FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Collaborative interaction in immersive 360º experiences

Pedro Hugo Lima Noevo

Mestrado em Engenharia Informática e Computação

Supervisor: Prof. Rui Rodrigues

Second Supervisor: Prof.ª Teresa Matos

March, 2022

Collaborative interaction in immersive 360º experiences

Pedro Hugo Lima Noevo

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Prof. Daniel Mendes
Referee: Prof. Rui Rodrigues
Referee: Prof. Paulo Dias

March, 2022

Resumo

Os sistemas de reprodução de vídeo tornaram-se, a cada dia, mais habituais e utilizados. Consequentemente, foram criadas extensões desta tecnologia permitindo colaboração multipessoal de modo a poder assistir remotamente e sincronamente. Exemplos conhecidos são o *Watch2gether*, *Sync Video* e *Netflix Party*, que nos permitem assistir vídeos síncrona e remotamente com amigos. Estas aplicações de visualização conjunta, apesar de bem desenvolvidas, estão limitadas ao clássico formato 2D, não se estendendo a vídeos 360º. O principal objetivo deste projeto é então expandir a pesquisa nesta área ao desenvolver um sistema colaborativo para visualização de vídeos 360º.

Foi realizada uma investigação sobre as vantagens e desvantagens de assistir a um vídeo 360º, de maneira a conseguir descobrir o que é a essência destes vídeos e mantê-la, integrando também a inclusão de outros utilizadores. Ao tentar obter respostas relativamente ao tipo de atividades colaborativas a aplicar num leitor de vídeos 360º, é imprescindível analisar o estado em que os sistemas colaborativos se encontram hoje em dia e posteriormente afunilar a pesquisa para a colaboração em ambientes virtuais e em vídeos. De maneira a compartimentalizar e facilitar esta pesquisa são considerados os seguintes temas de forma individual: a visualização de vídeos 360º, a generalidade dos sistemas colaborativos, a aplicação de colaboração em ambientes virtuais e os sistemas de vídeo colaborativos.

Já foram direcionados vários esforços na área de vídeos 360º, sendo um deles o projeto AV360, aplicação que permite ao utilizador editar e visualizar este tipo de vídeos com anotações e guias. A exploração feita no contexto desta dissertação usará como base as tecnologias utilizadas no projeto AV360. Dentro de todos os métodos analisados só os adaptáveis a ambientes imersivos e a vídeos são escolhidos e desenvolvidos neste projeto, de modo que foi criado um sistema de colaboração em vídeos 360º. O software permite que os utilizadores assistam em simultâneo a um vídeo, comunicando de forma ativa através das features desenvolvidas. Estas features assistem o utilizador na interação com os outros, permitindo a partilha de pontos de interesse no vídeo e ajudando na orientação dos participantes.

O plano de desenvolvimento deste sistema permitiu a realização de pequenos testes durante a sua implementação. No final, foram realizados testes de utilizador. Os participantes experimentaram ver vídeos 360º em diferentes condições, mais especificamente, sem colaboração, e com colaboração mas com possível limitação nas funcionalidades disponibilizadas pelo software. Após a visualização dos vídeos, responderam a questionários relativamente à experiencia. Os resultados obtidos demonstram que a interação com outras pessoas ao ver vídeos imersivos tem as suas vantagens e que certas features podem ser adicionadas ao sistema em questão, de maneira a melhorar a experiência do utilizador.

Palavras Chave: Computação centrada no ser Humano, Interação Humano-Computador, Dispositivos de interação, Colaboração, Vídeos 360º, Ambientes virtuais colaborativos

Abstract

Video players have become usual in our everyday devices. Consequently, extensions of this technology were created allowing multiple people to collaborate and watch videos remotely and synchronously. Well-known examples are Watch2gether, Sync Video and Netflix Party, which let us watch videos synchronously and remotely with friends. These applications, although well developed, are limited to the typical format, not extending to 360º videos. The main objective of this project is then to expand the research in this area by developing a collaborative system for 360º videos.

An investigation was carried out on the advantages and disadvantages of watching a 360º video, in order to discover what is the essence of these videos and maintain it when including the inclusion of other users. When trying to obtain answers regarding the type of collaborative activities to be applied in a 360º video player, it is essential to analyze the state in which collaborative systems are and subsequently narrow the research for collaboration in virtual environments and videos. In order to compartmentalize and ease this research, the following topics are considered individually: the viewing of 360º videos, the generality of collaborative systems, the application of collaboration in virtual environments and collaborative video systems.

Several efforts have already been made in the area of 360º videos, one of them being the AV360 project, an application that allows the user to edit and view this type of videos with annotations and guides. The exploration made in the context of this dissertation will be based on the technologies used in the AV360 project. Among all the methods analyzed, only those adaptable to immersive environments and videos are chosen and developed in this project. A 360º video collaboration system was then created. The software allows users to simultaneously watch a video with other users, while actively communicating through the developed features. These features assist the user in interacting with others, allowing the sharing of points of interest in the video and helping to guide the participants.

The development plan of this system allowed carrying out small tests during its implementation. In the end, user tests were carried out. Participants experienced watching 360º videos under different conditions, more specifically, without collaboration, and with collaboration but with possible limitations in the features provided by the software. After viewing the videos, they answered questionnaires regarding their experience. The results obtained demonstrate that interacting with other people when watching immersive videos has its advantages and that certain features can be added to the system in question, in order to improve the user experience.

Keywords: Human-centered Computing, Human Computer Interaction (HCI), Interaction devices, Collaboration, 360º Videos, Collaborative Virtual Environments (CVE)

Acknowledgments

To my mother and father, who have always been there for me and supported me in every decision I made.

To my girlfriend that helped me in every moment and handled my worst moods with encouraging words.

To my supervisors, Rui Rodrigues and Teresa Matos who helped at any moment and guided me, not only in my theoretical knowledge, but also through the fatigue of the work.

To my friends that kept me joyful and entertained throughout all the phases and helped me whenever they could.

Pedro Noevo

"If you think you are too small to make a difference,
try sleeping with a mosquito."

Dalai Lama

Contents

List of Figures

List of Tables

Abbreviations

AR	Augmented Reality
AV360	Augmented Virtuality 360
COVID19	Corona Virus Disease 2019
CSCW	Computer-Supported Collaborative Work
CSIG	Center for Information Systems and Computer Graphics
CVE	Collaborative Virtual Environments
DCVV	Distributed Collaborative Video Viewing
HCI	Human-Computer Interaction
HMD	Head-mounted Device
INESC TEC	Institute for Systems and Computer Engineering, Technology and Science
MQTT	Message Queuing Telemetry Transport
MR	Mixed Reality
QoE	Quality of Experience
RQ1	Research Question 1
RQ2	Research Question 2
RQ3	Research Question 3
UEQ	User Experience Questionnaire
URL	Uniform Resource Locators
VE	Virtual Environment
VR	Virtual Reality

Chapter 1

Introduction

With the release of the Google Cardboard in 2014 and the introduction of low-cost VR headsets to the market, immersive experiences have become much more affordable and appealing to the general population. Experiences like three hundred sixty-degree videos became popular. YouTube created opportunities for this technology by providing support for such media and allowing the upload and display in March 2015 [AEB20]. Beyond the increase in the easiness to experience immersive content, its creation also became more accessible. Better and cheaper 360º cameras are on the agenda of companies related to the technology industry. These relatively new cameras allow regular users to create their "homemade" content.

Consolidated with the right equipment, 360-degree videos have the possibility to create unique experiences of immersion in different stories where the viewers have the freedom to explore their own point of view in the narrated adventures. During this document, it is possible to understand that this type of videos enhances the user experience in more than one way. This versatility made the technology expand to different areas in the market. Besides storytelling, this media form has gained value in different branches like education, marketing, journalism, and clinical training [SWGW18].

The technologies available to display traditional 2D videos that are part of our everyday life have been developed to the point that we can from one side of the world to the other visualise them synchronously with one or more people. The ability to watch videos collaboratively can increase the quality of user experience and even be extremely useful in some professional scenarios. The impact of collaboration while visualising immersive videos is yet to be studied in more detail. Is it worth developing the same collaborative features implemented in the traditional videos for the 360-degree videos? Moreover, are these features enough for a good user experience?

1.1 Context and Motivation

Augmented Video 360 (AV360) is a project supported by Google's Digital News Initiative and developed at the Center for Information Systems and Computer Graphics (CSIG) of INESC TEC. It allows anyone that desires to edit 360-degree videos to create dynamic annotations, narrative attention focus points, oriented 3D subtitles and off-the-shelf visual effects. Besides that, the project provides the edited video through a web player [av3]. The work presented in this document uses the AV360 as a starting point for its technologies.

Since the appearance of COVID19, remote systems have become a crucial alternative for tasks that once required physical presence. If it is a compulsory assignment or just the casual "hang out" with friends, humans have the need to stay connected as a society and interact between themselves. As introduced before, 360-degree videos have gained significant popularity over the years. The possibility of watching one of these videos and socialising with someone else who is watching the video simultaneously with us is attractive. It can create new experiences for the users. Possible scenarios to use are virtual tours and events, crime scenes, clinical cases, education and even the funny video that two friends watch to laugh together. These scenarios are in different areas and might have different needs when users try to fulfil their objectives. The study of what a collaborative 360-degree video visualiser might need in each of these situations is necessary, and this need drives us to continue the research presented in this document.

1.2 Research Questions

There are numerous possibilities when interacting with someone. When trying to do that on a device, users can lose this freedom because technology has its limits. If not co-located, physical contact is denied and depending on the devices available, vision and other senses can be denied to several of the parties involved. The research goal of this project is to understand the impact of collaboration in 360-degree videos. To achieve this goal, the first research question defined for this research is the following:

RQ1: Does collaboration with other users help to improve the quality of experience in 360º videos?

While doing the different phases of the project, we acknowledge that the different collaborative features available in a 360-degree video visualiser can have a significant impact on the user's quality of experience. Thus, the research introduced another question: What are the best collaboration features to implement in a software system that reproduces 360º videos? The downside of this question is that it is too broad and almost impossible to be answered because there is always the possibility to create a different collaborative feature. So, we decided to narrow the possibilities and evaluate the necessity of voice communication and the spatial orientation features developed and described in chapter 4.

The second and third research questions are:

RQ2: Does voice communication help to improve the quality of experience while watching 360-degree videos collaboratively?

RQ3: Do Radar and Point of Interest Share (features explained in chapter 3) help to improve the quality of experience while watching 360-degree videos collaboratively?

The acronyms RQ1, RQ2 and RQ3 will be used to simplify identifying the research questions in this document.

1.3 Objectives

The main goal is to investigate and answer the research questions. However, more specific objectives were set to help us define a path to achieve the answers needed.

- Investigate the current state of virtual environments, 360º videos, collaborative systems and collaborative 360º videos.
- Determine which collaborative techniques to apply during the visualisation of 360º videos.
- Design the previously determined collaborative techniques.
- Implement the designed features using the same technologies as the AV360 project.
- Evaluate and test the implemented collaboration techniques in order to find answers to RQ1, RQ2 and RQ3.

The first topic is a state of the art review where the main areas related to collaborative visualisation of 360-degree videos are analysed and help us establish an initial idea of what is needed for the implementation of such system. It helped in the definition of RQ2 and RQ3. The second and third topics are essential for the proposal of the collaborative visualiser. The fourth objective is implementing the system that supports the results of the research questions that are answered in the fifth and final objective, the execution, evaluation and discussion of the system tests.

1.4 Main features of the proposed solution

Collaboration was implemented for the visualiser of the AV360 project. The final system developed connects users on different locations, allowing them to simultaneously visualise the same 360-degree video.

The system continues to have the same functionalities that AV360 had: an editor where users can edit their videos and add annotations; a 360-degree video player that displays these edited annotations with the video and can be used in a VR headset, mobile device or even a desktop computer. The system is web-based, which means that it is easy to access by anyone with an internet connection.

Adding to these functionalities that already existed in AV360, the users can watch the videos synchronously and benefit from other features like a shared radar to help with the spatial orientation and widgets that they can share to point to detail in the others viewport.

1.5 Document Structure

Chapter 1: The current chapter introduces the context of the dissertation, evidencing the rise of popularity in 360-degree videos and the opportunities to incorporate collaboration features in them. The purpose of this research is also discussed in this chapter with the mention of the research question, objectives and a description of what was implemented.

Chapter 2: The second chapter is dedicated to explaining the research made in the area of virtual environments, 360-degree videos, collaborative systems, more in-depth collaborative virtual environments and collaborative videos, and awareness.

Chapter 3: The third chapter describes the proposal of a collaborative 360-degree player. It is done a detailed description of the collaborative techniques that are possible to be developed after the presentation of some user scenarios for this system. It is also proposed the visual design and system architecture having the system requirements in consideration.

Chapter 4: The fourth chapter details the development of the system proposed in the previous chapter. It describes the technologies used, challenges and solutions found, and thoroughly explains the collaborative features implemented. The system modifications for the tests are identified and explained in this chapter.

Chapter 5: The fifth chapter explains the tests done, their structure and details like the data collected from the users and the videos that were available to them. It also discusses the results obtained in the tests and addresses the research questions.

Chapter 6: Finally, in the last chapter, we review the document and gather the conclusions for the dissertation.

Chapter 2

State of the Art Review

This chapter approaches virtual environments and collaborative systems history, evaluating previous and relevant developments and discoveries.

We introduce in section 2.1 the concept of virtual environments, discussing their current state, how to classify them and pointing out some singularities. Secondly, we address 360º videos in section 2.2, as a topic of main relevance for this work. So we research on what makes 360-degree different and what influences their user experience. After that, in section 2.3, collaborative systems are analysed, investigating what influences the quality of experience on them and more specifically in collaborative virtual environments and collaborative videos. The last topic introduced, present in section 2.4, is about user experience and awareness, and how to improve them both in the context of collaborative systems and virtual environments. Finally, we conclude this chapter in section 2.5, with a summary of all the previously mentioned matters.

2.1 Virtual Environments

Technology expansion led to a significant development in **virtual environments** (VE). As a result, VEs gained considerable fame, mostly recognized because of **virtual reality** (VR) technology. These two terms are commonly misunderstood to have the same meaning, but Annie Luciani wrote an article proposing the difference between them [Luc07]. According to the author: "A Virtual Environment may faithfully recreate an existing real environment or can be completely fictional", and "virtual reality is synonymous to the meaning of virtual environments, in the sense of worlds surrounding the user and being explored by him.". This is just one definition among countless others. For example, Janor Lanier, sometimes recognized as the father of virtual reality, proposes in 1988 a definition of VR: "virtual reality uses the approach of designing clothing devices, 'computer clothing', which is worn directly over the sense organs. The objective is to instrument this in such a way that you can provide exactly the stimulus to the person's sense organs that they would receive if they were in an alternate reality" [Lan88]. This was one of the first definitions, and it was polished by other authors like Jonathan Steuer in 1992, introducing different concepts like presence [Ste92]. Similar to VR, **augmented reality** (AR) adds a digital layer of computer graphics on top of the physical space enabling users to view and interact in real-time [PK20]. Thus, AR is part of the spectrum of **mixed reality** (MR) that allows interaction between physical and digital objects in real-time. Representing the RV continuum, an image containing the different spectrum of the mixed reality is displayed in Figure 2.1 [MTUK95].

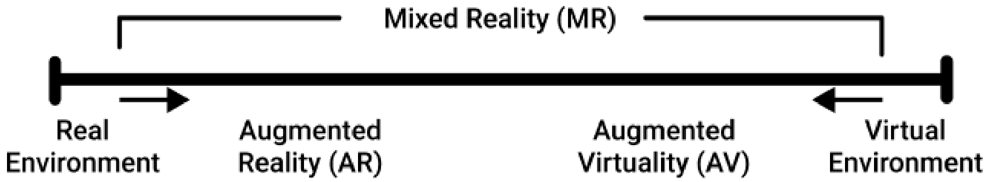


Figure 2.1: Reality-Virtuality (RV) Continuum [MTUK95]

The topic of MR lead us to understand and study new concepts. One frequently referred is **immersion**. In 1975, psychologist Csikszentmihalyi gave a series of definitions for flow [ZD09]: "The state of being completely involved in an activity for its own sake. The ego falls away. Time flies. Every action, movement, and thought follows inevitably from the previous one, like playing jazz. Your whole being is involved, and you are using your skills to the utmost." Flow refers to a psychological state of devoting oneself to a specific activity. Depending on the degree of involvement in an activity, flow falls into immersion, half-immersion, and apartness. Psychologists use the word "immersion" to describe the unique experience in which people are completely attracted by the activity and involved in it [ZD09]. Therefore, it is possible to consider immersion as an intense state of flow. High levels of immersion can create a feeling of **presence**, which is defined as "the extent to which a user feels that he or she is in a particular place, even while physically situated in another place" [SWGW18]. This definition helps us understand the relation between a good user experience with the feeling of presence and a raised sense of immersion.

Lanier of American VPL Research Inc first presented the virtual reality concept in 1989, which described the computer simulation technology [ZD09]. Nevertheless, the first virtual reality system was created by Ivan Sutherland and Bob Sproull and called "the ultimate display" in 1965. A few years pass by, and the endless possibility of new worlds and interactions delight the entertainment industry and their customers. Consequently, virtual reality is closely related to games in the current days. However, besides this business, VR impacts society in other crucial fields like medicine [JBB+20] and education [PFPP21].

Virtual environments can use more than just the typical computer to interact with users. It can encompass a different set of hardware, being those divided into input and output devices [SG96]. An example of the appearance of some of these devices is displayed in Figure 2.2. For the **input devices**, we can find pointing, tracking and speech recognition devices. Pointing devices help indicate where a particular point of focus is and can be crucial for some applications. They can vary from the well-known joysticks and space balls to the less traditional gloves. Tracking devices can either represent a portion of the user's body in the VE or update the image displayed to the user. They are typically mounted with a glove, flying mouse or any operator body part. Finally, speech recognition systems learn the user's speech patterns as he/she reads a predetermined list of words, after which that person may issue a wide range of voice commands. **Output devices** are vital for the user to perceive the system, and they can come in the form of visualization, audio and haptic devices. The essence of three-dimensional visualization is a topic of great importance and achieved with visualization devices like head-mounted display devices (HMDs), multi-wall displays (CAVEs), and shutter glasses combined with traditional CRT displays. To enlarge the sense of immersion in the VE, audio devices should have three-dimensional sound effects, these are rather standard and used worldwide. Lastly, haptic devices provide a physical sensation of touch, significantly providing a better sense of immersion. Having all these devices in consideration is imperative to define future requirements and limits in projects within the area of virtual reality.

Figure 2.2: Input and output devices example [AA06]

360-degree videos are a small part of VE but a big focus for this project. The following section exhibits more detailed research about these video's specific elements and limitations.

2.2 360-Degree Videos

360-degree videos, or **Immersive videos**, provide users with a spherical view and an immersive experience of the camera's surroundings allowing the viewer to control its orientation, as exemplified in Figure 2.3. This remarkable trait increased this video's type popularity in many contexts, including education, marketing, journalism, and clinical training [SWGW18]. Furthermore, the potential for immersive experience provision led to the creation of numerous 360° videos and online videos. Besides that, popular social media platforms such as YouTube and Facebook allowed viewers to upload and view 360° videos [LCH+17].

The production of a 360º video is not as simple as doing the traditional 2D video footage. Besides the unusual cameras used that benefit from their multiple lenses (often two wide-angle

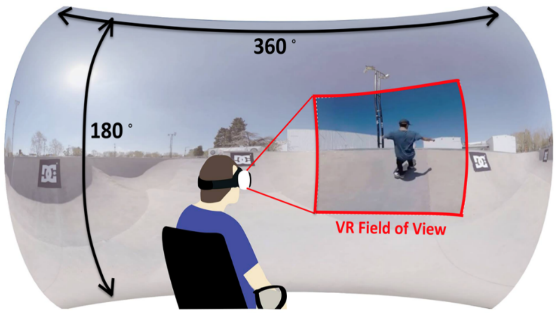


Figure 2.3: 360-degree video spherical view [LCH+17]

lenses), allowing footages of the environment all around the device, there is also a wide range of new features that the immersive environment can provide. For example, Argyriou, Economou, and Bouki [AEB20], describe 360º immersive video applications design aspects and propose a workflow for their development, as presented in Figure 2.4. First, pre-defined video scripts determine the video content and editing phases. The video is subsequently recorded, followed by the production stage creating the VR scenes using a game engine that supports application development for VR headsets. If it is pre-planned in the video development, the programming of the video functionality takes place after all the other steps, leading to the final product. The described workflow has steps like creating VR scenes using a game engine and gamification that are not strictly essential for the development of these videos. These actions might create a more complete user experience if implemented.

Apart from a good production of a 360º video, additional elements should be considered to enhance **audience engagement**, for example, awareness regarding the *display type* (including HMDs, mobile devices, and personal computers) and the *viewport dynamic* (the area of the 360-degree video frame that is displayed at a given time) [SWGW18]. In this study, following O'Brien and Toms' work [OT10], audience engagement is conceptualised as the extent to which an audience achieves deep cognitive, affective, and behavioural involvement with 360-degree videos. Some studies argue that sense of presence and *motion sickness* are the two main factors determining audience engagement in 360º videos [NBM+19]. "Motion sickness occurs if there is a conflict between visual, vestibular, and proprioceptive signals in response to a motion stimulus" [NBM+19]. However, Ayoung Suh et al. [SWGW18] suggest that motion sickness does not have that much influence on audience engagement, justifying it with "the high degree of presence may override the negative effect of motion sickness on audience engagement". Finally, the research also concludes that HMDs (Head-mounted devices) outperformed mobile devices in creating more significant degrees of presence, like MVPs (moving viewports) also improve presence compared to SVPs (static viewports).

