1998.
- [106] Holger T. Regenbrecht, Thomas W. Schubert, and Frank Friedmann. Measuring the sense of presence and its relations to fear of heights in virtual environments. *International Journal of Human-Computer Interaction*, 10(3):233–249, November 1998.
- [107] Mel Slater. Measuring presence: A response to the witmer and singer presence questionnaire. *Presence: teleoperators and virtual environments*, 8(5):560–565, December 1999.
- [108] Valentin Schwind, Pascal Knierim, Nico Haas, and Niels Henze. Using Presence Questionnaires in Virtual Reality. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–12, Glasgow, Scotland, 2019.

- [109] Desarrollo y DiseñoLab RTVE.es. ALENTO, Escena 360°. <https://lab.rtve.es/escena-360/alento/360/>, Accessed: 2023-02-18.
- [110] AirPano VR. 360°, Angel Falls, Venezuela. Part I. Aerial 8K video. <https://www.youtube.com/watch?v=8rUwdtERUOM>, Accessed: 2023-02-18.
- [111] Sara Martin Flamenco and Nokia. CLASE DE FLAMENCO EN 360. <https://www.youtube.com/watch?v=ZSPnFxDj7gg>, Accessed: 2023-02-18.
- [112] Disney on Broadway. Circle of Life in 360° - THE LION KING on Broadway. <https://www.youtube.com/watch?v=7T57kzGQGto&t=12s>, Accessed: 19-July-2008.
- [113] National Geographic. Lions 360° - National Geographic. <https://www.youtube.com/watch?v=sPyAQQklc1s>, Accessed: 2023-02-18.
- [114] Blick. 360° cockpit view - Fighter Jet - Patrouille Suisse - Virtual Reality. <https://www.youtube.com/watch?v=NdZ02-Qenso>, Accessed: 2023-02-18.
- [115] JCT-VC and ISO/IEC JTC1/SC29/WG11. Common HM Test Conditions and Software Reference Configurations. Output doc. M27343, 102th MPEG Meeting, Shanghai, China, October 2012.
- [116] Robert C. Streijl, Stefan Winkler, and David S. Hands. Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives. *Multimedia Systems*, 22(2):213–227, March 2016.
- [117] Manish Narwaria, Lukáš Krasula, and Patrick Le Callet. Data Analysis in Multimedia Quality Assessment: Revisiting the Statistical Tests. *IEEE Transactions on Multimedia*, 20(8):2063–2072, 2018.
- [118] Grupo de Tratamiento de Imágenes. Evaluating the Influence of the HMD, Usability, and Fatigue in 360VR Video Quality Assessments - Supplemental material. <https://www.gti.ssr.upm.es/data/360VR>, Accessed: 2023-02-18.
- [119] Paul Pürcher and Margit Höfler. Technology Meets Psychology: Psychological Background in Virtual Realities. In *International Convention on Information and Communication Technology, Electronics and Microelectronics*, pages 633–637, Opatija, Croatia, 2018.
- [120] Ashutosh Singla, Stephan Fremerey, Frank Hofmeyer, Werner Robitza, and Alexander Raake. Quality Assessment Protocols for Omnidirectional Video Quality Evaluation. *Electronic Imaging*, 2020(11):69–1–69–7, January 2020.
- [121] M-N García, F. De Simone, S. Tavakoli, N. Staelens, S. Egger, K. Brunnström, and A. Raake. Quality of Experience and HTTP Adaptive Streaming: A Review of Subjective Studies. In *International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 141–146, 2014.
- [122] Margaret Pinson, Marc Sullivan, and Andrew Catellier. A New Method for Immersive Audiovisual Subjective Testing. In *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2014.

- [123] Francesca De Simone, Jesús Gutiérrez, and Patrick Le Callet. Complexity Measurement and Characterization of 360-degree Content. In *Human Vision and Electronic Imaging*, volume 2019, pages 216–1–216–7, 2019.
- [124] Hanseul Jun, Mark Roman Miller, Fernanda Herrera, Byron Reeves, and Jeremy N Bailenson. Stimulus Sampling with 360-Videos: Examining Head Movements, Arousal, Presence, Simulator Sickness, and Preference on a Large Sample of Participants and Videos. *IEEE Transactions on Affective Computing*, 13(3):1416–1425, June 2020.