Quality of experience (QoE) varies from user to user. An attempt to define it is "QoE is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application

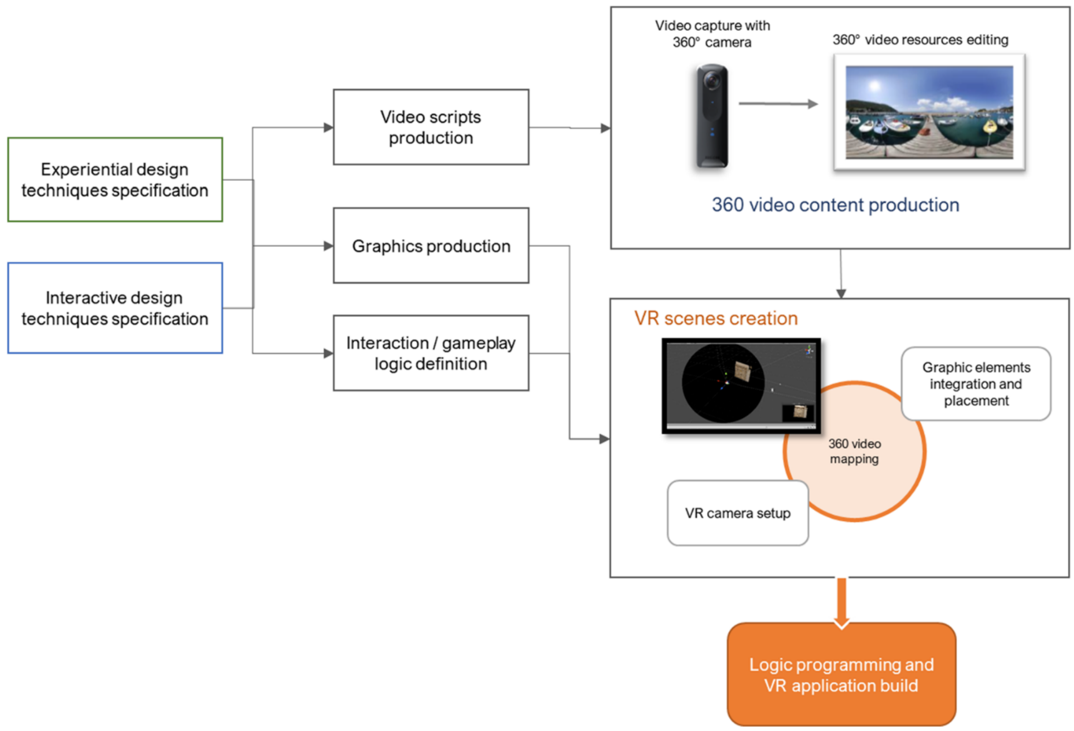


Figure 2.4: 360º immersive video application development workflow [AEB20]

or service in the light of the user's personality and current state" [CMP12]. Factors that have a meaningful impact on the QoE of 360-degree videos are bitrate, screen resolution and stalling. *Bitrate* is the amount of data encoded for a unit of time, and *screen resolution* is the number of pixels spread across a display. *Stalling* is when a video can not be seen continuously. Usually, the video stops because the data needed to continue the reproduction is not yet downloaded.

Streaming 360-degree videos brings up new challenges. They are enormously bandwidth-intensive, particularly with high resolution viewed with HMDs. Therefore, it is inevitable to avoid stalling and bitrate limitations while measuring the quality of experience for 360-degree videos in VR. Studies have considered the impact of stalling on the QoE for 360-degree videos [SSTG17], but Muhammad Shahid Anwar et al. [AWU+20] considered the various stalling impact under different bitrate levels on end-users. The authors assert: "Stalling always impacts the QoE of 360-degree videos, but the strength of this negative impact depends on the video bitrate level. The adverse effect of stalling events is more profound when bitrate level approaches to the high and low end". They conclude that viewers are comfortable with medium quality video when there is any disturbance in playback, either in terms of different stalling events or quality changes. Imagining a perfect scenario where stalling is not present, screen resolution is analysed in a research written by Wenjie Zou, Lihui Yang, Fuzheng Yang, Zhibin Ma, and Qiyong Zhao [ZYY+20]. They state that the user's perception of quality increases with screen resolution. However, it reaches a threshold where the influence stagnates.

Throughout the previous paragraphs, enhancing QoE and audience engagement was the focus. These two elements can be highly affected by **emotions**. Emotions play a vital role in the perception of everything. Five researchers conducted a study about the affective appraisal of immersive videos [THB+20]. Emotions were divided into two main dimensions: valence and arousal. "Valence refers to the degree of positive or negative affective response to a stimulus, while arousal refers to the intensity of the affective response (i.e., the degree of activation or deactivation)." [THB+20]. The authors concluded, as expected, that 360-degree videos had a strong influence over both dimensions. Furthermore, the method that they used to evaluate emotions proved to be successful. It was a simple EmojiGrid, as displayed in Figure 2.5. It can be helpful to receive feedback from users related to their emotions in future works associated with 360-degree videos.

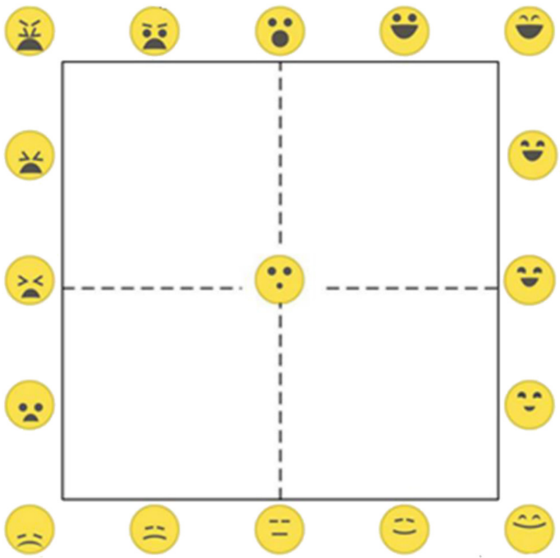


Figure 2.5: EmojiGrid [THB+20]

Emotions are part of each of us individually but can be enhanced when we interact with each other. Shared experiences can create a more pleasant user experience. That is why we introduce a more detailed investigation of **collaborative systems** in the next section.

2.3 Collaborative Systems

Web technology and electronic networks have mitigated numerous disadvantages of physical distance. Manifesting in almost every area, collaborative tools have a notable impact on society with advantages like tremendous time and cost savings, decreased travel requirements, faster and better decision making and improved communication flows. An example of these advantages is the situation of COVID19 crisis that lead to people's inability to be physically present. Collaborative systems helped the world, providing the possibility to fast communication and decision making. "Broadly defined, the field of collaborative computing, otherwise known as computer-supported cooperative work (CSCW), encompasses the use of computers to support coordination and cooperation of two or more people who attempt to perform a task or solve a problem together (Borenstein 1992)." [Sch96].

In 1992, Kari Kuutti and Tuula Arvonen tried to identify potential **computer-supported cooperative work** (CSCW) applications [KA92]. The authors propose this identification employing activity theory concepts. In their research, they defined CSCW and Activity Theory, respectively: "CSCW is defined in this paper as work by multiple active subjects sharing a common object, supported by information technology" and "Broadly defined, Activity Theory is a philosophical framework for studying different forms of human praxis as developmental processes, both individual and social levels interlinked at the same time.". Explained by the authors: "The solution offered by Activity Theory is that there is a need for an intermediate concept - a minimal meaningful context for individual actions – which must form the basic unit of analysis. This unit - better defined and more stable than just an arbitrarily selected context, but also more manageable than a 'whole' social system – is called an activity." [KA92].

Besides identifying when a collaborative system can be developed, research was performed regarding their classification. A simple taxonomy is to distinguish by time and space, as represented in Figure 2.6. The interaction can happen **synchronously** or **asynchronously** and can be **co-located** or **remote**. Four scenarios are identifiable with these two dimensions: synchronous and co-located, synchronous and remote, asynchronous and co-located and asynchronous and remote. Additionally, several taxonomies were proposed based on group size, predictability, application-functionality, coordination process, and others. Figure 2.7 summarises some existing literature about this topic.

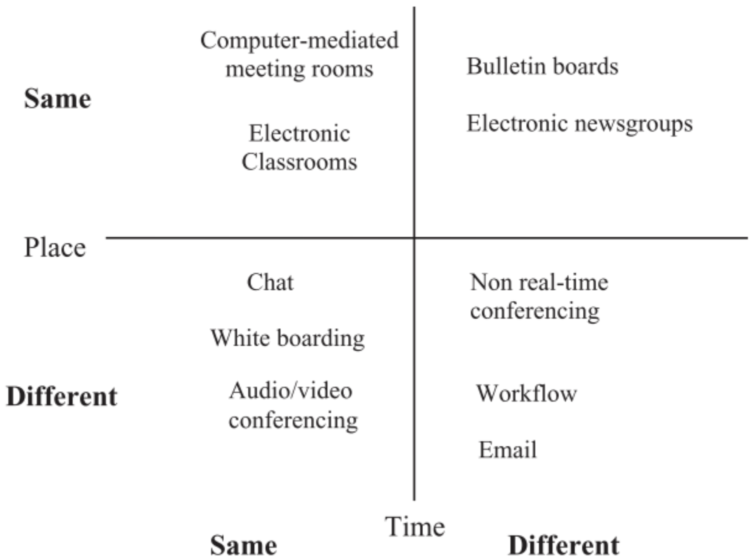


Figure 2.6: Time/space classification [BM02]

Classification dimensions in the literature

Other		Technical			Group issues			Application	Time/Space	Classification criteria/authors
Usability and ergonomics	Mode of interaction	Scalability	Software	Hardware	Types of group tasks	Characteristics of group	Group size			
					X		X			DeSanctis and Gallupe (1987)
			X	X		X				Kraemer and King (1988)
									X	Johansen (1998)
	X							X	X	Ellis, Gibbs and Rein (1991)
X	X	X	X	X			X	X		Jarczyk, Loffler and Volksen (1992)
	X		X						X	Mentzas (1993)
					X					McGrath and Hollingshead (1994)
									X	Grudin (1994)
					X					Malone and Crowston (1994)
								X		Coleman (1995)
			X					X		Ellis (2000)

Figure 2.7: Classification dimensions in the literature. Adapted from [BM02]

Each year, new technologies that apply new styles of collaboration are released. In the research written by Georgia Bafoutsou and Gregoris Mentzas [BM02], the most commonly encountered services collaborative systems provide were recognised as the following: Bulletin board, Discussions, E-mail, E-mail notifications, Online paging/messaging, Chat, Whiteboard, Audio/Video conferencing, Task list, Contact management, Screen sharing, Surveys/polling, Meeting minutes/records, Meeting scheduling tools, Presentation capability, Project management, File and document sharing, Document management and Synchronous work on files/document.

We go deeper into collaborative systems in this project, studying how awareness is affected when collaboration is present and more specific collaborative systems related to virtual environments and 360-degree videos.

2.3.1 Collaborative Virtual Environments

Virtual environments have great potential for the implementation of collaborative tools. **Collaborative virtual environments** (CVEs) are described as "distributed virtual reality systems that offer graphically realised, potentially infinite, digital landscapes." by E. F. Churchill and D. Snowdon [CS98]. Other articles report it as a convergence between VR and CSCW or virtual worlds shared across a computer network [BGRP01].

Collaboration can be incorporated with VE in a variety of ways. Blending the notions addressed in this work of VR, AR and the two dimensions of place and time of collaboration, it is possible to create several scenarios for CVEs [PK20]. First, addressing the current state of **VR** mixed with collaboration techniques, it is possible to identify a reasonable amount of studies in the different dimensions:

- **Co-located and Synchronous:** A physical environment with immersive projection screens enables people to work together in the same workspace. Such environments surround the user with 2D and 3D information. The user can interact and share visual elements in a face-to-face setting, in which communication is supported via speech, gestures, gaze, and non-verbal cues [ISI+17].

- **Co-located and Asynchronous:** In the literature, asynchronous co-located VR collaboration mainly focuses on semi-immersive projection screens and tabletop systems where work is supported by shifts, by "handing over", and "taking over" work [BCBM18]. Such systems provide the same view for collaborators in the physical workspace in which face-to-face interaction is possible.

- **Remote and Synchronous:** It supports remote users to work together in a shared virtual environment by immersing users into a co-located setting and is often referred to as an immersive virtual environment. They support remote and real-time multi-user collaboration, easy interaction, information and data sharing [OP07]. Different tools, features, and functions can directly manipulate objects, navigate, encounter people, and share visual artefacts. Also, collaborative Web 2.0 tools and sharing mechanisms like instant messaging, audio, video, teleconferencing and multimedia presentations can be featured.

- **Remote and Asynchronous:** Where most VR systems support synchronous activities, the asynchronous mode is lacking and not always supported. However, examples exist by leaving data and messages for later review, recording the VE or replaying messages in immersive virtual environments [GPB+00, MGFS13].

Regarding AR and collaboration, there is very little research. Multiple challenges exist, like the role that time plays in the interactions, how to capture annotations and different inputs and revisualise them, and how other forms of communication influence the collaboration. Nevertheless, the opposite facet is explored, and the following topics provide an explanation of **synchronous** collaboration in **AR**:

- **Co-located:** See-through HMDs have been employed to show graphic objects and allowing real-time interactions. It has been used within education by adding annotations in real-time and within engineering, allowing participants to observe and interact with dynamic visual simulations and CAD models.

- **Remote:** Remote AR has its application in multiple industries like factory planning, maintenance, product design, and education, where a huge focus has been on assistance, work instructions and training.

Each of these scenarios has its challenges when being developed. Extracting some **issues and challenges** in CVEs from previous research [CS98, BGRP01], we try to name and explain them:

Individual-group task transitions: Transitions between shared and individual activities can become complex. CVEs are commonly used where a group activity is the main focus, but inverting the situation can be complicated and needs to be handled carefully.

Information overflow: Too many stimuli at once can be overwhelming. In some scenarios, the user does not need to receive all the information from the other participants. It might not be relevant.

Individual viewport dynamic: Researchers support that virtual environments should support subjective views (an unique view for each user), otherwise, the users would be forced to agree on a common, possibly non-optimal, visualisation style. However, different studies supported by the CSCW community suggest that this possibility of different perspectives may hinder people's ability to collaborate. The viewport dynamic should be carefully chosen according to the goal of the system.

Asynchronous interactions: When actions are not physically co-located and synchronous, providing shared context is complex. Tools that provide shared context in asynchronous work contexts are crucial to create awareness of others within the system.

Awareness control: Both in collaborative systems and virtual environments, awareness is a crucial topic and challenge. It is studied with more extensive detail in section 2.4.

Scalability: Limitations on scalability arise from a variety of system bottlenecks. Large numbers of active participants generate high volumes of network traffic, especially movement updates and audio packets. Even if the core network and server facilities can sustain a CVE, the network connection to each participant's machine can become a bottleneck. Finally, the user's local computer must process it and render the shared virtual world at a satisfactory quality while maintaining a sufficiently rapid response to the participants' movements and other actions.

Communication system architecture: The architecture for the communication between users of the system is complex since users worldwide can participate. The three most common ones are client/server, peer-to-peer unicast and peer-to-peer multicast. The first can be overwhelming to the server because it is in the centre of all the communication. The second usually is the most bandwidth-intensive but introduces fewer network delays. Finally, the third is similar to the peer-to-peer unicast but for more than one user, typically using a better bandwidth-efficient network mechanism. These architectures are commonly mixed and used in the same system in different parts of the communication.

The unknown: New kinds of human factors that typically shared systems are not used to deal with are also challenging. These are not possible to be all identified beforehand because they greatly depend on the system built. An example is: Users assume that the other collaborators have the peripheral view of a human being. However, the CVE technology might limit it. Systems intended to support collaborative activities should be designed to explicitly consider the tasks to be achieved and the intended users' social and cognitive characteristics.

For the last section of collaborative systems, we narrow the research to **collaborative 360-degree videos** in an attempt to help with the answers to RQ1, RQ2 and RQ3.

2.3.2 Collaboration in Videos

Collaboration in both 2D and 360-degree videos has been investigated. In this section, we start by exploring the theory of collaboration in traditional 2D videos. Following that, we study how previous attempts of 360-degree videos have endeavoured and the most notable challenges.

Typical 2D videos

Watching a video with a group of people can be enjoyable and has its benefits. For example, previous research on Tutored Video Instruction shows that learning is enhanced when small groups of students watch and discuss lecture videos together [CBS+00]. This paper analyses **Distributed Collaborative Video Viewing** (DCVV) more in-depth. The system allowed groups of students to watch and discuss, together, pre-recorded lectures from multiple locations.

Even though the study is old (2000), the researchers gathered some curious results that should be paid attention to these days. First, the authors acknowledged that users were uncomfortable pausing the video, especially if no bigger/more important entity was present. Second, the communication channel was also rated. Communication via text-chat was considered flawed because the attention from the video would have to shift, and the viewers could not keep track of everything happening simultaneously. Communication via audio proved to be the most efficient, and the video component helped to fix minor issues, like creating empathy and understanding others feelings but was not deemed essential.

One can pause, go backwards, go forward, change the video's speed, communicate through a voice or text channel and even create live annotations. However, the collaborative possibilities in 2D videos are limited and made possible without much problems with the help of the technologies that we have available nowadays. Nevertheless, improvements to satisfy the user are constantly being explored. Asaad Alghamdi, Younes Balah, Mohammad Albejadi and Muhamad Felemban [ABAF20], just like other authors [LKS+16, ZWWZ18], tried to improve the QoE by fixing poor internet coverage problems. Unlike the usual client/server where the server provides a video through streaming, in their software system called BeeCast [ABAF20], they also connect devices from the same network to share the packages received and increase the quality in each node.

360-degree videos

Virtual environments are promising technologies when we talk about collaboration. Do 360-Degree videos, being part of VEs, make them appealing for collaboration? Anthony Tang and Omid Fakourfar were sceptic and wrote a study [TF17] where they say that "such an interface is unlikely to work well in multi-person scenarios.". This quote was their first assumption; after the results, they discuss several problems that may help future developers meaning to build applications involving the joint viewing of 360-degree videos. Their advice and warnings were:

- To better understand what others are looking at in relation to the communication happening, the ability to point, gesture or otherwise reference objects in the video must be somehow given to the users. A feature that can help with this is the provision of a compass to the viewers.
- When co-located, the users tend to point to each other screen. With HMDs, it is impossible and may create a significant obstacle in communication.

- Just like the spatial freedom provided to the user, time can also be at the viewer's disposal to switch. The ability to go backwards and have a label to "come back" to the labelled place is suggested.

- Give the user the possibility to choose between his/her view and the others.

Other researchers worked on the development of these kinds of software systems. 360Anywhere [SCY+18] is one of these applications, described by the authors as "a framework for 360 video–based multi-user collaboration that, in addition to allowing collaborators to view and annotate a 360 live stream, also supports projection of annotations in the 360 stream back into the real-world environment in real-time."

While building the application, several challenges were identified: *Gaze*, when it is not clear which portion of the 360 video collaborators are seeing; *Out-of-sync*, the fact that 360-degree video collaborators do not necessarily share the same view; *Gestures*, gestures performed by one collaborator may be missed by others not sharing the same view. Other challenges were discussed but specifically related to the streaming and annotations created on the video.

For these challenges, solutions were proposed: *Gaze Awareness*, displays of coloured cones that indicate where each user is looking; *Follow Me*, which enables one collaborator to gain control of everyone's 360 feed, synchronizing the view of all users; *Audio/Video Chat*, provides a separate Skype-like channel; *Back in Time*, which enables the remote collaborator to rewind the live stream by 10 seconds; *Annotations*, which provides the functionality to draw, place images, and write text directly into the 360-degree stream; *Calibration*, which enables the user to define one or more projections in the 360-degree live feed.

Finally, the authors matched that Gaze was fixed by Audio/Video Chat, Gaze Awareness, Follow Me, and Annotations; Out-of-sync by Audio/Video Chat and Follow Me; Gestures by Audio/Video Chat.

2.4 User Experience and Awareness

According to the Cambridge dictionary, **awareness** is defined as: "knowledge that something exists, or understanding of a situation or subject at the present time based on information or experience" [Awa21]. Many articles tried to define awareness concerning specific fields that they were involved. In this section, we clarify its importance and meaning when used in collaborative systems and immersive environments.

CSCW software systems try to provide users with **awareness information**, information about the presence, activities, and availability of community members. This type of information is critical for a better user experience, but it should only be broadcasted if the concerned group members have agreed to transmit those data. The understanding of how it should be implemented can be hard to achieve. Various "types" of awareness can be found in a system. For example, group awareness or workspace awareness is the information about one another, shared artefacts, and group processes. "Group awareness, therefore, can be broadly defined as consciousness and information of various aspects of the group and its members." [GST05]. A visual representation of this definition exposing 3 questions that can help with awareness is displayed in Figure 2.8. CSCW propose the division into four types of awareness:

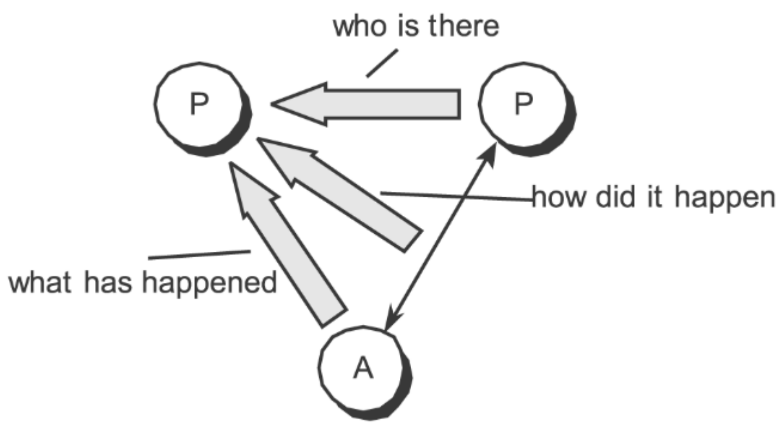


Figure 2.8: Group awareness visual representation [GST05]

- **Informal awareness** - the experience of who is around, what these people are doing and what they will do.
- **Social awareness** - the availability of information such as interest and attention or the emotional state of a conversation partner.
- **Group-structural awareness** - the information about the group and its members, like roles and responsibilities and their status or positions on specific issues.
- **Workspace awareness** - knowledge about the workspace in general, like information about other participants' interactions with the shared space and the artefacts it contains. Some researchers also divide this type into synchronous and asynchronous awareness. Synchronous is the understanding of what co-workers are doing, their availability and related things at the moment. Asynchronous is the possibility to understand when an artefact has changed, by whom, when and in what way.

Social science researchers, more specifically sociologists and psychologists, have their definition of awareness that differs slightly from CSCW's community. They divide it into group awareness, social awareness, task-specific awareness, situational awareness and objective self-awareness. **Group awareness** is defined as "a specific set of behaviors as characteristics of intimate, primary groups and maintains that these behaviors will occur more often in those groups that have attained an enhanced level of (the group's) self-awareness" [Bar91]. **Social awareness** is an essential component when empathizing with another. It is considered the ability to take the perspective of another. **Task-specific awareness** can be identified when someone is able to adequately describe a used strategy and create a detailed report on the difficulties in understanding the task. **Situational awareness** allows the decision-makers to function, and it is "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" [EB94]. **Objective self-awareness** is when we become conscious of ourselves when we are the focus of our attention.

Comparing CSCW concepts with those coming from social science, we can identify similarities and differences. Both group awareness and social awareness match the two areas, and workspace awareness can be compared to situational awareness. On the other hand, no connections can be found in informal awareness and group-structural awareness of CSCW and task-specific awareness and objective self-awareness of social sciences.

Interpreting the definitions of social sciences can help users feel more comfortable with applications in future developments, so Tom Gross, Chris Stary, and Alex Totter [GST05] propose some means to grant the user each kind of awareness established in social sciences, as presented in Figure 2.9, achieving a better user experience.

Type of Awareness	*Recommended Means*	*Expected Benefit*
Group awareness	Provision of information for task completion	Establishment of transactive memory system facilitating collaboration and increasing group satisfaction (particularism)
	Provision of simultaneous source of information	Symmetry of awareness
Social awareness	Visualization through diagrams or symbols (2D, 3D)	Provision of information about one's own appearance enabling feedback on the effect of the appearance on other group members
	Context-sensitive feedback mechanism	Reflection on communication rules through feeding back on one's own or other experience from another group member
		Establishment of sense for community (informal awareness)
	Mutual sensing of presence	Respect of other's spheres or interaction spaces (peripheral awareness)
	Motivation for and control of engagement	Each member is heard and can be brought into the group
Task-specific awareness	Visualization: bar charts, flow diagrams	Mutual transparency of working process leading to increased understanding of group members
Situation awareness	Extrapolation tools	Provision of explicit projections into the future of collaborative activities leading to increased understanding of individual behavior of group members and allowing sound revision of individual expectations
	Feedback provision at an informal level	Improving the sense of community
	Presentation of group member's location	Improving the spatial feeling of community (location awareness)
Objective self-awareness	Mirroring video camera	Increased awareness of oneself leading to accurate support of self-reflection

Figure 2.9: Suggestions in awareness improvement. Adapted from [GST05]

Virtual environments bring both advantages and the need for research in different notions. The idea of **spatial awareness** is highly relevant when it comes to these systems. A scenario that exemplifies the privileges of spatial awareness is when someone says, "Look! behind you!". This simple sentence gives the ability to identify that something is happening behind. However, the affirmation not only gives the person the information but also influences him or her to look in the direction indicated; it is called **attentional orienting** [Pos80].

> “Attentional orienting improves visual research and guides the learners to the important elements. Therefore, it could improve their indexing in memory, reduce the cost of processing and the cognitive load, improving the transformation of declarative information into actions.“ [SHP19]

Attentional orienting can be done with the help of different techniques. **Visual guidance** is one of these techniques that can come in numerous forms. Arrows, paths, ripples and targets were used by Samuel Cosgrove Jr. and Joseph J. LaViola Jr. [CL20] to test which was better to give orientation through an explorable 360 VR environment. Figure 2.10 is a representation of these visual guides. The faster the users found the objective in the virtual world, the better was considered the orientation. Paths resulted in being the most efficient method, targets and ripples were identical, and arrows were considered the worst. Nevertheless, the authors suggest that hybrid combinations like arrows and targets can be efficient by simultaneously showing exact location and direction.

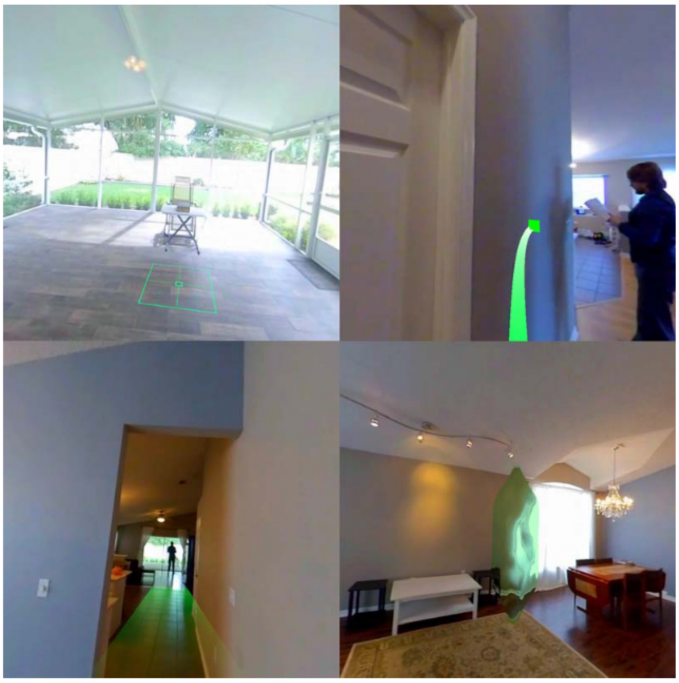


Figure 2.10: Visual guides. Top-left: Target; Top-right: Arrow; Bottom-left: Path; Bottom-right: Ripple [CL20]

AutoPilot is another focus assistance tool. It takes the viewer directly to target without the help of visual guides or the user's physical effort turning the neck or moving the controller device. Yen Chen Lin et al. [LCH+17] compared AutoPilot with Visual Guidance and got results favouring the first one. However, in some situations, the freedom granted by the visual guidance system was too significant and valued by viewers. It is proposed in the article that in high-level systems, a hybrid style is implemented where users have the possibility to choose what they prefer. Also related to visual guidance, the writers suggest that the cues should be customizable by the viewers in terms of speed, size, colour and so on.

Visual techniques are great attention orienting tools; nevertheless, other senses can be explored to obtain the same results. In an attempt to study the enhancement of visual perception, the authors stated, "perceptual sensitivity of subthreshold masked visual stimuli was indeed improved by concurrent acoustic stimuli." [FBL02]. In other words, just like the visual instructions, **sounds** can and should be used as cues to guide the user spatial orientation.

2.5 Summary

In this chapter, immersion is discussed as an inherited characteristic of VEs and how it influences the feeling of presence for a better user experience. To help in this advantage of immersion, VEs have their set of hardware devices, both input and output devices.

More specifically, in the VEs, 360-degree videos provide users with an experience of immersion where they have the freedom to explore the view all around them. Studies concluded that audience engagement is higher when the suitable elements are taken into consideration. These elements can vary from physical conditions like the display device and resolution of the screen to the video itself, how it is filmed, viewport dynamic, the bitrate, stalling effect, and the emotions it creates.

Emotions can be intensified when shared with other humans. Collaborative systems have been around for a while, and after many attempts to define them, a straightforward categorization is made with the help of the notions of synchronous, asynchronous, remote and co-located. Virtual environments are suitable to include collaboration; at least most of them are, AR is considered very challenging and complex to implement for the technology that we have available at the moment. Nevertheless, challenges appear when we try to add multi-viewer functionalities to a 360-degree video, for example. Learning from typical 2D collaborative video viewers implemented in the past, features like pause, go backwards, go forward, change the video's speed, communicate through a voice or text channel and even create live annotations are shared. This is not enough to mitigate all the problems presented when sharing the view in 360-degree videos, so other solutions and traits like coloured cones that indicate where each user is looking and enabling one collaborator to synchronize the view of all users were tested.

Awareness, similar to emotions, have a significant impact on the user experience. Different types of awareness have been researched and compared to the meaning of awareness in social sciences. Multiple techniques to improve a collaborator's awareness can be implemented in a software system. Virtual environments bring one more style of awareness compared to other software, spatial awareness. It is suggested that collaborative systems, when combined with virtual environments, implement audio effects and visual guidances and/or autopilot.

From the investigation in this chapter, we can conclude that collaboration in 360-degree videos is an area that lacks research. Numerous features employed in collaborative systems can be applied in these types of videos. However, a good system is built with the proper techniques, with the goal to augment the feeling of immersion, presence and awareness of the viewer. A solid understanding of collaborative systems and virtual environments helps us define valid requirements to enhance the user experience when designing collaborative interactions in immersive 360-degree videos. This chapter prepares us to approach the proposed work described in chapter 3 with a wider background about the main areas of the project.

Chapter 3

Proposed system for a collaborative experience with 360-degree videos

To present answers for the research questions established in section 1.2, and to complete the objectives of the dissertation, we propose the implementation of a collaborative immersive visualizer so that tests are performed over it. Learning from the previous research done on virtual environments and collaborative systems introduced in chapter 2, we provide an insight into what is expected to be implemented in this chapter.

We start with a general description of what was proposed in this work, delivering an overview of the main goals in section 3.1. Secondly, an explanation of the role of the AV360 project in the software proposed is clarified in section 3.2. Afterwards, some possible user scenarios are described in section 3.3. These user scenarios help us create the requirements for the system, represented in section 3.4. The features proposed are explained in section 3.5 and the plan for its implementation is exposed in section 3.6. In the following sections 3.7 and 3.8, the visual design and system architecture are proposed correspondingly. Finally, section 3.9 presents a summary closing this chapter.

3.1 General Description

As reviewed in chapter 2, the implementation of collaborative features in 360-degree videos may increase users' quality of experience. As stated previously, to answer the research questions, we propose the development of an immersive video visualizer with the addition of collaborative techniques. A first approach idealizes the ability to develop collaboration for different devices to increase accessibility: desktop, mobile and VR headsets. Different devices have different limitations, which creates the need for an adaptation of the features according to the hardware available. For example, the mobile devices in VR mode do not have access to a keyboard, making it hard for users to write. Two or more users should be able to connect and share the experience with the help of several collaborative techniques. The main focus is on synchronous collaboration, where the users have a shared timeline controlled by all. However, asynchronous collaboration is also a possibility. All the techniques that imply collaboration between the users are described in more detail in section 3.5 and explain their primary goal, having the research from chapter 2 as a basis. In the following section, we present a brief explanation regarding how we will take advantage of the previously developed 360-degree visualizer for the AV360 project.

3.2 AV360 context

Chapter 1 mentions that the AV360 project is the starting point for the application that is planned to be developed. The project focuses on 360-degree content and tools. During several iterations, a visualiser of 360-degree videos was created and improved for the AV360 project. The project also includes an editor for 360-degree videos. This editor allows the creation of different kinds of annotations in the videos to guide the user throughout the experience. The visualizer, as illustrated in Figure 3.1, displays these annotations and has some features already developed:

- Sound management
- Full-screen mode
- Radar
- VR toggle mode
- Basic voice commands while in VR mode
- Interaction to rotate the scene in several ways. Some of them are: touch and drag in mobile, mouse control by dragging and arrow keys in desktop, and rotating the device around in VR mode.

Figure 3.1: AV360 visualizer example

Using the tools that AV360 provides as groundwork for our research, we intend to add collaboration to the visualizer. The AV360 project is web-based, which implies some conditioning in the development of the collaborative features such as the tools. In the following sections, we discuss what is proposed to be developed in more detail.

3.3 User Scenarios

To gather the main requirements and features for an application of this nature, we first started by visualising and describing some system usage scenarios.

The user scenarios can be divided into two main categories. The first one contains field trips, touristic tours, medicine class and football games revisions. This first type of scenario has users with different roles watching the video, and we call it hierarchical scenarios. Users like professors have access to more functionalities, like muting and removing others from the room. In the second category, non-hierarchical scenarios, all the users have the same roles. Despite the focus of this work being in the non-hierarchical scenarios, specifically in a group of friends watching a video together, scenarios from both categories were imagined in the planing of this work. One example scenario of each type is presented below. Other more extensive scenarios can be found in the appendix A.

Hierarchical Scenario - Medicine Class

Scenario:

Students take a medical class where a specific procedure is learned through a 360-degree video.

The motivation of the users:

Students want to learn and question what is displayed in the video. The teacher is present to explain the environment and critical points in the video.

Example:

Students will watch a 360-degree video of a successful heart transplant. Some students are together in the classroom, and others are in their respective homes.

The professor prepares the video and shares the room's link with all the students.

When the professor checks that everyone is ready, he starts the video and explains every step.

In the end, there is time to question the professor, and students can come back in the video to review specific parts.

Specific interactions:

- The professor verbally communicates with all the students to explain what is being reproduced in the video.

- A student with a question raises the hand emoji to have the professor's attention.
- The professor rewinds the video timeline to explain the operation's details.
- While asking a question, a student uses a visual beacon to point to a specific place in the video and direct the other users' attention.
- A student has difficulty keeping track of what the professor is referring to in the 360-degree scenario. Therefore he changes the view of their screen to the professor's view.
- The professor notices that the group's attention is not in the right place by checking the radar and corrects them.
- The professor uses the compass directions displayed in the radar to orient the students to a specific area.
- The professor mutes one, some or all students to avoid disturbances.
- The professor removes one or more students from the room because they are disturbing the class.
- The professor forces the students' attention towards an area of interest with the help of a vignetting effect which manifests in the field of view of the students.
- The professor slows the video down to explain a detail happening in the operation.
- The professor pauses the video and draws in the paused image to explain a student's question adequately.

Non-hierarchical Scenario - Entertainment

Scenario:
Six friends hang out remotely and decide to watch a 360-degree video together.

The motivation of the users:
The six friends just want to entertain themselves and hang out with each other.

Example:
The users will watch a 360-degree video related to skating. All of them are in their respective homes.

One of the friends creates a room and shares the room's link with the others.

When they agree, one of them starts the video.

Specific interactions:

- The users talk and laugh with each other.
- In the video timeline, one of the friends returns to a previous timestamp to point out something funny that he saw.
- One of the friends uses a ping with a reaction to react to a specific space in the video.
- One of the friends mutes himself due to a lot of background noise.
- One of the friends needs to be absent for some time, so another member of the group pauses the video for everyone to wait for his/her return.

3.4 Requirements

With the help of the user scenarios detailed in section 3.3, it was easier to collect requirements for stable development of the upcoming software system.

These requirements, which have the enhancement of the user's quality of experience as a primary goal, are divided into functional and non-functional requirements.

A degree of importance was defined for the **functional requirements**, from the most important to the least important: critical, important, useful, extra. It is crucial to notice that functionalities already implemented in the AV360 visualizer, like playing and pausing the video, are considered closed and not stated here. As described before, the AV360 project is web-based, automatically conditioning the requirements.

The critical functional requirements are essential for the system's basic functioning, even if not related directly to collaboration:

1. The user should be able to connect with other users remotely and connect a specific group of users.
2. The user should be able to see the video synchronously with other users.
3. The user should be able to instruct the program verbally, using voice commands.
4. The user should receive feedback regarding the commands he gives to the machine.

The important functional requirements are vital for the user tests that will help us answer the research questions proposed:

5. The user should be able to know where the others are looking at.
6. The user should be able to identify the other's field of view and differentiate each other in the VR environment.
7. The user should be able to hear and speak to other users.

8. The user should be able to mute and unmute himself.

9. The user should have information about their microphone: if it is on or off for other users.

10. The user should be able to direct attention to a specific point in the viewport of others.

The useful functional requirements are functionalities that could help gather extra information for future applications that intend to combine immersive systems with collaboration:

11. A user with higher permissions should be able to control the video timeline and deny others from doing so.

12. A user with higher permissions should be able to remove users from the experience.

13. A user with higher permissions should be able to mute and unmute others.

14. A user with higher permission should be able to force the view of the other users towards a specific area.

The extra functional requirements are functionalities that could only be developed for non-VR or, like the name says, act as extra accessories to complete already existing functionalities:

15. The users should be able to customize the colour that represents them in the radar and the displayed name.

16. The user should be able to know and control the volume of the remaining users' sound for his device.

17. While not in VR mode, the user should be able to desynchronize from the others. This is not available for VR because of the complexity that involves navigating through the video with voice added to the asynchronous commands, it would be too overwhelming for the user.

18. While not in VR mode, the user should be able to view the other users' view in a small window. This is not available for VR because of viewport size, such feature would occupy too much of the screen.

19. While not in VR mode, the user should be able to share a reaction with emojis or a sound clip. This is not available for VR because of the complexity of choosing an emoji or sound clip through voice commands.

20. While not in VR mode, the user should be able to communicate via chat with other users. This is not available for VR because of the lack of keyboard and space in the viewport.

21. The user should be able to ask for the video to pause without interrupting the other users verbally.

22. The user should be made aware of all available voice commands.

All these requirements were afterwards transformed into a set of features. The correspondence can be seen in the appendix in Table B.1 for critical requirements, Table B.2 for important requirements, Table B.3 for useful requirements and finally Table B.4 for extra requirements. Note that most of these features can be developed for desktop, mobile and VR. However, our primary focus is on immersive 360-degree videos, so the development of these features for Desktop and mobile without the VR mode is considered beyond the scope of this dissertation.

Besides the functional requirements, the following **non-functional requirements** were established:

- The user interface should be easy to operate.
- The user interface should have a quick and smooth response.
- The user interface should be intuitive.
- The system must be scalable.
- The system must be compatible with different hardware.
- The system must be easily maintainable.
- The system must have high availability.

3.5 Features

During the visualization of a 360-degree video, numerous events can be happening all around the scene. For example, in the second scenario present in the user scenarios section, watching it with a friend or a group of friends can be enjoyable when everyone is participating and pointing out funny or different points of interest that they found in the video. However, if the display of the information shared is not controlled, the interaction can become confusing and overwhelming. The opposite scenario can occur, where the collaboration between users might become too limited, and the interaction may not reach its full potential. Therefore, to create a balance in the information available, the following subsections describe interactions that were considered to be explored and implemented in this project.

Sharing information with others is a sensitive topic, and some users might not be comfortable with all the forms of communication described here. Taking this into consideration, when developing an advanced system that implements these interactions, the users should have the possibility to deactivate any interaction through a settings menu. If the communication does not have a way to be deactivated, the users should be warned of it before using the system.

The features were divided into direct communication, indirect communication and personal features. For every features it is assigned a priority degree and the corresponding requirement that helped with the feature definition. Besides these ways to group the features, some are identified as **spatial orientation features**. These features help with the user spatial orientation in relation to others and are tagged as such in their description.

3.5.1 Indirect communication features

Indirect communication allows users to constantly and unconsciously share information about themselves without any effort.

- **Radar**: With the spherical view that 360-degree videos provide, giving directions to localize a specific point of interest can be challenging. A radar with cardinal points indicating north, east, south and west mitigates this issue. The directions should be implemented in a way that the north points to the initial angle of the video, that is previously defined by the content creator. Cones with unique colours drawn in the radar represent each user's field of vision. **Tags:** Important feature, Requirements 5 and 6, **Spatial orientation feature**.

- **Mini-view**: As reviewed in chapter 2, autopilot can be beneficial in some situations. Mini-view is a miniature view of another collaborator chosen by the user that is displayed on the screen, over the 360 video. This miniature can be expanded to full screen. While expanded it is like being in the described autopilot state. **Tags:** Extra feature, Requirement 18.

- **King of the room**: In section 3.3, we introduce the concept of hierarchical scenarios. The king of the room is the feature that creates the hierarchical difference and opens the path for other features that depend on the different types of users. **Tags:** Useful feature, Requirements 11 and 12.

- **Personalize display and profile**: Changing the profile is destined to help the users distinguish from each other and feel more attached to the system because they customize it in their own way. **Tags:** Extra feature, Requirement 15.

3.5.2 Direct communication features

Direct communication allows users to actively and consciously share information with other collaborators. The immoderate usage of this type of features by the users can be disturbing and lower the quality of experience. Considering the possibility of spam, its excessive usage should be blocked.

- **Voice communication**: Voice can be a powerful tool to share information. Human beings evolved to communicate vocally, and it is one of our main ways to share information with others. The users should control their sound volume and individually change the other's volume just for their devices. **Tags:** Important feature, Requirements 7, 8 and 9.