- [125] Lucjan Janowski, Ludovic Malfait, and Margaret H. Pinson. Evaluating Experiment Design with Unrepeated Scenes for Video Quality Subjective Assessment. *Quality and User Experience*, 4(2), June 2019.
- [126] Benjamin J. Li, Jeremy N. Bailenson, Adam Pines, Walter J. Greenleaf, and Leanne M. Williams. A Public Database of Immersive VR Videos with Corresponding Ratings of Arousal, Valence, and Correlations between Head Movements and Self Report Measures. *Frontiers in Psychology*, 8, December 2017.
- [127] Xavier Corbillon, Francesca De Simone, and Gwendal Simon. 360-degree Video Head Movement Dataset. In *ACM on Multimedia Systems Conference (MMSys)*, pages 199–204, 2017.
- [128] Wen-Chih Lo, Ching-Ling Fan, Jean Lee, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. 360 Video Viewing Dataset in Head-Mounted Virtual Reality. In *ACM on Multimedia Systems Conference (MMSys)*, pages 211–216, Taipei, Taiwan, 2017.
- [129] Jiachen Yang, Tianlin Liu, Bin Jiang, Houbing Song, and Wen Lu. 3D Panoramic Virtual Reality Video Quality Assessment based on 3D Convolutional Neural Networks. *IEEE Access*, 6:38669–38682, July 2018.
- [130] Riva, Giuseppe and Mantovani, Fabrizia and Capideville, Claret Samantha and Preziosa, Alessandra and Morganti, Francesca and Villani, Daniela and Gaggioli, Andrea and Botella, Cristina and Alcañiz, Mariano. Affective Interactions using Virtual Reality: the Link Between Presence and Emotions. *CyberPsychology & Behavior*, 10(1):45–56, February 2007.
- [131] Andrew K Przybylski, Kou Murayama, Cody R DeHaan, and Valerie Gladwell. Motivational, Emotional, and Behavioral Correlates of Fear of Missing Out. *Computers in Human Behavior*, 29(4):1841–1848, July 2013.
- [132] Alexandra Voinescu, Liviu-Andrei Fodor, Danaë Stanton Fraser, Miguel Mejías, and Daniel David. Exploring the Usability of Nesplora Aquarium, a Virtual Reality System for Neuropsychological Assessment of Attention and Executive Functioning. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 1207–1208, Osaka, Japan, 2019.
- [133] Nicola S Schutte and Emma J Stilinović. Facilitating Empathy Through Virtual Reality. *Motivation and Emotion*, 41(6):708–712, October 2017.
- [134] Iis P Tussyadiah, Dan Wang, Timothy H Jung, and M Claudia Tom Dieck. Virtual Reality, Presence, and Attitude Change: Empirical Evidence from Tourism. *Tourism Management*, 66:140–154, June 2018.

- [135] Tuuli Keskinen, Ville Mäkelä, Pekka Kallioniemi, Jaakko Hakulinen, Jussi Karhu, Kimmo Ronkainen, John Mäkelä, and Markku Turunen. The Effect of Camera Height, Actor Behavior, and Viewer Position on the User Experience of 360 Videos. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 423–430, Osaka, Japan, 2019.
- [136] Ingwer Borg. Facet Theory. In *Encyclopedia of Statistics in Behavioral Science*. 2005.
- [137] Simon NB Gunkel, Hans M Stokking, Martin J Prins, Nanda van der Stap, Frank B ter Haar, and Omar A Niamut. Virtual Reality Conferencing: Multi-user Immersive VR Experiences on the Web. In *ACM Multimedia Systems Conference (MMSys)*, pages 498–501, Amsterdam, Netherlands, 2018.
- [138] Alexandra Covaci, Ramona Trestian, Estêvão Bissoli Saleme, Ioan-Sorin Comsa, Gebremariam Assres, Celso AS Santos, and Gheorghita Ghinea. 360 Mulsemmedia: A Way to Improve Subjective QoE in 360 Videos. In *ACM International Conference on Multimedia (MM)*, pages 2378–2386, Nice, France, 2019.
- [139] Pablo Pérez, Ester González-Sosa, Redouane Kachach, Jaime Ruiz, Ignacio Benito, Francisco Pereira, and Álvaro Villegas. Immersive Gastronomic Experience with Distributed Reality. In *IEEE Workshop on Everyday VR (WEVR)*, pages 1–6, Osaka, Japan, 2019.