- **Text chat**: Some situations, like being in a library, are not favourable to use the voice chat, or a participant might not have a working microphone. For this reason, for users that have a keyboard available, text chat is helpful to include them in any situation. **Tags:** Extra feature, Requirement 20.

- **Reactions**: Emojis shortcuts can be used to quickly express feelings and help users share their experience without interrupting the video. Furthermore, simple shortcuts to produce sound clips are also a reaction to add extra interaction between viewers. **Tags:** Extra feature, Requirement 19.

- **Synchronized view**: Users can start the video synchronously but can also stop it, go forward or backwards without becoming asynchronous. **Tags:** Critical feature, Requirement 2.

- **Ask to pause**: Besides the shared basic video controls, "Ask to pause" grants users who feel uncomfortable interrupting the video access to a shortcut button that notifies the other users to pause the video if or when they feel it is suitable. **Tags:** Extra feature, Requirement 21.

- **Point of interest share (or Ping users)**: The user will create a visual beacon on the view of the other participants. This has the intention of helping to identify small details in the display. **Tags:** Important feature, Requirement 10, **Spatial orientation feature**.

- **Link room share**: To connect to the same video and synchronize, the users share a link provided by the system. **Tags:** Critical feature, Requirement 1.

- **Tunnel Vision**: Depending on the "king of the room" feature, tunnel vision is destined for users with high permissions, and it forces the other users' vision to a specific point in the scene by blurring the surrounding area. **Tags:** Useful feature, Requirement 14, **Spatial orientation feature**.

- **Mute and unmute users**: Just like the "tunnel vision" feature, "mute and unmute" users is destined for users with high permissions and allows them to mute and unmute other users that are lower in the hierarchical level. **Tags:** Useful feature, Requirement 13.

3.5.3 Personal features

Personal features, don't create interaction or impact others, besides the user that is using them.

- **Users volume**: Allows a user to control the sound volume of the others individually, balancing the sound and improving the quality of experience. **Tags:** Extra feature, Requirement 16.

- **Individual mode**: Contrary to the "Synchronized view" feature, it allows users to desynchronize and navigate the video on its own while in the same room as the other users, not disturbing the shared video timeline. **Tags:** Extra feature, Requirement 17.

- **Voice recognition**: It is how users interact with the system. It recognizes speech and interprets the given commands. **Tags:** Critical feature, Requirements 3 and 4.

- **Voice commands menu**: Informs the users about the available commands to use with the voice recognition feature. **Tags:** Extra feature, Requirement 22.

3.5.4 Summary of features

This subsection presents table 3.1 with all the proposed features ordered by importance. The table has two other columns besides the "Importance". One is the "Type", indicating the type of feature according to the previous subsections and the last one is "Others" that displays the missing information regarding the feature's tag.

Features	Type	Importance	Others
Link room share	Direct communication	Critical	Req. 1
Synchronized view	Direct communication	Critical	Req. 2
Voice recognition	Personal	Critical	Req. 3 and 4
Radar	Indirect communication	Important	Spatial Orientation feature; Req. 5 and 6
Voice communication	Direct communication	Important	Req. 7, 8 and 9
Ping users	Direct communication	Important	Spatial Orientation feature; Req. 10
King of the room	Indirect communication	Useful	Req. 11 and 12
Mute and unmute users	Direct communication	Useful	Req. 13
Tunnel vision	Direct communication	Useful	Spatial Orientation feature; Req. 14
Personalize display and profile	Indirect communication	Extra	Req. 15
Users volume	Personal	Extra	Req. 16
Individual mode	Personal	Extra	Req. 17
Mini view	Indirect communication	Extra	Req. 18
Reactions	Direct communication	Extra	Req. 19
Text chat	Direct communication	Extra	Req. 20
Ask to pause	Direct communication	Extra	Req. 21
Voice commands menu	Personal	Extra	Req. 22

Table 3.1: Proposed features ordered by importance.

3.6 Development planning

Considering the time needed to execute, analyse and discuss the user tests, and the dissertation delivery date, a plan was created to organize the development stage. It was adopted a form of informal agile development through sprints. Each sprint had new tasks. Besides these new tasks, informal tests were done to find issues in the recently implemented features, and if any issue was found, it was fixed. It was defined that each sprint had a duration of 1 week.

The development of the features was planned to start with the highest priority ones and proceed to the following. If the implementation took less time than expected, low priority features were also supposed to be developed. The plan was the following:

- **Sprint 1**: Implement the Critical Features.
- **Sprint 2**: Implement the Important Features.
- **Sprint 3**: Implement Useful Features or use this sprint to improve Critical and Important Features if needed.
- **Sprint 4**: Implement changes for user tests.
- **Sprint 5**: Fix any final issue and prepare the deployment of the application.

3.7 Visual design

The quality of the user experience is our main way to evaluate the final system and gather results. However, numerous things can impact the user experience. Our goal while designing the visual structure of the system was to make it the most intuitive and accessible, while also enhancing the user experience. The reason for this concern was to facilitate the user's learning process of the application so that this factor would negatively influence as little as possible the final results. The visual design was created in mockups that also helped with the planning of the system.

All the different iterations of them can be found in Figma[1], an online designing tool. However, Figure 3.2 illustrates some screens from the final iteration of the mockups:

- Screen A is the desktop view. The others are in VR mode.
- Screen B displays a notification of a change in the video timeline for plus 10 seconds.
- Screen C displays the application giving feedback regarding the user's speech. The detected speech in the example is: "Best project".
- Screen D displays a selection menu.
- Screen E displays the voice commands menu.
- Screen F displays the crosshair from the "Ping users" feature being used.
- Screen G displays a user changing the video time.
- Screen H displays a user changing the general volume.

The mockups were created to be semi-interactive which made them closer to the real implementation. These mockups went through preliminary usability tests, and several versions were created out of users' opinions. These tests were done informally and one main goal was to make the system interaction with the voice commands menu more intuitive. The primary version's first main change was related to the position of the visual elements, like the radar. Initially, these elements were too close to the borders. While using a VR headset, elements on the side of the screen are harder to see and require more effort in moving eyes to an uncomfortable angle. Finally, the second main change and challenge along the various versions was related to the feature allowing users to change the volume. It was not intuitive, and users kept giving wrong voice commands until the last iteration that had a high success rate with non-experienced users.

Even though all the screens went through these preliminary tests, not all of them were expected to be implemented in time of the dissertation conclusion. Low priority features like the desktop implementation or voice commands menu representing most of the figure's screens (A, D, E, G, H) were not developed in the final prototype. However, its planning and design can be used for future developments.

[1]URL to access the mockups: https://www.figma.com/file/Pms4JnAQKsJEepWSFtBbVo/Theses?node-id=620%3A1075

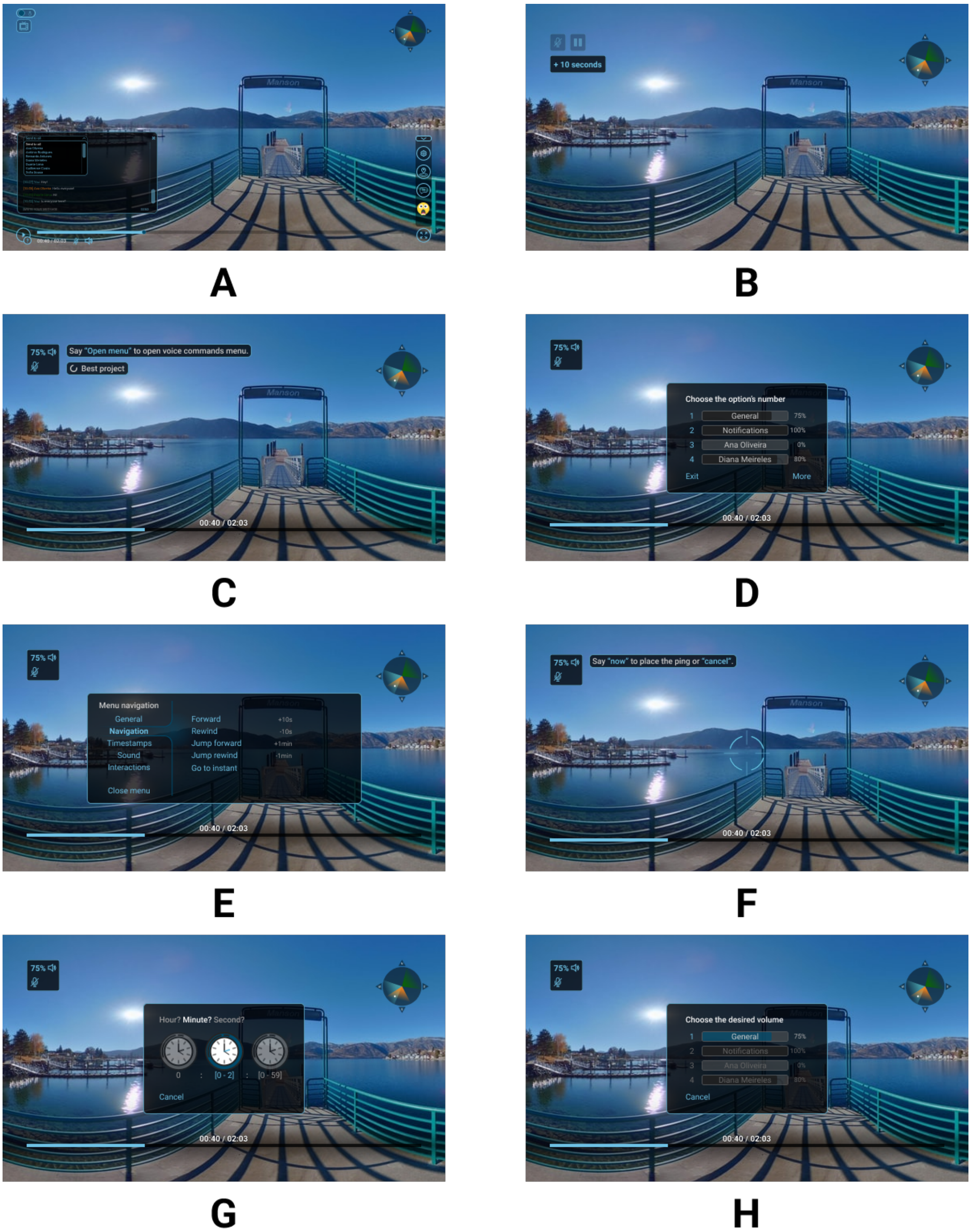


Figure 3.2: Design mockups

3.8 System Architecture

In this section, we present the initial idea for the software architecture based on the requirements established previously.

Due to some challenges during the development phase, two different iterations were proposed for the system architecture. The objective of both of them is clear: the system should find a way to make different clients communicate and use the AV360 visualizer as a 360-degree video player. To fulfil this objective, the first system architecture designed had a broker to make the clients communicate between themselves and a server that indicates them to the right room in the broker. Figure 3.3 illustrates this client-server software architecture with the existing AV360 visualizer represented in red.

The second iteration was thought during the development phase, after encountering some challenges when implementing the system. Opposite of what was proposed in the first architecture, it is composed of a peer to peer data exchange. This second option contains a server that coordinates the clients to their respective rooms and a connection broker that manages the peers' connections. The data exchange is done directly between peers through a connection that is initially created with the help of the connection broker. More details about it are detailed in section 4.2.

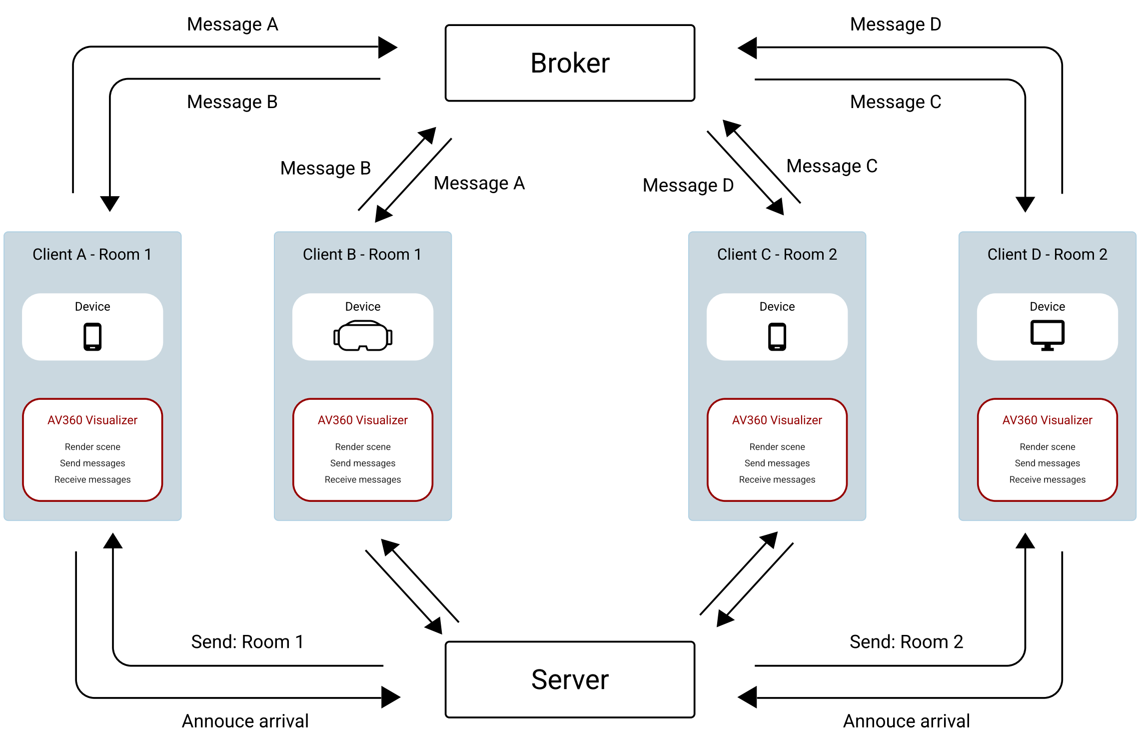


Figure 3.3: Client-server software architecture.

3.9 Summary

This chapter starts with a general description of the proposed system, afterwards, it frames the AV360 project and the purpose of its visualizer in the prototype that is planned to be developed. It is followed by the presentation of some possible user scenarios, essential to identify the requirements that helped us define the features for the software system.

The planning of the development the application's visual design are proposed, and some examples are displayed. Finally, the system architecture expected to be developed is explained. The software proposal serves as a guide for the development described in chapter 4. The next chapter describes the development process, the used tools and technologies, and the implemented system itself.

Chapter 4

Development of collaborative interaction for 360-degree videos

Chapter 3 is the proposal of a 360-degree video visualizer that allows users to watch the same video synchronously. This software system and its development process is described in this chapter.

The implemented system starts by presenting the main page that allows the user to create a new room or access an existing room. Once the user proceeds from the first page, the visualizer is displayed with the video ready to play. From that moment on, the users can watch the video with whomever they desire, and anyone who joins the same room will catch up with the latest state of the video. While watching the video, the users on a computer can enjoy the availability of voice communication. Besides that, features described in the previous chapter, like the collaborative radar and the point of interest share, are also available to all the users.

The system itself and its development is described in this chapter. It starts by presenting the overall system, its usage and general workflow, in section 4.1. Section 4.2 introduces the software architecture and the technologies used. Afterwards, in section 4.3 a more detailed view of the communication implemented is presented with the explanation of the collaborative features developed. All the challenges and solutions found during the implementation of this system are described in section 4.4. This software serves as a tool for research, so some modifications were done to the system for testing purposes, and they are described in section 4.5. Finally, a summary presented in section 4.6 points out the main conclusions of the development phase and closes this chapter.

4.1 System usage and general workflow

This section introduces the developed system, explaining the workflow and how to use it. The first contact of the user with the application was also the first feature developed: the "Link Room Share", that allowed users to connect with each other. To construct this feature, it was necessary to assemble the server, the client-side of the communication, and a main page for the web app. As exhibited in Figure 4.1, the main page gives the user two options: create a new room or access an existing room through a code. After the users choose an option, they are redirected to a room where the 3D scene is rendered, and a default video with ballerinas dancing is presented.

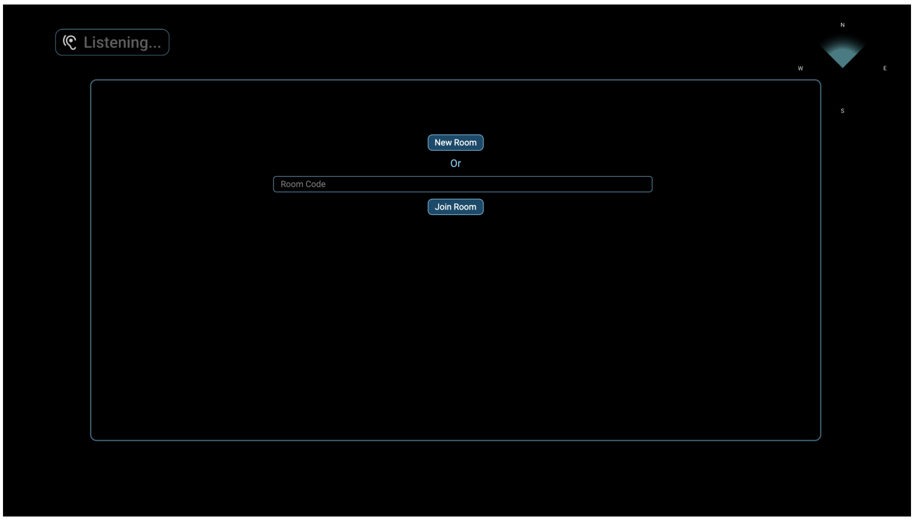


Figure 4.1: Illustration of the main page

When the users joins a room, they will catch up to the video state. For example, if the video is already playing in 1 minute and 20 seconds, they will jump directly to that time and start playing automatically. If it is the first user in the room, the video is always paused in the beginning waiting for instructions. From there on, every action taken by that user influences the others; they are synchronously connected. This is the **Synchronised view** feature that was implemented.

When the client joins the video room, they will be able to perceive the direction where the other users are watching. It is displayed in the **radar** feature where the field of view of everyone connected to the room is displayed on the radar on the top right of the screen by a cone of a unique colour. It is illustrated in Figure 4.2.

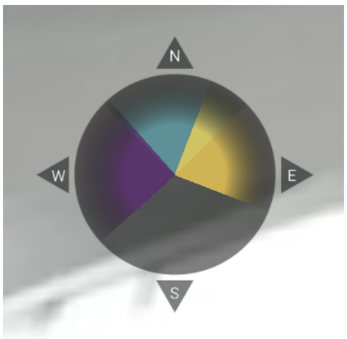


Figure 4.2: Radar with 3 users connected.

To interact with the system in VR mode, the users use the **Voice recognition** feature. The users pause the video with two taps on the mobile device that detects the motion and stops the video. When the video is paused, the microphone that transforms speech to text is activated. Once the microphone is on, they can give the following commands to the system:

- "play" to reproduce the video.
- "forward" or "skip" to skip 10 seconds of the video.
- "rewind" or "back" to rewind 10 seconds of the video.
- "jump forward" or "jump skip" to skip 1 minute of the video.
- "jump rewind" or "jump back" to rewind 1 minute of the video.
- "target" or "ping" to activate the ping users feature.
- "louder" to increase the volume of the video.
- "softer" to decrease the volume of the video.
- "mute" to mute the video.
- "unmute" to unmute the video sound.

During some informal tests performed during the development of the speech recognition feature, we noticed that the strict British English pronunciation of the Web Speech API can be hard to achieve for non-native speakers. Taking that into consideration, more than one keyword was established for the same command, for example: "target" and "ping" trigger the same feature because the word target can be easier to pronounce for some users.

The voice recognition feature is directly related to the human-computer interaction component of this project. We created voice recognition feedback to achieve good system performance when interacting with the user. This feedback is displayed in the top left of the user's screen, as illustrated in Figure 4.3. In the Figure, the system interpreted the command "Louder". The feedback message can be in one of three states:

- Placeholder state: Displays "Listening..." when the system is listening to the user speech; Displays "Playing" when the video is playing; Displays "Paused" when the video is a pause.
- Recognized state: Displays the command recognized.
- Error state: Displays "Can you repeat?" when the system could not associate the user speech to a command.

The **Ping Users** feature was also implemented and, as described before, is triggered by the "target" or "ping" voice command. It allows a user to share an interesting point in the video with the others with the help of a widget, as illustrated in Figure 4.4.

Following on from this general section, the following ones detail the technical aspects of the implementation, starting by describing the system architecture and technologies used.

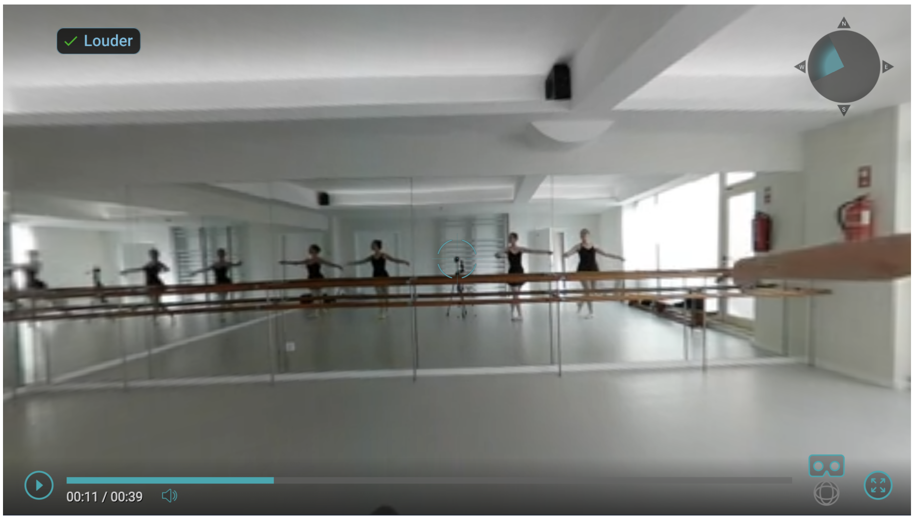


Figure 4.3: Example of the the voice recognition feature recognizing the command "Louder".

Figure 4.4: Example of a ping from the yellow user.

4.2 System Architecture & technologies

The main difference between the system developed for this research and the last version of the AV360 software is the collaboration between users, which means that the main focus is the implemented communication system. The initial proposal of this communication was through a broker. A Message Queuing Telemetry Transport broker, or just MQTT broker, was used to make the clients communicate between themselves and a server connected by a WebSocket to the clients was also implemented and used to indicate the clients to the right topic in the broker. Topics are the rooms where the users would communicate.

Among the messages, encoded MP3 files were to be exchanged between users to create the possibility for users to communicate live through voice. After some analyses, it was noticed that the data transiting through a broker before reaching the clients is not the best option for the objective planned to achieve. The new solution swapped the exclusive client/server architecture to a peer-to-peer connection. This way, the communication is made faster without going through a server that can potentially accumulate traffic and consequently add more delay in the messages. However, there are downsides to this approach, as each client has now to deal with more than one connection. If N users are in a room, each client must create at least N-1 connections. This amount of connections can become heavy and create the need for better download and upload bandwidth. It is not expected to be a problem for the practical usage of the software in this research because, as explained in chapter 5, we expect to have only 2 to 5 users connected simultaneously, which is not a significant amount, especially with the low-sized messages sent.

The system is composed of more than one service and technology, as illustrated in Figure 4.5. **WebRTC**[1] is an API that provides web browsers and mobile applications the possibility of real-time communication without the need for plugins or the installation of native applications. It allows peer-to-peer communication. Our system uses **PeerJs**[2], an API that tries to find the fastest path between peers and simplifies the usage of WebRTC. This is how we construct peer-to-peer communication, exchanging messages and streaming media between clients.

Node.js[3] is used together with Express. Node.js is an open-source server environment with built-in mechanisms that can handle multiple incoming network connections of the system. **Express**[4] is a reasonably famous framework used over Node.js to manage the incoming connections. These technologies are used in a server that functions as a client coordinator. We consider this the **main server** and refer to it just as the server. It informs the clients about the arrival and departure of other peers to the same room and their respective id so that the already attending peers can connect or disconnect to them. The way this server connects to the clients is through WebSockets. More specifically, we use **Socket.io**[5], an API that makes it possible to open a two-way interactive communication session between the client and the server.

[1] WebRTC main page: https://webrtc.org/
[2] PeerJs main page: https://peerjs.com/
[3] Node.js main page: https://nodejs.org/
[4] Express main page: https://expressjs.com/
[5] Socket.io main page: https://socket.io/

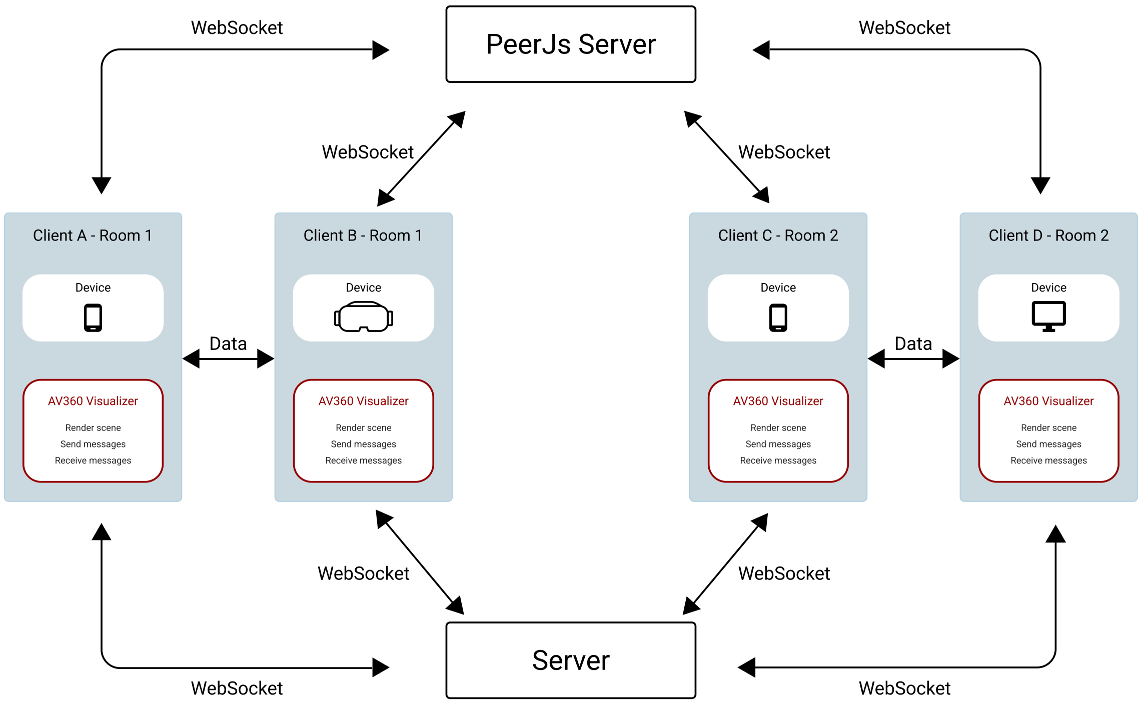


Figure 4.5: Final system architecture exemplified with 4 clients.

The architecture also has a **PeerJs server** that serves as a **connection broker**. It is different from the previously planned broker, it does not interfere with the data shared between users and only manages the pool of connections, linking each peer through their id and enabling rapid reuse of these connections by short-lived processes without the overhead of setting up a new connection each time. Meaning that when a connection fails, the PeerJs server handles it and reestablishes it. The data itself is directly exchanged between clients in the same room without the need for any other service.

While being used, the system can be divided into three phases:

- **The initial connection:** The first phase consists of a client-server architecture where the clients communicate with the servers to connect with other clients.
- **The information exchange**: The second phase forms a peer to peer architecture where the peers are linked with each other and communicate directly.
- **Failure and reconnection**: The third phase is a client-server architecture that occurs when a connection fails, and there is the need for the PeerJs server to interfere and help the clients reconnect with each other.

Besides the previously described technologies, the ones from the av360 project were maintained and reused in the client's side (represented in the figure by the red outline). **Three.js**[6], is a

[6]Three.js main page: https://threejs.org/

JavaScript library that was used to create the scene that displays the 360-degree video. To switch to virtual reality, the AV360 project uses **WebXR**[7], an API that connects the 3D scene created by Three.js with the VR headset. Finally, React and Webpack are also reused. **React**[8] is a JavaScript library used to build the video editor interface, and **Webpack**[9] is a module bundler that compiles the source code files, as well as libraries, into an optimized bundle that can easily be shipped to the end-user.

The following section explains how each collaborative feature was implemented using the described technologies and architecture.

4.3 Specification of the implemented collaborative features

The current section details each collaborative feature implemented and how the messages are exchanged between peers to achieve an immersive video synchronous and shared experience. The first subsection 4.3.1 explains the critical "Synchronized View" feature, and the following 4.3.2 and 4.3.3 describe the development of the spatial orientation features essential for the study of RQ3.

4.3.1 Synchronized View

The synchronized view feature is achieved through a set of messages exchanged by the system in several keypoints of user interaction. First, the client connects with peerJs and receives its own peer id. Afterwards, a signal is sent to the server informing its id. The server informs all the other clients in the same room about the new user. From there, each of the clients already connected to the room uses the help of the PeersJs server to create a new connection with the new user. Once the connection is established, all the users send a "greetings message" to the new one. This message contains three important pieces of information which are: the colour of the peer that is sending the message so that the new one does not use the same colour; the current time of the video so that the new user can update and sync with the others; the state of the video so that the new user knows if the video is supposed to be playing or paused. Every message sent between users has a similar format that includes the message type, timestamp, and corresponding data. An example of the greetings message is illustrated in Figure 4.6. At this point, the users are supposed to be synchronized.

Considering different internet speeds and the usage of different devices, we assume that some users can experience a delay in the video. For this reason, we resync every user when one of the following actions is taken: play, pause or the video time is manually changed. The resync is done in relation to the user with less time of video seen so that no one loses segments of the video. The peers that take action send in the action message their current time. Each receiving peer does one of two things: they either update their time if the sending peer has seen less time of

[7] WebXR main page: https://immersiveweb.dev/
[8] React main page: https://reactjs.org/
[9] Webpack main page: https://webpack.js.org/

```
messageObject = {
  type: messageType,
  timestamp: timestamp,
  data: {
    color: color,
    videoTime: videoTime,
    isPaused: isPaused,
  },
};
```

Figure 4.6: Greetings message example.

the video or send a message with their time for the other users to update. Figure 4.7 illustrates a sequence diagram of User 2 joining User 1 in a room and resyncing once User 1 pauses the video. In the example, we notice that User 2 falls behind in the video timeline because there is a need to exchange the "Update time" message once he receives the User 1's time in the "Pause" message.

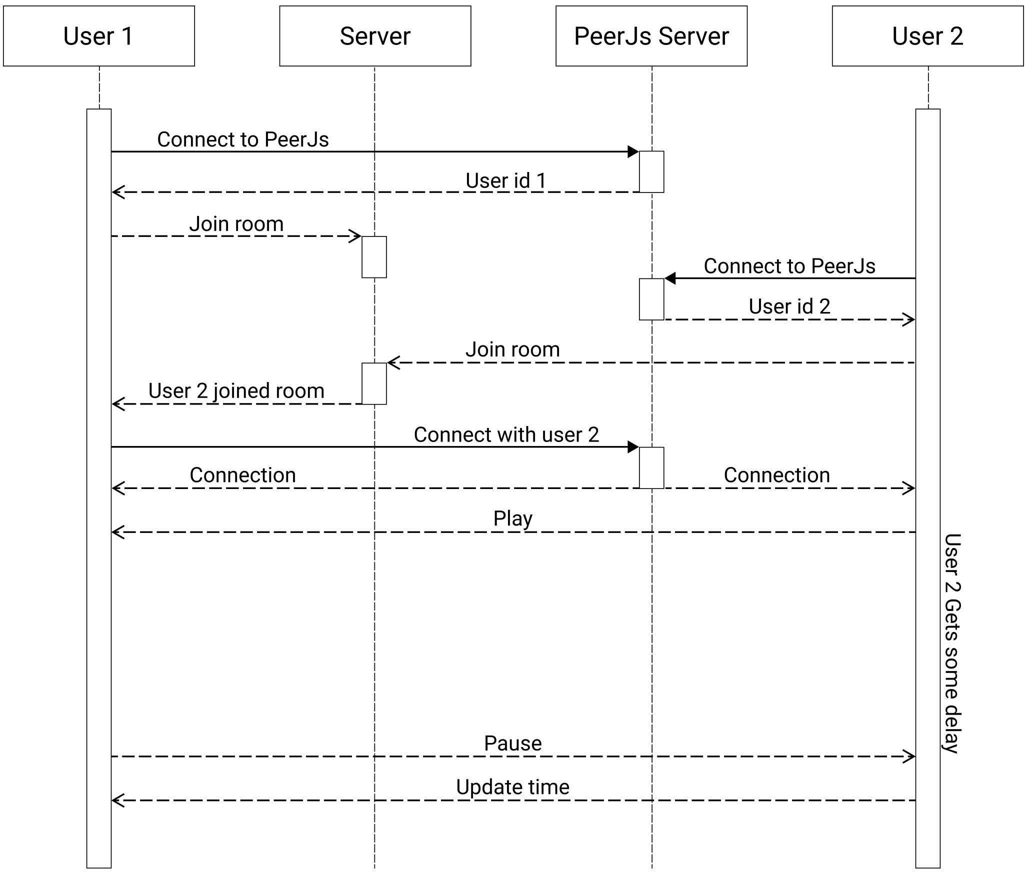


Figure 4.7: Sequence diagram of a communication scenario.

4.3.2 Point of Interest Share

The point of interest share feature is activated with the command voice "target" or "ping". Figure 4.4, illustrated in section 4.1 displays the screen of the purple user receiving a ping from the yellow user. Lets assume that yellow user is User 1 and the purple user is User 2. User 1 activates the target/ping voice command. After the activation of the command, two things happen:

- The point of interest widget (the yellow target displayed in the Figure) is rendered in the center of User's 1 field of view. Each user has a cross-hair when the video is paused that represents the center of the field of view and helps to aim. The ping is represented by a 2D plane in a shape of a square and its texture is an image of the white target illustrated in Figure 4.8. The system changes the colour of the widget's texture to be the same as the colour of the user that created it.
- A message is broadcasted to all the other users. The message contains information about the user's colour, in this case it is yellow, and the target quaternion coordinates. With this information, all the peers can recreate the targets in their scene. Besides the target, an arrow guide is rendered indicating the direction of the point of interest in case it is not in the user's viewport. The message sent adds one more piece of information, the distance to the camera. Its purpose is for future developments that might want to zoom in or out the ping. It is possible by sharing the target's distance to the camera.

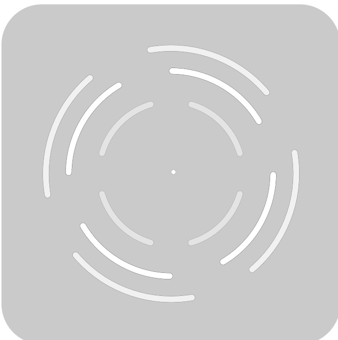


Figure 4.8: Point of interest default texture with grey background.

4.3.3 Radar

The radar feature is always present to the users, and the information is expected to be crucial for their orientation. This feature is reused from the AV360 project and is composed of different pieces. The main piece is the base mesh, a default circle texture with the compass directions for orientation. On top of this base, two different elements can be rendered. The first element is the light angle representing the user field of view that rotates over the centre of the radar base to adjust to the user's head movement. The second type of element is little dots representing annotations on the map. In figure 4.9 we can recognize each of these elements separated from each other.

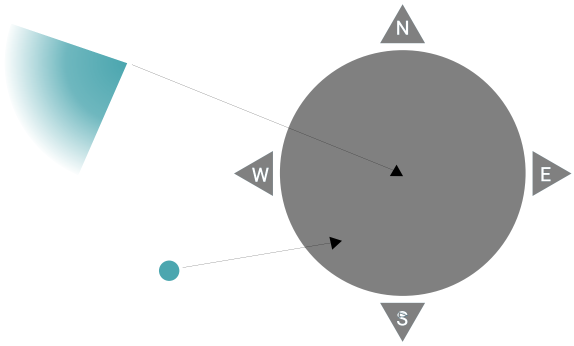


Figure 4.9: Example of a dismantled radar.

When engaging in a shared experience, the coordinates used to calculate the light angle are shared with the others. Once a connection is established between peers, they both proceed to send a "radar coordinates" message to each other every 50ms. It contains their respective colour, peer id and light angle. The light angle is the angle where the peer is looking in the scene. With this information, every user can recreate the other's viewport light angle in the radar with the respective colour. Different intervals between "radar coordinates" messages were tried, but 50ms was the longest time found before starting to feel lag in the other's viewport light angles.

4.4 Challenges and solutions

This section explains some of the challenges found during the development phase. Just like the change in the system architecture, these changes created an involuntary divergence between some aspects of the initial proposal and the developed system. Besides the challenges, we try to explain the workaround found for these issues and our thought process.

When the users are interacting with software, it must respond so that they know that the system is not frozen or stopped working. In our application, when a user gives a voice command to the system, he needs to receive some feedback regarding the command given. That is what we call "voice recognition feedback". If the system is not quick, it can be inconvenient and become a less fluid interaction. Initially, this feature was supposed to display all the words captured by the microphone and afterwards inform the user if the command was recognised or not with the help of colours and symbols, as illustrated in Figure 4.10. The system needed to process the speech, convert it to string and display it. Since the speech processing takes some time, the display should be as fast as possible. An easy and high-performance method would be to create an HTML element and change the innerHTML depending on the value of the speech. However, when using VR mode, only the Three.js scene canvas is rendered in a way that the image is doubled (one image for each eye), making the view confusing if there is not an HTML element for each eye. There are alternatives to present text in Three.js, but rerendering objects in the scene or recreating textures and changing them every time a command is received can be exhaustive and time-consuming. Finally, we concluded that feedback personalised for each speech might not be the most efficient way. It could be possible if the most common recognised expressions were saved in the cache, improving performance after the first usages. However, as described in section 4.1, we opted to provide simple feedback. The system has previously created and loaded textures for each one of the existing commands and other four default textures that have written: "Can you repeat?", "Listening...", "Paused" and "Playing". When a speech is recognised the software either interprets the speech as an existing command and displays the previously loaded texture or displays "Can you repeat?" if the command does not exist. Furthermore, the "Listening..." texture is used as a placeholder whenever the microphone is on and the user can interact with the system. Two other placeholders exist to keep the user informed: "Playing" and "Paused". It is rendered in the scene in a way that minimizes the impact on the performance. It renders a 2D rectangular plane object on the top left of the screen that is always facing the user camera and changes texture depending on the command received. All the textures are saved as default images, as illustrated in Figure 4.11.

Figure 4.10: Example of the initial prototype for voice recognition feedback

Figure 4.11: Example of the final textures for voice recognition feedback

The URL can be used as a more advanced tool to interact with our system. For example, the user can change the video by changing the link. The URL comprises the room code and the video desired to be displayed: https://WebsiteName.com/roomCode/videoId/FeaturesRestrictionId; the default video is selected if the video id is omitted. To facilitate the choice of the video on the main page, if the users decide to write "room1/2" on the room code section, they will be redirected to room1 that displays the video with id 2. As displayed in the example, the URL has an additional parameter named "FeaturesRestrictedId" which exists due to the inability to add the **voice communication** feature in all devices. For this parameter, id 1 turns on voice communication and can be used on desktop devices to communicate between users remotely. If there is no id for the "FeaturesRestrictedID", the default is no voice communication. In our software, there are two purposes for the use of voice: voice communication between users and voice commands to the system. The voice commands' feature uses the free Web Speech API to transform speech to text. The voice communication feature collects the user audio stream from the browser media devices and, with the help of PeerJs API, creates calls that send the audio to other users. While on the desktop, these two technologies work. However, on mobile devices, they are not compatible due to the operating system's limitations over the microphone usage. The solution found for the user tests that would be performed in this project is to use a different way of communication, an external application to communicate. Besides the incompatibility with the voice communication feature, the Web Speech API also created some difficulties with the reproduction of the video. The video could not be reproduced with the microphone on simultaneously. To tackle this issue, we decided to allow the speech recognition to start only when the video was paused and stop the recognition when the video started playing. This created the need for the users to pause the video every time they wanted to interact with the system while on VR display.

4.5 System instrumentation for user tests

An important aspect taken in consideration while developing the application was the user tests. A log system was developed to help with the test phase of the research, which whenever a user performed a voice command the system saved it in a Map object. In this object, the keys correspond to the commands, and the values are the number of times the respective commands were given. When the video end event is triggered, this Map object and the user test id are printed into a **log file** in .txt format, and an automatic download of it is done on the user's device. This **user test id** is unique to each participant during the test phase and it is given by the test manager in an initial phase of the tests.

While considering the environment for tests, it was a possibility for them to be conducted with all the users in the same room. Recreating a user test scenario with all the users in the same room leads us to conclude that when everyone has the video paused, the microphones are on, and if someone gave a command to the system, it could be captured by the other devices. To avoid this, we added an extra step to **turn the microphone on and off**. The microphone does not turn automatically on when the user pauses the video. Once the video is paused, the user can tap the device twice again to turn on or off the microphone, depending on its state. To simplify the tests, the "FeaturesRestrictedId" parameter present in the URL that is described in section 4.4 can receive four valid ids:

- 1 is for tests done remotely and with voice communication on.
- 2 is for tests done remotely and with voice communication off.
- 3 is for tests done in the same room, which allows to turn the microphone on and off.
- 4 was specially developed for Phase C Configuration CV of the tests, explained in section 5.2, which disables the Radar and Point of Interest Share features.

Besides the changes on the "FeaturesRestrictedId", one more parameter was added to the URL representing the user test id. This last parameter is used to link the log files with the questionnaire answers of the testers. **The final URL** has the following structure: https://WebsiteName.com/roomCode/videoId/FeaturesRestrictionId/userTestId

4.6 Summary

This chapter presents a description of the implemented system and explains its usage. Following that, the best choice for the software architecture is discussed and we conclude that in our case it is a peer-to-peer approach. Along with this discussion, we present the technologies used for the software. The synchronized view, point of interest share and radar features are explained in more detail, since these are the collaborative features implemented into the existing AV360 visualizer. The development of each feature is explained along with the main challenges and solutions found to create a user-friendly interaction.

The chapter also presents the changes made in the system specifically for the tests that are described in the following section and will help us answer the research questions introduced in chapter 1.