- [140] Julia Himmelsbach, Stephanie Schwarz, Cornelia Gerdenitsch, Beatrix Wais-Zechmann, Jan Bobeth, and Manfred Tscheligi. Do We Care About Diversity in Human Computer Interaction: A Comprehensive Content Analysis on Diversity Dimensions in Research. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–16, Glasgow, Scotland Uk, 2019.
- [141] Tabitha C Peck, Laura E Sockol, and Sarah M Hancock. Mind the Gap: The Underrepresentation of Female Participants and Authors in Virtual Reality Research. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1945–1954, February 2020.
- [142] Lucjan Janowski and Margaret Pinson. The Accuracy of Subjects in a Quality Experiment: A Theoretical Subject Model. *IEEE Transactions on Multimedia*, 17(12):2210–2224, October 2015.
- [143] Nicola Cranley, Philip Perry, and Liam Murphy. User Perception of Adapting Video Quality. *International Journal of Human-Computer Studies*, 64(8):637–647, August 2006.
- [144] Deepti Ghadiyaram, Janice Pan, and Alan C Bovik. A Subjective and Objective Study of Stalling Events in Mobile Streaming Videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(1):183–197, November 2017.
- [145] Philip Kortum and Marc Sullivan. The effect of Content Desirability on Subjective Video Quality Ratings. *Human factors*, 52(1):105–118, May 2010.
- [146] Samira Tavakoli, Sebastian Egger, Michael Seufert, Raimund Schatz, Kjell Brunnström, and Narciso García. Perceptual Quality of HTTP Adaptive Streaming Strategies: Cross-experimental Analysis of Multi-laboratory and Crowdsourced Subjective Studies. *IEEE Journal on Selected Areas in Communications*, 34(8):2141–2153, August 2016.

- [147] Christos George Bampis, Zhi Li, Anush Krishna Moorthy, Ioannis Katsavounidis, Anne Aaron, and Alan Conrad Bovik. Study of Temporal Effects on Subjective Video Quality of Experience. *IEEE Transactions on Image Processing*, 26(11):5217–5231, July 2017.
- [148] Alexander Raake, Marie-Neige Garcia, Werner Robitza, Peter List, Steve Göring, and Bernhard Feiten. A Bitstream-based, Scalable Video-Quality Model for HTTP Adaptive Streaming: ITU-T P. 1203.1. In *International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, Erfurt, Germany, 2017.
- [149] Skyler T Hawk, Loes Keijsers, Susan JT Branje, Jolien Van der Graaff, Minet de Wied, and Wim Meeus. Examining the Interpersonal Reactivity Index (IRI) among Early and Late Adolescents and their Mothers. *Journal of Personality Assessment*, 95(1):96–106, September 2013.
- [150] Anne-Laure Gilet, Nathalie Mella, Joseph Studer, Daniel Grünh, and Gisela Labouvie-Vief. Assessing Dispositional Empathy in Adults: A French Validation of the Interpersonal Reactivity Index (IRI). *Canadian Journal of Behavioural Science*, 45(1):42, January 2013.
- [151] Noah J Goldstein, I Stephanie Vezich, and Jenessa R Shapiro. Perceived Perspective Taking: When Others Walk in our Shoes. *Journal of personality and social psychology*, 106(6):941, 2014.
- [152] Tooba Ahsen, Christina Yu, Amanda O’Brien, Ralf W Schlosser, Howard C. Shane, Dylan Oesch-Emmel, Eileen T. Crehan, and Fahad Dogar. Designing a Customizable Picture-Based Augmented Reality Application For Therapists and Educational Professionals Working in Autistic Contexts. In *International ACM SIGACCESS Conference on Computers and Accessibility*, Athens, Greece, 2022.
- [153] Adrián Borrego, Jorge Latorre, Mariano Alcañiz, and Roberto Llorens. Embodiment and Presence in Virtual Reality After Stroke. A Comparative Study with Healthy Subjects. *Frontiers in neurology*, 10:1061, October 2019.