Chapter 5

Testing and evaluation

Trying to find the most accurate answers for the research questions described in section 1.2, we have conducted user tests. Recapitulating, our primary goal is to determine if watching 360-degree videos simultaneously with other people can improve the quality of experience and what is the importance of some collaborative features, particularly radar orientation, pings, and voice communication.

We start the current chapter with a general description of the tests in section 5.1. Afterwards, the test structure is presented in section 5.2, explaining the different phases of the test sessions. Section 5.3 contains an explanation of the data that was collected during the experiments and section 5.4 exposes the information gathered and analyzes it. Finally, in section 5.5 we discuss the analyses executed previously.

5.1 General description

The developed system allows us to compare the quality of the user experience when visualising 360-degree videos individually and in a group. Besides video synchronization, the system also has some collaborative tools that can be used to communicate and have a better spatial awareness of the other users. We analyse and discuss the importance of these features for the user experience.

During this phase of the dissertation, we collected data from 44 participants. Each participant was part of a test session where they viewed two different videos, each video twice. To watch these videos, the participants used mobile devices and a Google Cardboard to enable the VR mode. When watching the videos, they had different tasks and conditions. They experienced individual visualization, synchronous visualization with others, and synchronous visualization without the availability of certain features like the voice communication or the radar and point of interest share. After visualizing a video, a questionnaire was given to the participants. The questionnaire's answers and the logs that recorded the voice commands given by the users were used to evaluate the user quality of experience.

The following section introduces the structure of the test sessions and explains the intention of each phase that composes them.

5.2 Test structure

The structure of the tests was created to achieve a good collection of data that allowed for the comparison of individual versus collaborative visualization of immersive videos, as well as the comparison of the radar and ping features with the voice communication feature.

The test sessions required at least two participants and they could not be repeated during tests: person "A" during test session number one cannot participate in any other test related to this study, because the answers to the second session could be influenced by the first one. The system is prepared to run on phones using Android as their operating system (OS), so the devices gathered and used for the experiments were a Samsung A20e, a Samsung Galaxy A5, and a Redmi Note 7. We could not gather any more phones, so, when the group from a test session was composed by more than 3 persons, the phones from the participants that had Android as their OS were used. For the VR interaction, Google Cardboards were available for use. When the participant's devices were used, we reviewed them to make sure they were consistent and had the system specifications required for the tests to be performed without issues. We used Google Chrome as the chosen browser and made sure it was in the highest version possible in all devices.

This section explains the structure of the test sessions and describes the videos used to execute them. Each test session consists of 3 phases described in more detail ahead: **Phase A** - Experience explanation and system tutorial (section 5.2.1), **Phase B** - Individual Vs Collaborative experience (section 5.2.2), and **Phase C** - Collaborative visualization and features comparison (section 5.2.3). The videos displayed and the tasks executed in each of the phases are described in subsection 5.2.4. Phase B and phase C are composed of two configurations each. The order in which phases B and C and their configurations are executed changes between sessions, to avoid order bias, as described in section 5.2.5.

5.2.1 Phase A - Experience explanation and system tutorial

This phase is the starting point of the tests for every user. First, the users receive an explanation of what is intended to be evaluated and the research context.

The devices used in the experiment are presented to the participants. Warnings regarding motion sickness and discomfort followed by a declaration of consent are given for the users to sign. The users are also informed that they can leave the experiment anytime.

The information gathering system is explained to the participants, and they are informed about the demographic data that will be disclosed to the public.

Finally, the participants have access to the devices used to visualise the immersive videos (mobile device and Google cardboard). The software and its usage is explained, and the users have some time to familiarise themselves with the system. In this phase, the participants receive help, and any questions they have are answered. The goal is to make the system intuitiveness less impactful on the user's experience.

5.2.2 Phase B - Solo visualisation VS Collaborative Visualisation

The second phase focuses on RQ1 and compares the quality of the user experience of a solo visualisation with a collaborative visualisation. It is composed by two different configurations. The following configurations were executed in varying order:

- *Configuration I (Individual)* - The users visualise a 360-degree video independently. Its focus is on creating a comparison baseline for the collaborative visualization.
- *Configuration C (Collaborative)* - The users view a video with access to all the system's collaborative tools and connected with, at least, one more viewer. The focus is to compare the user experience with the solo visualisation. It also evaluates whether any collaborative feature is underused or preferred when all tools are available.

5.2.3 Phase C - Collaborative visualisation and features comparison

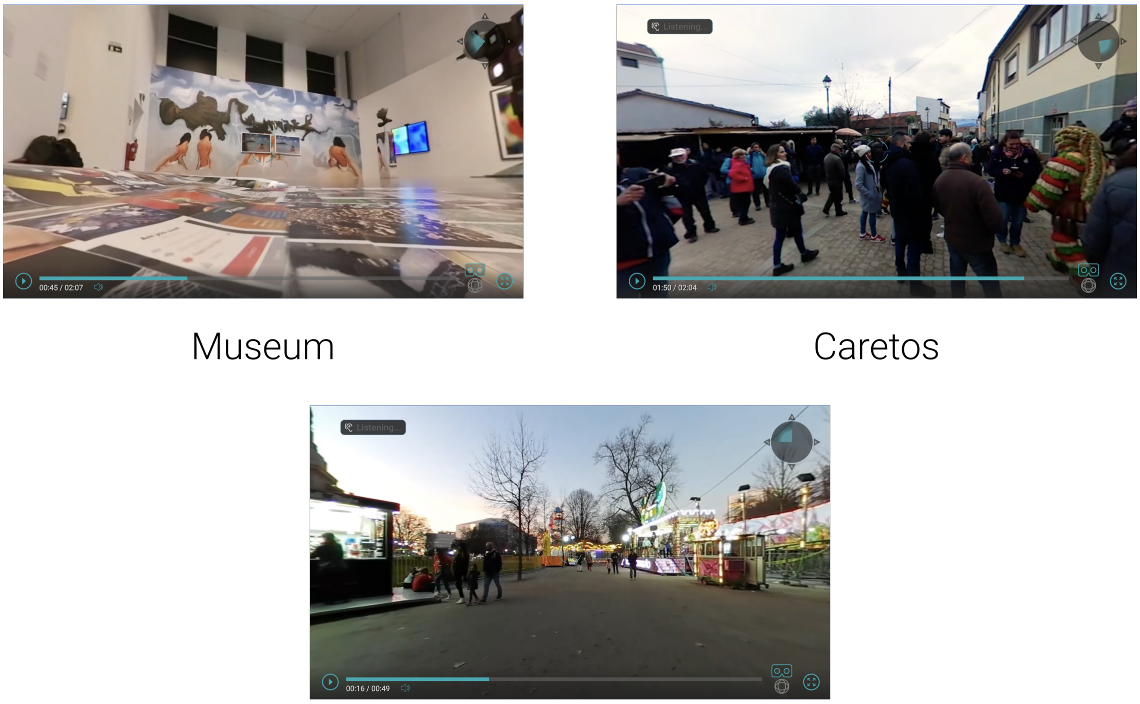
Focusing on RQ2 and RQ3, the third phase is also composed by two configurations. In this phase, the visualisations are done collaboratively, and in each configuration, the collaborative functionalities that they have access to are different. The following configurations were executed in varying order:

- *Configuration CV (Collaborative Voice)* - The users view a video with only the voice communication feature and not the spatial orientation collaborative features. This configuration aims to understand the efficiency of the voice communication feature in the collaboration and test if orientation is an issue when explicitly designed orientation features are unavailable.
- *Configuration CS (Collaborative Spatial)* - The users view a video with only the spatial orientation collaborative features and not the voice communication feature. This configuration is the opposite of the previous one. It aims to understand the efficiency of the orientation features in improving the quality of collaborative visualisation of 360-degree videos and test if the lack of verbal communication creates a different result for the quality of experience.

5.2.4 Participant tasks and selected 360-degree videos

The content provided to the users can significantly impact the user experience. An example of it is that an art enthusiast could have a much better experience and interest while watching a video that occurs in a museum. Three different videos were selected for the tests to minimize the impact of their content in the final result. They are illustrated in Figure 5.1. This selection was made through a set of requirements:

- The video should have several points of interest in the same instant.
- The video time should not be more than two minutes and a half.



Museum

Caretos

Christmas activities

Figure 5.1: Videos illustration

- The video time should not be less than thirty seconds.

The first topic was the most difficult to guarantee with the videos available in the dataset. To overcome the difficulty of encountering natural points of interest in the video, we decide to add an extra task to the participants in each video:

- *Video 1* - Caretos traditional festivities. The task assigned to the users was to count how many Caretos they could find in the video. Caretos are masked young men dressed in colourful suits, that are part of an ancient tradition.

- *Video 2* - Public space with Christmas activities. The task assigned to the users was to count how many children they could find in the video.

- *Video 3* - Museum visit. The task assigned to the users was to count how many fire extinguishers they could find in the video.

Additionally, the participants knowing each other was not a requirement, meaning that they could be more introverted and not share any input, creating some discomfort. Considering these conditions, the tasks were added to give the participants a reason to speak to each other and also to encourage them to look more to their surroundings.

5.2.5 Order of phases and 360-degree videos

Due to the fact that participating in phase B first can alter the results obtained in phase C (due to learning behaviour) and vice versa, we decided to make these two phases permutable. Meaning that in one test session phase B would be executed first and in another it would be phase C. The same happened with the configurations inside each phase. For example, in phase B, if configuration I was executed first it could influence the results of configuration C. The same solution was implemented and the configurations inside each phase were made permutable. The only fixed phase is Phase A, as it is always the start of the test session.

Besides the phases and configurations, the videos also changed between test sessions. This permutation was done to rotate the visualization between the three videos, and reduce the effect of their content on the experience. Table 5.1 represents the tests cycle order with Phase B, C and the videos. After every eight test sessions the cycle restarts for the phases and configurations.

Test 1	Phase B - I	Phase B - C	Phase C - CV	Phase C - CS
	Video 1		Video 2	
Test 2	Phase B - I	Phase B - C	Phase C - CS	Phase C - CV
	Video 3		Video 1	
Test 3	Phase B - C	Phase B - I	Phase C - CV	Phase C - CS
	Video 2		Video 3	
Test 4	Phase B - C	Phase B - I	Phase C - CS	Phase C - CV
	Video 1		Video 2	
Test 5	Phase C - CV	Phase C - CS	Phase B - I	Phase B - C
	Video 3		Video 1	
Test 6	Phase C - CV	Phase C - CS	Phase B - C	Phase B - I
	Video 2		Video 3	
Test 7	Phase C - CS	Phase C - CV	Phase B - I	Phase B - C
	Video 1		Video 2	
Test 8	Phase C - CS	Phase C - CV	Phase B - C	Phase B - I
	Video 3		Video 1	

Table 5.1: Tests cycles.

5.3 Collected data

During each test session, two methods were used to extract data for posterior analysis. The first method was through questionnaires. We gathered demographic data and tried to understand each participant's experience with the help of a set of well-defined questions. With the second method, we gathered quantitative data with system logs that were triggered with user interaction. Both methods are relevant, allowing us to compare the data obtained via the questionnaires with the actions of the users during the experiment.

5.3.1 Questionnaires

The questionnaires are essential in gathering various types of data for this experiment. A questionnaire was given to the user after phase A and each of the configurations of Phase B and C. The first questionnaire, right **after Phase A**, assembles a group of **demographic questions**: age, gender and VR experience. In addition, the user test ID is recorded by the test manager in order to identify and match the questionnaire responses with the respective logs.

We used the validated **User Experience Questionnaire (UEQ)** [LHS08]. With Likert-scale based questionnaires, a set of sentences are provided, to which the users can provide a quantitative value on a scale that represents a subjective dimension. This is how we evaluate the participant's experience **after each of the configurations from Phase B and C**. The answers go from 1 to 7. Some questions consider seven a good outcome, while others are structured the opposite way. To understand this difference between questions, table 5.2 shows the question number and the correspondent of 1 and 7.

Question	Left (1)	Right (7)	Scale
1	annoying	enjoyable	Attractiveness
2	understandable	understandable	Perspicuity
3	creative	dull	Novelty
4	easy to learn	difficult to learn	Perspicuity
5	valuable	inferior	Stimulation
6	boring	exciting	Stimulation
7	interesting	interesting	Stimulation
8	unpredictable	predictable	Dependability
9	fast	slow	Efficiency
10	inventive	conventional	Novelty
11	obstructive	supportive	Dependability
12	good	bad	Attractiveness
13	complicated	easy	Perspicuity
14	unlikable	pleasing	Attractiveness
15	usual	leading edge	Novelty
16	unpleasant	pleasant	Attractiveness
17	secure	secure	Dependability
18	motivating	demotivating	Stimulation
19	meets expectations	does not meet expectations	Dependability
20	inefficient	efficient	Efficiency
21	clear	confusing	Perspicuity
22	impractical	practical	Efficiency
23	organized	cluttered	Efficiency
24	attractive	unattractive	Attractiveness
25	friendly	unfriendly	Attractiveness
26	conservative	innovative	Novelty

Table 5.2: UEQ questions and scales.

Researching on how to evaluate user experience lead us to create **Extra Questions** besides the 26 present in the UEQ. The main article used as foundation to write these questions was "A Quality of Experience Model for Haptic Virtual Environments" written by A. Hamam, A. Saddik and J.Alja'am [HSA14].

Two Extra Questions were added after the UEQ in the four configurations. One related with VR: "On a scale of 1 to 7, how immersive was the experience?", where 1 is referring to "not immersive" and 7 to "very immersive". The last question is an open question that lets users add anything else that they felt but was not mentioned in the questionnaire.

Considering that Phase B - Configuration C, Phase C - Configuration CV and Phase C - Configuration CS have the collaborative component in the experience, three more Extra Questions were added to have a more specific feedback from the user related to this topic:

- On a scale of 1 to 7, how aware were you in relation to the other users?
- On a scale of 1 to 7, how spatially orientated were you in relation to the other users?
- On a scale of 1 to 7, how much did you feel the need to communicate through voice with other users?

5.3.2 Logs

The Logs have the function to gather the number of times the users use each voice command. Of all the commands, we intend to focus the analysis in the number of times the "play" and "target" commands are used and relate it with the stage of the experience. The tracked commands are downloaded at the end of each video in a text (.txt) file, as illustrated in Figure 5.2.

Figure 5.2: Logs example

5.3.3 Data processing

The data collected for each of the stages in the tests can be divided into three types for analysis. The UEQ Data, corresponding to the data collected through the questions from the UEQ questionnaire. The Extra Data, corresponding to the Extra Questions. And finally, the Logs Data.

The UEQ questionnaire has already developed a way to analyse the **UEQ Data** and we follow its data analysis tools that are available online. These tools transform the answers between 1 and 7 into values between -3 and 3, where -3 is always a bad outcome and 3 a good one. To achieve these values the tools use a simple process: if the good outcome correspond to 7 then $Final\_value = Answer - 4$; else $Final\_value = 4 - Answer$. The UEQ does not have an overall score. However, the questions affect the value of a corresponding scale. The existing scales and the correspondence to each question are described in table 5.2. The existing scales are **attractiveness**, **perspicuity**, **novelty**, **stimulation**, **dependability** and **efficiency**. For example, question 1 affects the attractiveness scale, and question 2 affects the perspicuity scale. The value for each scale is obtained for each participant. With these values we can calculate the average of each scale. The UEQ analysis tool makes available some benchmarks to compare and evaluate the final results. Besides the average for each scale, the answer distribution and the mean value per question is also computed by the analysis tool.

The **Extra Data** is analysed similarly to the UEQ data, and each question is considered a new scale. For Phase B - Configuration I, the only new scale is immersion. For the other configurations, the new scales are **immersion**, **awareness**, **orientation** and **voice need**.

For **Logs Data** it is calculated an average of the usage of every command in each stage of the tests.

5.4 Tests

This section presents the gathered data. The first subsection refers to the demographic data regarding the participants. The following sections describe each of the test phases and configurations. It displays the data assembled in every situation, analysing and discussing it.

5.4.1 Phase A

The test sessions started by explaining the research's area and goal to the users. They also got warned about all the side effects that VR usage could have, like motion sickness, and consented before proceeding with the experiment. After a short time for the participants to get used to the system, they filled out a form that gathered the following demographic data:

- 38,6% are male, and 61,4% are female. Represented in table 5.3.
- 6,8% of the participants were underage or recently turned 18 years old, 77,3% of the users were between 19 and 38 years old, and the remaining 15,9% were more than 39 years old. The oldest participant was 57 years old. Represented in table 5.4.
- No participants used VR technologies regularly or worked in the area. 59,1% of the participants never tried it, and 40,9% already had contact with it but not too much. Represented in table 5.5.

Gender	Participants
Female	27
Male	17
Other	0

Table 5.3: Gender of the participants.

Age	Participants
0-18	3
19-38	34
39-90	7

Table 5.4: Age of the participants.

VR Experience	Participants
Never tried	26
Just tried once or few times	18
Regular user for games and immersive experiences	0
Work or research VR	0

Table 5.5: The participant's previous experience with VR.

5.4.2 Phase B

As explained previously, phase B is composed by two configurations. The users watch the same video in both configurations but with different conditions. In Configuration I, the participants watch it individually, without any collaboration. In Configuration C, they watch the video with each other with access to all the system's functionalities. In both configurations, the task given to the users was the same depending on the video selected. This phase has the goal to gather information to answer RQ1. The analysis regarding the data collected in this phase are presented in the following subsections.

5.4.2.1 Configuration I

Configuration I of Phase B is expected to provide information about the user experience regarding individual visualisation so that this data can be compared with the one gathered from collaborative visualisation (Configuration C). The users answer the UEQ questionnaire mentioned in section 5.3.1 and one more question regarding immersion.

Figure 5.3 and Figure 5.4 present data per question of the UEQ. The first one illustrates the mean value per question and the second the answer distribution per question. Figure 5.5 displays the mean value per scale, including the immersion scale that is not present in the UEQ.

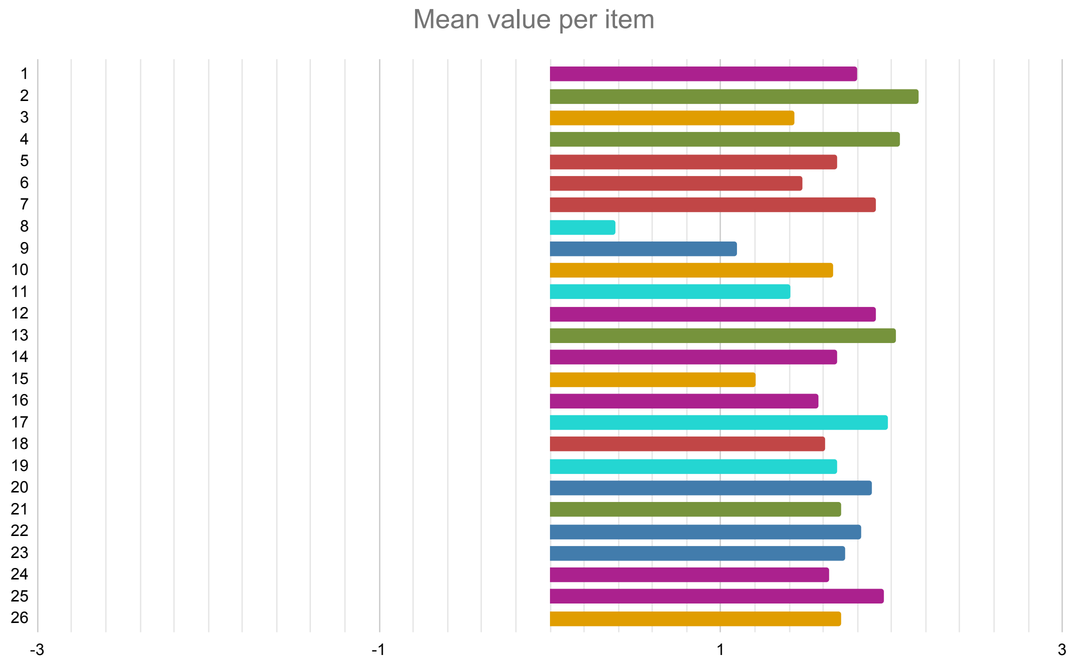


Figure 5.3: Phase B Configuration I - Mean value per question.

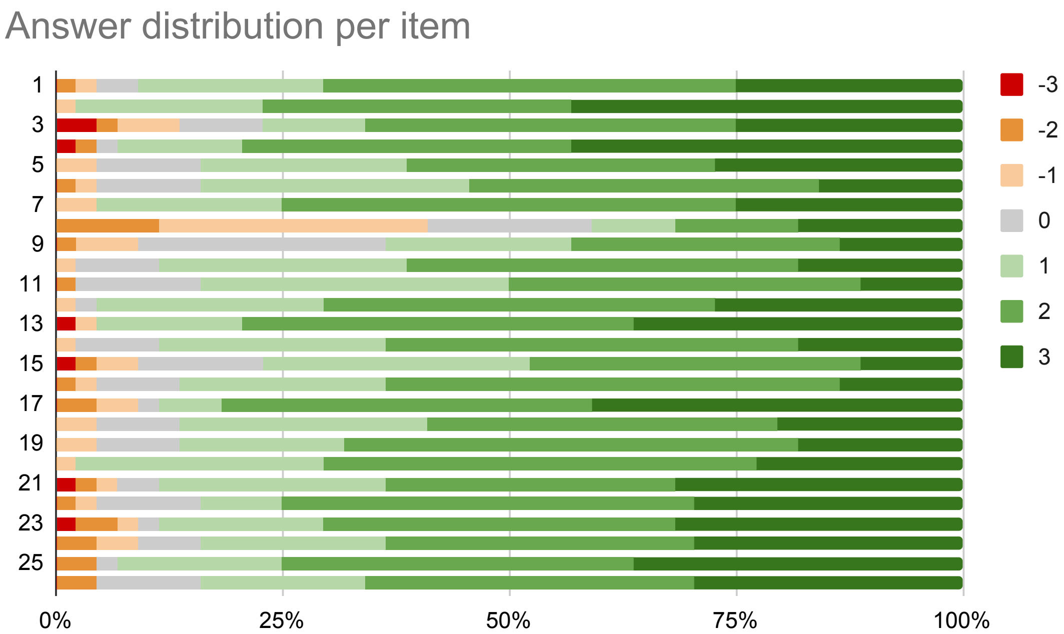


Figure 5.4: Phase B Configuration I - Answer distribution per question.