- [154] Ceenu George, Michael Spitzer, and Heinrich Hussmann. Training in IVR: Investigating the Effect of Instructor Design on Social Presence and Performance of the VR User. In *ACM Symposium on Virtual Reality Software and Technology (VRST)*, pages 1–5, Tokyo, Japan, 2018.
- [155] Kevin Yu, Gleb Gorbachev, Ulrich Eck, Frieder Pankratz, Nassir Navab, and Daniel Roth. Avatars for Teleconsultation: Effects of Avatar Embodiment Techniques on User Perception in 3D Asymmetric Telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 27(11):4129–4139, August 2021.
- [156] Carlos Cortés, Marta Orduna, Pablo Pérez, and Narciso García. Collaborative Interfaces for XR Immersive Learning. In *ACM International Conference on Interactive Media Experiences (IMX)*, pages 209–215, Aveiro, JB, Portugal, 2022.
- [157] Andrew Best, Sahil Narang, and Dinesh Manocha. SPA: Verbal Interactions between Agents and Avatars in Shared Virtual Environments using Propositional Planning. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 117–126, 2020.

- [158] Maia Garau, Mel Slater, Vinoba Vinayagamoorthy, Andrea Brogni, Anthony Steed, and M Angela Sasse. The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of Communication in a Shared Immersive Virtual Environment. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 529–536, 2003.
- [159] Carlos Cortés, María Rubio, Pablo Pérez, Beatriz Sánchez, and Narciso García. QoE Study of Natural Interaction in Extended Reality Environment for Immersive Training. In *IEEE Conference on Virtual Reality and 3D User Interface (VR)*, pages 363–368, Christchurch, New Zealand, 2022.
- [160] Jie Li, Yiping Kong, Thomas Röggl, Francesca De Simone, Swamy Ananthanarayan, Huib de Ridder, Abdallah El Ali, and Pablo Cesar. Measuring and Understanding Photo Sharing Experiences in Social Virtual Reality. In *ACM International Conference on Human Factors in Computing Systems (CHI)*, Glasgow, Scotland UK, 2019.
- [161] M. Rasel Mahmud, Michael Stewart, Alberto Cordova, and John Quarles. Auditory Feedback to Make Walking in Virtual Reality More Accessible. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 847–856, Singapore, 2022.
- [162] Jason Lawrence, Dan B Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G Desloge, Tommy Fortes, Eric M Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, et al. Project starline: A high-fidelity telepresence system. 2021.
- [163] Tara Behrend, Steven Toaddy, Lori Foster Thompson, and David J Sharek. The Effects of Avatar Appearance on Interviewer Ratings in Virtual Employment Interviews. *Computers in Human Behavior*, 28(6):2128–2133, 2012.
- [164] Edward De Bono. *Six Thinking Hats: The Multi-million Bestselling Guide to Running Better Meetings and Making Faster Decisions*. Penguin UK, 2017.
- [165] Arthur Aron, Elaine N Aron, and Danny Smollan. Inclusion of Other in the Self Scale and the Structure of Interpersonal Closeness. *Journal of personality and social psychology*, 63(4): 596, 1992.
- [166] Matias Volonte, Yu-Chun Hsu, Kuan-Yu Liu, Joe P Mazer, Sai-Keung Wong, and Sabarish V Babu. Effects of Interacting with a Crowd of Emotional Virtual Humans on Users’ Affective and Non-verbal Behaviors. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 293–302, 2020.
- [167] Edmund R Thompson. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (panas). *Journal of Cross-Cultural Psychology*, 38(2):227–242, July 2007.
- [168] Ruei-Che Chang, Chao-Hsien Ting, Chia-Sheng Hung, Wan-Chen Lee, Liang-Jin Chen, Yu-Tzu Chao, Bing-Yu Chen, and Anhong Guo. OmniScribe: Authoring Immersive Audio Descriptions for 360° Videos. In *ACM Symposium on User Interface Software and Technology (UIST)*, pages 1–14, Bend, OR, USA, 2022.

- [169] Mariachiara Rapuano, Antonella Ferrara, Filomena Leonela Sbordone, Francesco Ruotolo, Gennaro Ruggiero, and Tina Iachini. The Appearance of the Avatar Can Enhance the Sense of Co-Presence During Virtual Interactions with Users. In *PSYCHOBIT*, 2020.
- [170] Redouane Kachach, Sandra Morcuende, Diego González-Morin, Pablo Pérez, Ester González-Sosa, Francisco Pereira, and Álvaro Villegas. The Owl: Immersive telepresence communication for hybrid conferences. In *IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 451–452, Bari, Italy, 2021.
- [171] Katherine M Tsui, Munjal Desai, and Holly A Yanco. Towards Measuring the Quality of Interaction: Communication through Telepresence Robots. In *Proceedings of the Workshop on Performance Metrics for Intelligent Systems*, pages 101–108, 2012.
- [172] Divine Maloney, Guo Freeman, and Andrew Robb. Stay Connected in An Immersive World: Why Teenagers Engage in Social Virtual Reality. In *ACM Interaction Design and Children (IDC)*, pages 69–79, Athens, Greece, 2021.
- [173] Tanya Hill and Hanneke du Preez. A Longitudinal Study of Students’ Perceptions of Immersive Virtual Reality Teaching Interventions. In *IEEE International Conference of the Immersive Learning Research Network (iLRN)*, pages 1–7, Eureka, CA, USA, 2021.
- [174] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, Accessed: 2023-02-18.
- [175] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014. URL <https://arxiv.org/abs/1409.1556>.
- [176] Rufat Rzayev, Sven Mayer, Christian Krauter, and Niels Henze. Notification in VR: The Effect of Notification Placement, Task and Environment. In *ACM Annual Symposium on Computer-Human Interaction in Play (CHI Play)*, page 199–211, Barcelona, Spain, 2019.
- [177] Mazin Ali, Ferat Sahin, Shitij Kumar, and Celal Savur. 360° View Camera based Visual Assistive Technology for Contextual Scene Information. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2135–2140, Banff, AB, Canada, 2017.
- [178] Yeong Won Kim, Chang-Ryeol Lee, Dae-Yong Cho, Yong Hoon Kwon, Hyeok-Jae Choi, and Kuk-Jin Yoon. Automatic Content-aware Projection for 360 Videos. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4753–4761, Venice, Italy, 2017.
- [179] Wenyan Yang, Yanlin Qian, Joni-Kristian Kämäräinen, Francesco Cricri, and Lixin Fan. Object Detection in Equirectangular Panorama. In *IEEE International Conference on Pattern Recognition (ICPR)*, pages 2190–2195, Beijing, China, 2018.
- [180] Kevin Pfeil, Pamela J Wisniewski, and Joseph J Laviola Jr. The effects of gender and the presence of third-party humans on telepresence camera height preferences. In *ACM Symposium on Applied Perception (SAP)*, Virtual Event, USA, 2020.
- [181] Grupo de Tratamiento de Imágenes. EVENT-CLASS. <http://www.gti.ssr.upm.es/data/event-class>, Accessed: 2023-02-18.

- [182] ITU-T Recommendation P.910. Subjective video quality assessment methods for multimedia applications. April 2008.
- [183] Santiago González Izard, Juan A. Juanes Méndez, Francisco J. García-Peñalvo, Marcelo Jiménez López, Francisco Pastor Vázquez, and Pablo Ruisoto. 360 Vision Applications for Medical Training. In *ACM International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM)*, Cádiz, Spain, 2017.
- [184] Yuxuan Zhang, Hexu Liu, Shih-Chung Kang, and Mohamed Al-Hussein. Virtual Reality Applications for the Built Environment: Research Trends and Opportunities. *Automation in Construction*, 118:103311, October 2020.
- [185] Adam Hadhazy. Stanford course allows students to learn about virtual reality while fully immersed in vr environments. <https://news.stanford.edu/2021/11/05/new-class-among-first-taught-entirely-virtual-reality/>, 2022. Accessed: 2023-02-18.
- [186] Kilian Gloy, Paul Weyhe, Eric Nerenz, Maximilian Kaluschke, Verena Uslar, Gabriel Zachmann, and Dirk Weyhe. Immersive anatomy atlas: Learning factual medical knowledge in a virtual reality environment. *Anatomical Sciences Education*, 15(2):360–368, April 2022.
- [187] Khanh-Duy Le, Morten Fjeld, Ali Alavi, and Andreas Kunz. Immersive Environment for Distributed Creative Collaboration. In *ACM Symposium on Virtual Reality Software and Technology (VRST)*, pages 1–4, Gothenburg, Sweden, 2017.