Figure 5.3 displays a significant difference in question number 8 (unpredictable - predictable) compared to the others. Question number 9 (fast - slow) also presents a discrepancy even though it is not as significant as number 8. Both questions have a balanced distribution regarding the answers, as shown in figure 5.4. After receiving some feedback from the user in the form of comments, we believe that this is happening due to some participants' misunderstanding on what to evaluate. Even though the users were warned that the questionnaire was regarding the system and experience, they could not abstract from the video itself. Some questions like number 8 (unpredictable - predictable) might have been answered concerning the video. The videos were seen twice each, making that if the users saw it for the first time, they would tend to answer towards the unpredictable side of the scale, and if it were the second time watching the video, they would tend to the predictable side.

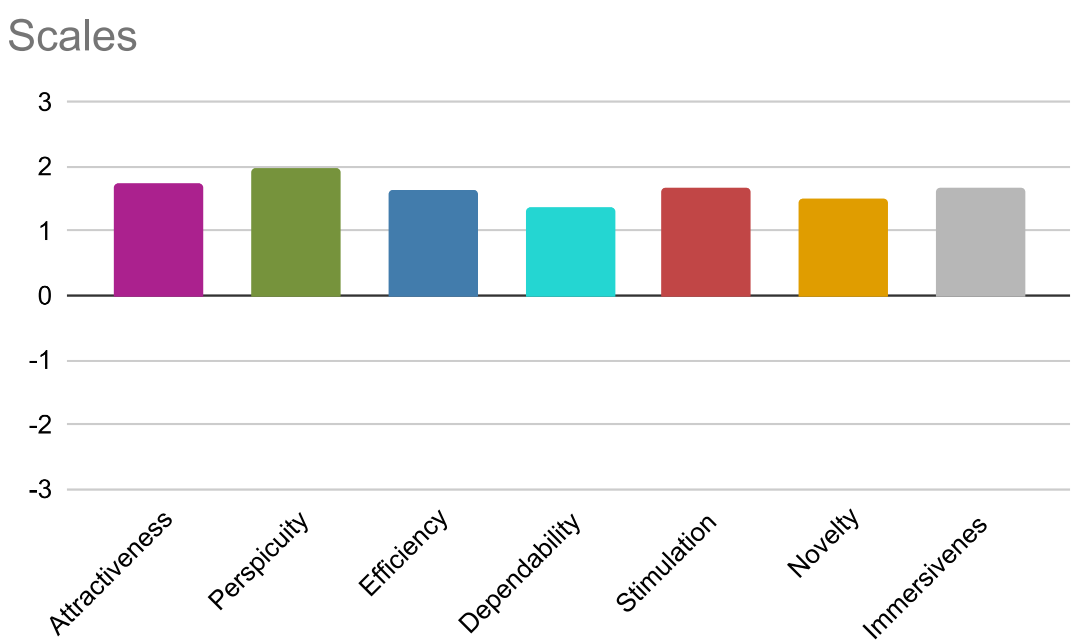


Figure 5.5: Phase B Configuration I - Scales average.

As a consequence of what was explained in the previous paragraph, the dependability has a lower average when compared to the other scales, as illustrated in figure 5.5. However, table 5.6 still shows a good overall result compared to the UEQ benchmarks. When receiving such positive responses from the participants, we split the users' data and calculated the average from people with experience in VR and with no experience. This analysis was conducted because we believed that the high values could be impacted by the enthusiasm of experiencing something for the first time. Surprisingly, the users with experience gave a better score than the other population, as shown in table 5.7 and 5.8.

	Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
Mean	1.758	1.983	1.631	1.364	1.670	1.500
Variance	0.911	0.786	0.740	0.917	0.892	0.654
Benchmark comparison	Good	Good	Good	Above average	Good	Good

Table 5.6: Phase B Configuration I - Benchmark comparison with all participants.

	Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
Mean	1.571	1.808	1.558	1.077	1.452	1.433
Variance	1.036	0.942	0.837	0.904	0.885	0.603
Benchmark comparison	Above average	Good	Good	Below average	Good	Good

Table 5.7: Phase B Configuration I - Benchmark comparison with participants not experienced in VR.

	Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
Mean	2.028	2.236	1.736	1.778	1.986	1.597
Variance	0.651	0.489	0.621	0.683	0.776	0.751
Benchmark comparison	Excellent	Excellent	Good	Excellent	Excellent	Good

Table 5.8: Phase B Configuration I - Benchmark comparison with participants experienced in VR.

The only point noticed regarding the logs is that some participants would use the target functionality just for personal amusement even though it did not have any practical advantages in individual visualisation. The average usage of each voice command is illustrated in table 5.9.

	Play	Target	Back	Skip
Average	2.727	0.091	0.136	0.045

Table 5.9: Phase B Configuration I - Average usage of each voice command.

5.4.2.2 Configuration C

Configuration C of phase B is expected to provide information about the user experience regarding collaborative visualisation so that this data can be compared with the one gathered from the individual visualisation. The users answer the UEQ questionnaire mentioned in section 5.3.1 and four more questions regarding immersion, awareness, orientation and the need to use the voice to communicate.

Figure 5.6 and Figure 5.7 present data per question of the UEQ. The first one illustrates the mean value per question and the second the answer distribution per question. Figure 5.8 displays the mean value per scale, including the immersion, awareness, orientation, and voice need scales that are not present in the UEQ.

Data from this configuration also supports the conclusions explained in the previous section regarding some participants misunderstanding the context of some questions, like numbers 8 and 9.

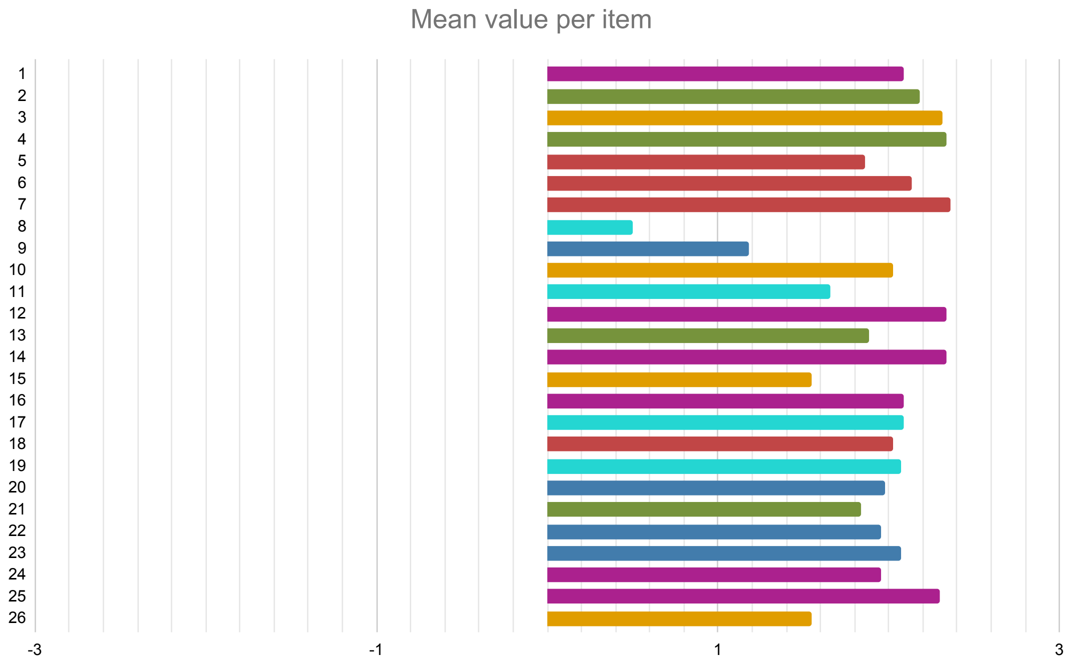


Figure 5.6: Phase B Configuration C - Mean value per question.

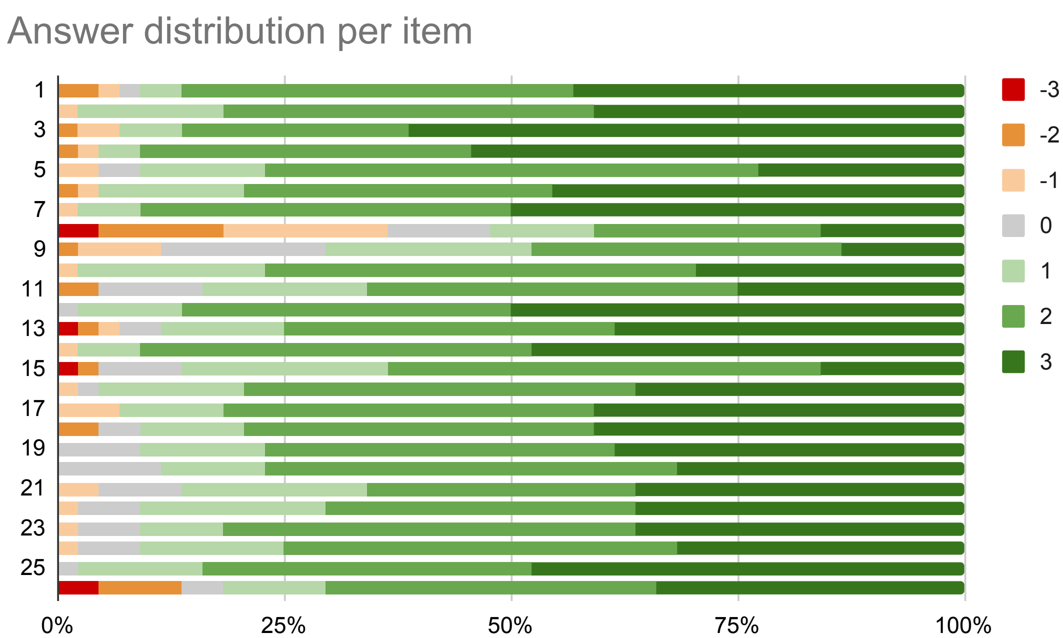


Figure 5.7: Phase B Configuration C - Answer distribution per question.

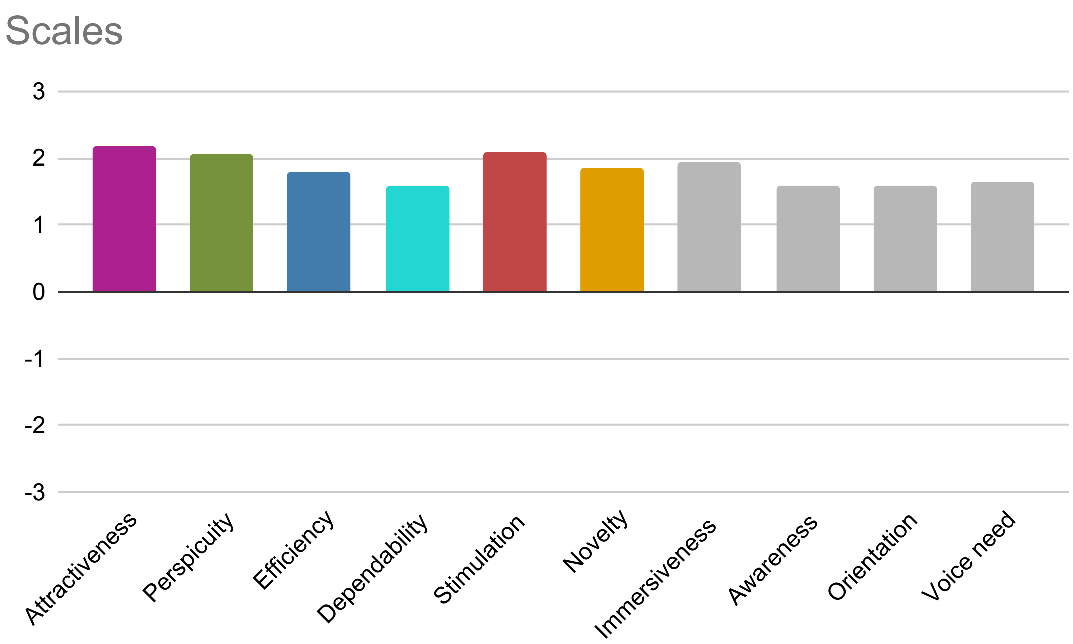


Figure 5.8: Phase B Configuration C - Scales average.

In Figure 5.8 it is observable that everything looks ordinary regarding the scales. There is no extreme variation. The participants felt aware and orientated in relation to each other. There was also a high recognition by the participants regarding the need of using voice as a means of communication. Comparing the UEQ questionnaire benchmarks in tables 5.10, 5.11 and 5.12, we can notice an impressive outcome in all the scales except the dependability scale in the users with no experience in VR.

	Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
Mean	2.186	2.063	1.795	1.580	2.097	1.858
Variance	0.471	0.617	0.716	0.531	0.666	0.812
Benchmark comparison	Excellent	Excellent	Good	Good	Excellent	Excellent

Table 5.10: Phase B Configuration C - Benchmark comparison with all participants.

	Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
Mean	2.147	2.048	1.798	1.423	2.019	1.663
Variance	0.341	0.590	0.720	0.499	0.510	1.005
Benchmark comparison	Excellent	Excellent	Good	Above average	Excellent	Excellent

Table 5.11: Phase B Configuration C - Benchmark comparison with participants not experienced in VR.

In the logs information, in table 5.13, we notice that all the voice commands were used. The subsequent section compares the logs from Configuration I and C, suggesting more findings.

	Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
Mean	2.241	2.083	1.792	1.806	2.208	2.139
Variance	0.684	0.691	0.752	0.519	0.914	0.435
Benchmark comparison	Excellent	Excellent	Good	Excellent	Excellent	Excellent

Table 5.12: Phase B Configuration C - Benchmark comparison with participants experienced in VR.

	Play	Target	Back	Skip
Average	3.773	1.523	0.114	0.068

Table 5.13: Phase B Configuration C - Average usage of each voice command.

5.4.2.3 Configurations comparison

This section will compare Configuration I and Configuration C of phase B. The graphic displayed in Figure 5.9 is achieved through tables 5.14 and 5.15.

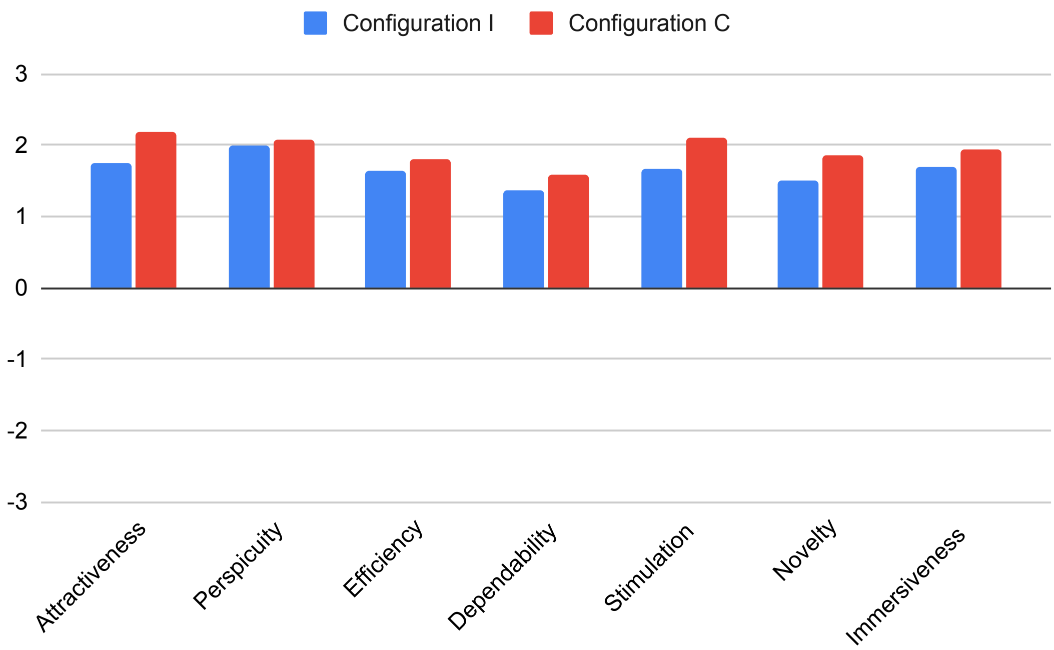


Figure 5.9: B - I Vs B - C. Scales averages.

We can notice that every single scale in the collaborative stage has a better score compared to the individual visualization. Even the perception of the system's attractiveness increases significantly with the collaboration between participants. One interesting result is the immersion levels that increase. After the experiment, some participants would comment that having other people participating in the same experience made it more real and, therefore, immersive. At the end of each group test, the participants were asked directly if they preferred to watch the videos

Scale	Mean	STD	N	Confidence (5%)	Confidence Interval	
Attractiveness	1.758	0.955	44	0.282	1.475	2.040
Perspicuity	1.983	0.887	44	0.262	1.721	2.245
Efficiency	1.631	0.860	44	0.254	1.377	1.885
Dependability	1.364	0.958	44	0.283	1.081	1.647
Stimulation	1.670	0.944	44	0.279	1.391	1.949
Novelty	1.500	0.809	44	0.239	1.261	1.739
Immersiveness	1.682	1.052	44	0.311	1.371	1.993

Table 5.14: Phase B Configuration I - Average, standard deviation and confidence.

Scale	Mean	STD	N	Confidence (5%)	Confidence Interval	
Attractiveness	2.186	0.686	44	0.203	1.983	2.388
Perspicuity	2.063	0.785	44	0.232	1.830	2.295
Efficiency	1.795	0.846	44	0.250	1.545	2.045
Dependability	1.580	0.729	44	0.215	1.364	1.795
Stimulation	2.097	0.816	44	0.241	1.855	2.338
Novelty	1.858	0.901	44	0.266	1.592	2.124
Immersiveness	1.932	1.169	44	0.346	1.586	2.277

Table 5.15: Phase B Configuration C - Average, standard deviation and confidence.

individually or with the group, and 100% of the answers were with the group. There was just one participant who added a comment saying that the group is better for entertainment purposes. However, it could be more challenging to process all the information in professional scenarios if everyone is not coordinated and "on the same page".

For the logs, we can notice in Figure 5.10 that the video was paused more times in Configuration C (and the target was more used) because the participants needed to explain their thought process and communicate with each other.

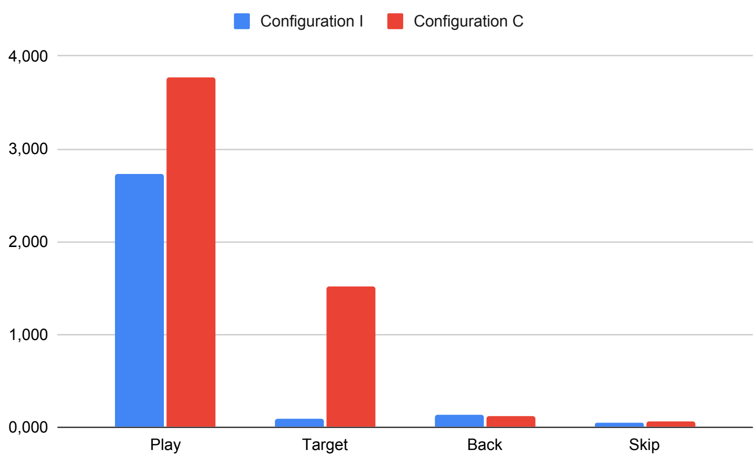


Figure 5.10: B - I Vs B - C. Logs comparison.

5.4.3 Phase C

Like phase B, phase C is also composed by two configurations. The users watch the same video in both configurations but with different conditions. Both visualisations are collaborative, but in Configuration CV, the participants watch it without the orientational features, and in Configuration CS, they watch the videos without the possibility of communicating through voice. In both configurations, the task given to the users was the same, depending on the video selected. This phase has the goal to gather information to answer RQ2 and RQ3. The analyses regarding the data collected in this phase are presented in the following subsections.

5.4.3.1 Configuration CV

Configuration CV of Phase C is expected to provide information about voice communication's efficacy and the need for spatial orientational features. The users answer the UEQ questionnaire mentioned in section 5.3.1 and four more questions regarding immersion, awareness, orientation and the need to use the voice to communicate.

Figure 5.11 and Figure 5.12 present data per question of the UEQ. The first one illustrates the mean value per question and the second the answer distribution per question. Figure 5.13 displays the mean value per scale, including the immersion, awareness, orientation, and voice need scales that are not present in the UEQ. Questions 8 and 9 follow the same pattern as in phase B.