- [188] Maria V Sanchez-Vives and Mel Slater. From Presence to Consciousness through Virtual Reality. *Nature Reviews Neuroscience*, 6(4):332–339, 2005.
- [189] Robert S. Kennedy, Norman E. Lane, Kevin S. Berbaum, and Michael G. Lilienthal. Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *The International Journal of Aviation Psychology*, 3(3):203–220, November 1993.
- [190] Pablo Pérez, Nuria Oyaga, Jaime J Ruiz, and Álvaro Villegas. Towards Systematic Analysis of Cybersickness in High Motion Omnidirectional Video. In *IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3, Cagliari, Italy, 2018.
- [191] Sarthak Ghosh, Lauren Winston, Nishant Panchal, Philippe Kimura-Thollander, Jeff Hotnog, Douglas Cheong, Gabriel Reyes, and Gregory D. Abowd. NotifiVR: Exploring Interruptions and Notifications in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1447–1456, January 2018.
- [192] Ceenu George, Manuel Demmler, and Heinrich Hussmann. Intelligent Interruptions for IVR: Investigating the Interplay between Presence, Workload and Attention. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–6, Montreal QC, Canada, 2018.
- [193] Marta Orduna, Jesús Gutiérrez, Carlos Manzano, David Ruiz, Julián Cabrera, César Díaz, Pablo Pérez, and Narciso García. EVENT-CLASS: Dataset of Events in the Classroom. In *International Conference on Quality of Multimedia Experience*, pages 81–84, Montreal, Canada, 2021.

- [194] John Brooke, Patrick W. Jordan, Bruce Thomas, Bernard A. Weerdmeester, and Ian McClelland. *Usability Evaluation in Industry*. Taylor & Francis, London, UK, 1996.
- [195] Olivier Augereau, Gabriel Brocheton, and Pedro Paulo Do Prado Neto. An Open Platform for Research about Cognitive Load in Virtual Reality. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 54–55, Christchurch, New Zealand, 2022.
- [196] Conor Keighrey, Ronan Flynn, Siobhan Murray, and Niall Murray. A Physiology-based QoE Comparison of Interactive Augmented Reality, Virtual Reality and Tablet-based Applications. *IEEE Transactions on Multimedia*, 23:333–341, March 2020.
- [197] Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le Callet. State-of-the-art in 360 Video/Image Processing: Perception, Assessment and Compression. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):5–26, January 2020.
- [198] Tanja Aitamurto, Andrea Stevenson Won, Sukolsak Sakshuwong, Byungdoo Kim, Yasamin Sadeghi, Krysten Stein, Peter G Royal, and Catherine Lynn Kircos. From FoMO to JoMO: Examining the Fear and Joy of Missing out and Presence in a 360 Video Viewing Experience. In *ACM International Conference on Human Factors in Computing Systems (CHI)*, pages 1–14, Yokohama, Japan, 2021.
- [199] Jianhao Du, Ha Manh Do, and Weihua Sheng. Human–Robot Collaborative Control in a Virtual-Reality-Based Telepresence System. *International Journal of Social Robotics*, 13(6): 1295–1306, November 2021.
- [200] Jeeyun Oh, Eunjoo Jin, Sabitha Sudarshan, Soya Nah, and Na Yu. Does 360-degree Video Enhance Engagement with Global Warming?: The Mediating Role of Spatial Presence and Emotions. *Environmental Communication*, 15(6):731–748, November 2021.
- [201] Lei Zhang, Yanyan Suo, Ximing Wu, Feng Wang, Yuchi Chen, Laizhong Cui, Jiangchuan Liu, and Zhong Ming. TBRA: Tiling and Bitrate Adaptation for Mobile 360-Degree Video Streaming. In *ACM International Conference on Multimedia (MM)*, pages 4007–4015, Virtual Event China, 2021.
- [202] Dmitry Alexandrovsky, Susanne Putze, Michael Bonfert, Sebastian Höffner, Pitt Michelmann, Dirk Wenig, Rainer Malaka, and Jan David Smeddinck. Examining Design Choices of Questionnaires in VR User Studies. In *ACM International Conference on Human Factors in Computing Systems (CHI)*, page 1–21, Honolulu, HI, USA, 2020.