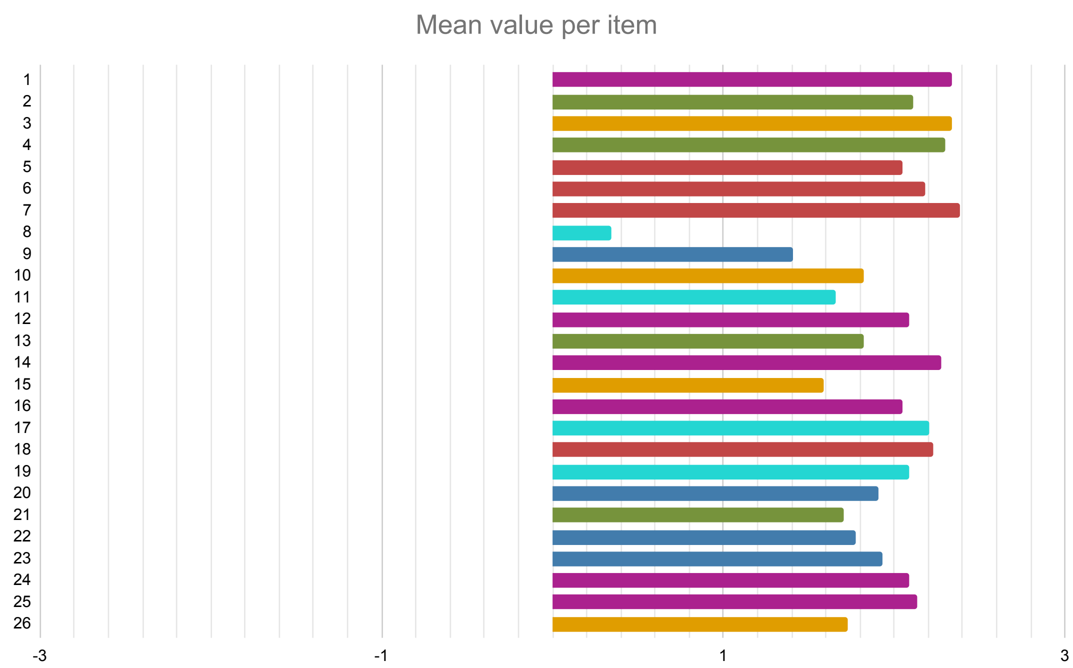


Figure 5.11: Phase C Configuration CV - Mean value per question.

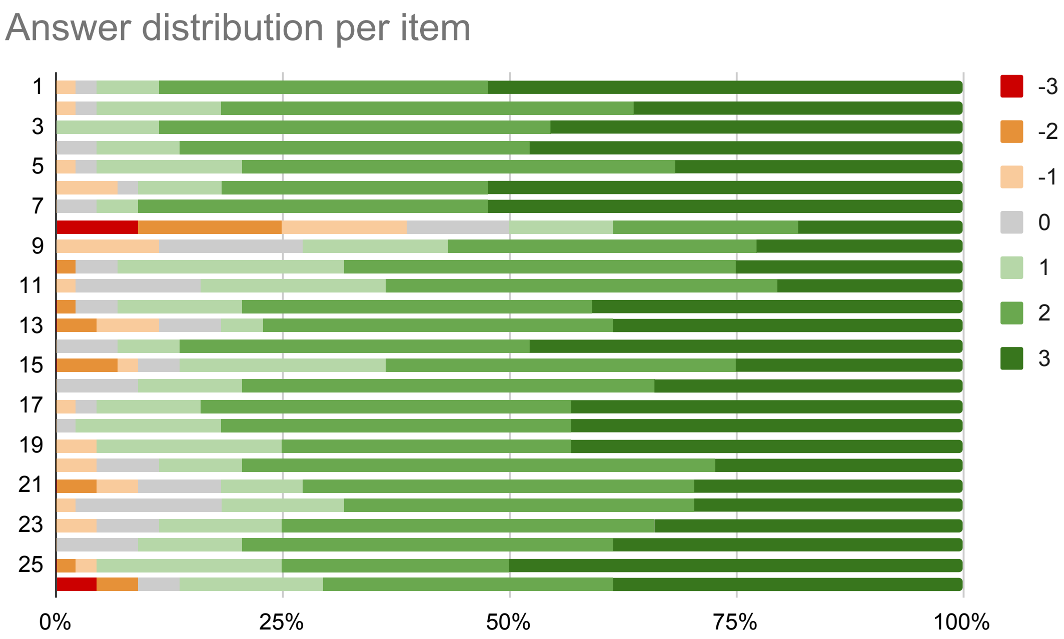


Figure 5.12: Phase C Configuration CV - Answer distribution per question.

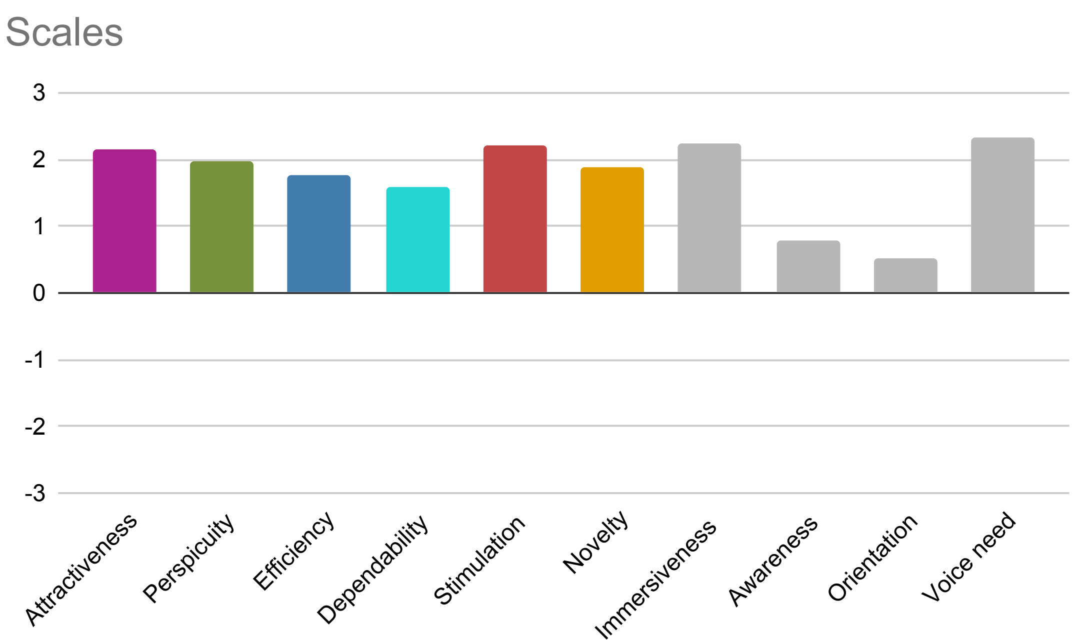


Figure 5.13: Phase C Configuration CV - Scales average.

Figure 5.13 and the comparison tables 5.16, 5.17 and 5.18 exhibit a good outcome regarding the user experience with the UEQ scales with good or excellent results. However, we can notice a significant discrepancy between awareness and orientation and the other scales. The users did not feel very aware of each other and were even less orientated. With the lack of spatial orientation features, the users' need to communicate through voice was very high. During the tests sessions, we noticed an extensive usage by the participants of the voice communication to orientate spatially, which corresponds correctly with the data achieved with the questionnaires.

	Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
Mean	2.163	1.983	1.756	1.574	2.210	1.869
Variance	0.636	0.798	0.751	0.717	0.575	0.583
Benchmark comparison	Excellent	Good	Good	Good	Excellent	Excellent

Table 5.16: Phase C Configuration CV - Benchmark comparison with all participants.

	Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
Mean	2.051	1.856	1.663	1.356	2.058	1.712
Variance	0.688	0.841	0.910	0.791	0.677	0.553
Benchmark comparison	Excellent	Good	Good	Above average	Excellent	Excellent

Table 5.17: Phase C Configuration CV - Benchmark comparison with participants not experienced in VR.

	Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
Mean	2.324	2.167	1.889	1.889	2.431	2.097
Variance	0.551	0.721	0.531	0.472	0.374	0.567
Benchmark comparison	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent

Table 5.18: Phase C Configuration CV - Benchmark comparison with participants experienced in VR.

As anticipated, table 5.19 illustrates the logs without using the voice command "target". While watching the test, it was noticed that the same comments had to be repeated and the video had to be often rewound by the participants.

	Play	Target	Back	Skip
Average	3.341	0.000	0.295	0.000

Table 5.19: Phase C Configuration CV - Average usage of each voice command.

5.4.3.2 Configuration CS

Configuration CS of phase C is expected to provide information about spatial orientational features efficacy and test the need for voice communication when it is missing. The users answer the UEQ questionnaire mentioned in section 5.3.1 and four more questions regarding immersion, awareness, orientation and the need to use the voice to communicate.

Figure 5.14 and Figure 5.15 present data per question of the UEQ. The first one illustrates the mean value per question and the second the answer distribution per question. Figure 5.16 displays the mean value per scale, including the immersion, awareness, orientation, and voice need scales that are not present in the UEQ. Questions 8 and 9 follow the same pattern as in phase B and the previous configuration described.

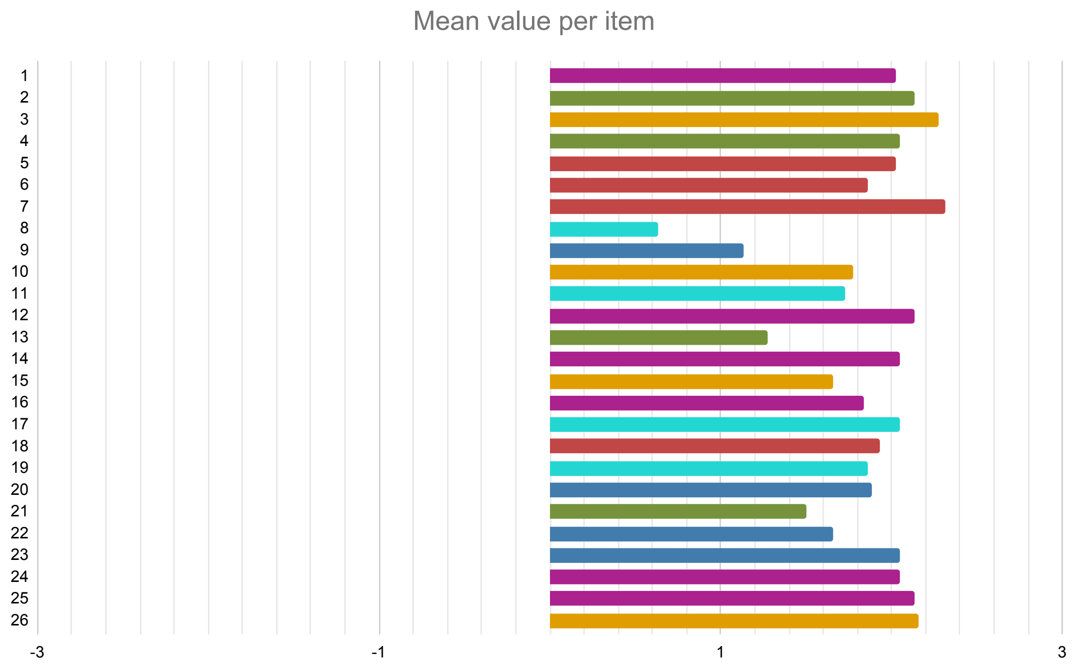


Figure 5.14: Phase C Configuration CS - Mean value per question.

Figure 5.16 and the comparison tables 5.20, 5.21 and 5.22 exhibit a good outcome regarding the user experience with the UEQ scales with above average, good or excellent results. We can notice low averages on the last three scales: awareness, orientation, and voice need. The participants felt some orientation but weren't much aware of the others. The results show that despite not being considered in this research as a spatial orientation feature, voice communication can improve the orientation of the users.

The tests sessions were executed in the same room, so the participants were asked not to speak to each other. The only sound that they could do was give voice commands to the system. In this situation, more than half of the participants could not resist to speak or leave some sort of sound clue to the others. This observation during the tests contradicts the answers given in the questionnaire. The answers cause us to believe that voice communication is important but it is not on a critical need.

As expected, there is a significant usage of the target feature in this stage, and correspondingly, there is a high need to stop the video and increase the number of both "play" and "target" voice commands, as demonstrated in table 5.23.

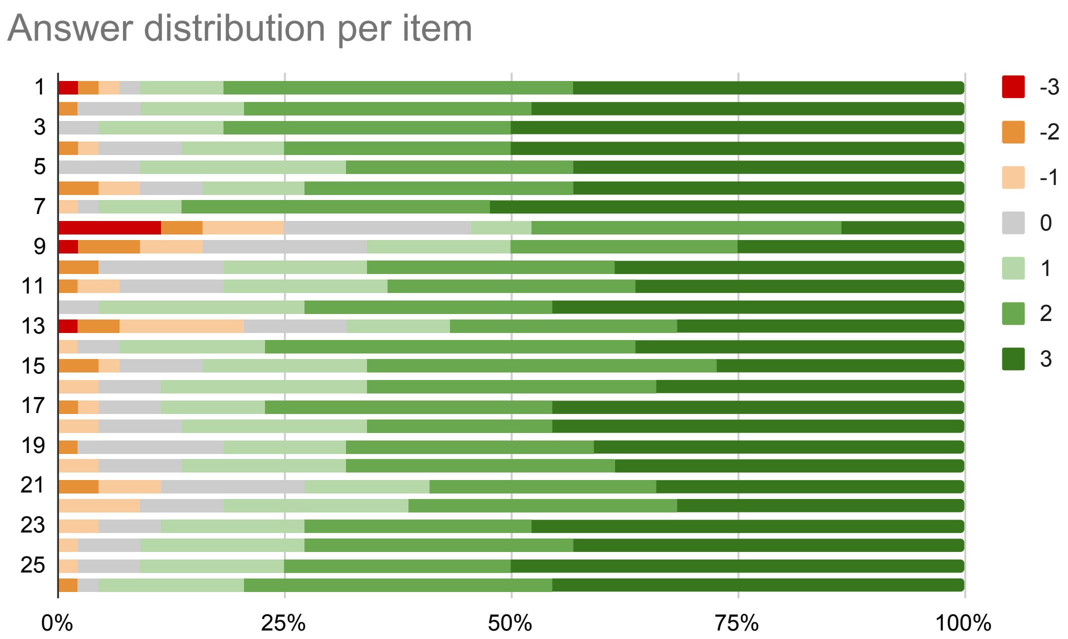


Figure 5.15: Phase C Configuration CS - Answer distribution per question.

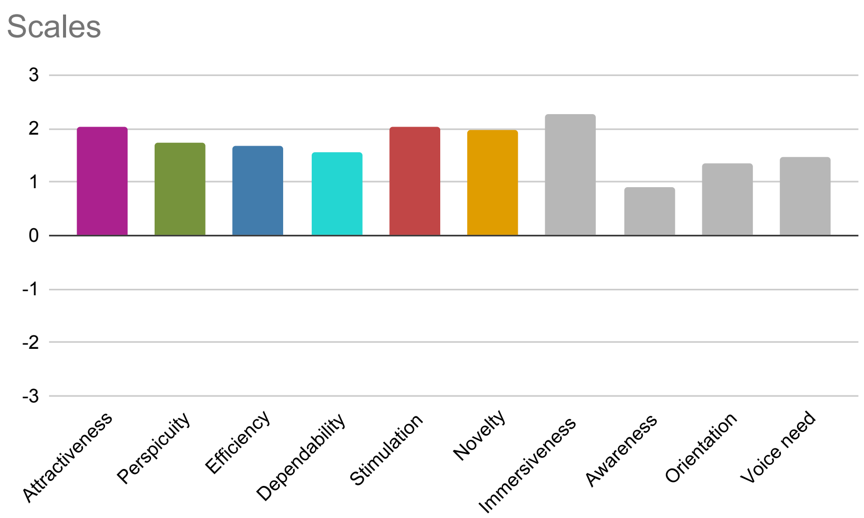


Figure 5.16: Phase C Configuration CS - Scales average.

	Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
Mean	2.038	1.739	1.682	1.568	2.034	1.966
Variance	0.816	1.203	1.222	0.783	0.941	0.810
Benchmark comparison	Excellent	Good	Good	Good	Excellent	Excellent

Table 5.20: Phase C Configuration CS - Benchmark comparison with all participants.

	Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
Mean	1.885	1.490	1.567	1.269	1.837	1.962
Variance	0.824	1.317	1.183	0.790	0.930	0.768
Benchmark comparison	Excellent	Above average	Good	Above average	Excellent	Excellent

Table 5.21: Phase C Configuration CS - Benchmark comparison with participants not experienced in VR.

	Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
Mean	2.259	2.097	1.847	2.000	2.319	1.972
Variance	0.765	0.876	1.302	0.485	0.866	0.918
Benchmark comparison	Excellent	Excellent	Good	Excellent	Excellent	Excellent

Table 5.22: Phase C Configuration CS - Benchmark comparison with participants experienced in VR.

	Play	Target	Back	Skip
Average	4.864	3.795	0.045	0.114

Table 5.23: Phase C Configuration CS - Average usage of each voice command.

5.4.3.3 Configurations comparison

This section compares Configuration CV and Configuration CS of phase C. The graphic displayed in Figure 5.17 is achieved through tables 5.24 and 5.25.

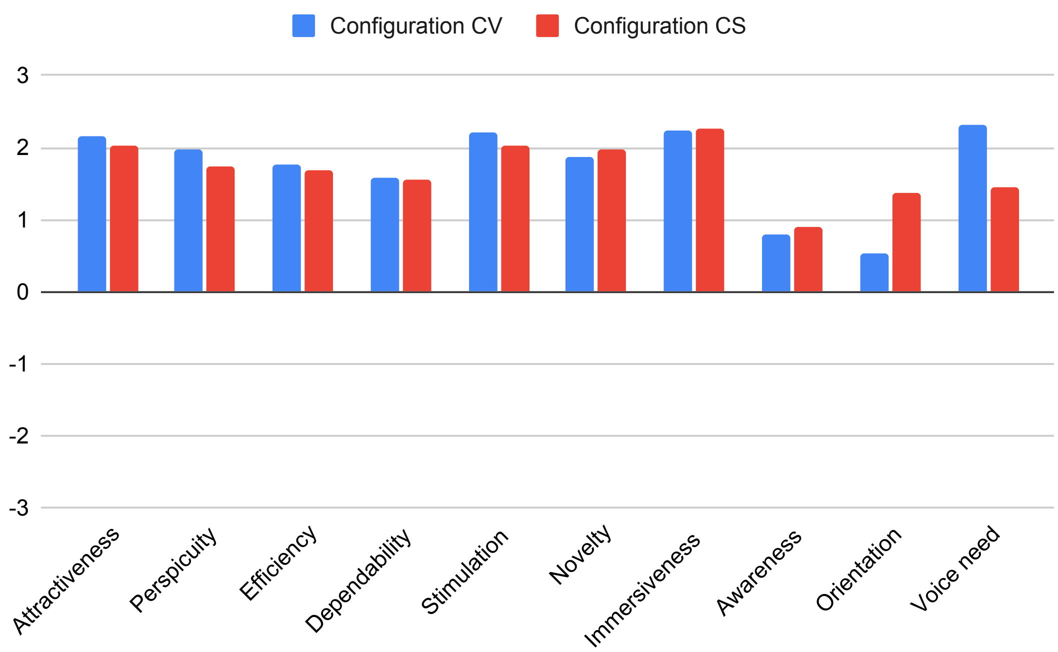


Figure 5.17: C - CV Vs C - CS. Scales averages.

Scale	Mean	STD	N	Confidence (5%)	Confidence Interval	
Attractiveness	2.1629	0.7977	44	0.2357	1.9272	2.3986
Perspicuity	1.9830	0.8931	44	0.2639	1.7191	2.2469
Efficiency	1.7557	0.8668	44	0.2561	1.4996	2.0118
Dependability	1.5739	0.8466	44	0.2502	1.3237	1.8240
Stimulation	2.2102	0.7586	44	0.2241	1.9861	2.4344
Novelty	1.8693	0.7758	44	0.2292	1.6401	2.0985
Immersiveness	2.2273	0.9115	44	0.2693	1.9579	2.4966
Awareness	0.7955	1.4560	44	0.4302	0.3652	1.2257
Orientation	0.5227	1.5475	44	0.4573	0.0655	0.9800
Voice need	2.3182	0.9092	44	0.2686	2.0495	2.5868

Table 5.24: Phase C Configuration CV - Average, standard deviation and confidence.

Scale	Mean	STD	N	Confidence	Confidence Interval	
Attractiveness	2.038	0.904	44	0.267	1.771	2.305
Perspicuity	1.739	1.097	44	0.324	1.415	2.063
Efficiency	1.682	1.105	44	0.327	1.355	2.008
Dependability	1.568	0.885	44	0.261	1.307	1.830
Stimulation	2.034	0.970	44	0.287	1.748	2.321
Novelty	1.966	0.900	44	0.266	1.700	2.232
Immersiveness	2.273	0.694	44	0.205	2.068	2.478
Awareness	0.909	1.552	44	0.459	0.450	1.368
Orientation	1.364	1.699	44	0.502	0.862	1.866
Voice need	1.455	1.823	44	0.539	0.916	1.993

Table 5.25: Phase C Configuration CS - Average, standard deviation and confidence.

When comparing the graphic in Figure 5.17, we can acknowledge that regarding the main difference between scales are in the perspicuity, voice need, stimulation, and orientation.

When using voice, the participants could ask each other questions if they did not understand something happening with the system. However, when voice communication was not allowed, their difficulties were enhanced, and therefore, the perspicuity got a lower score.

The voice communication need is an intriguing result where the participants felt that it was not critical when they did not have it available. However, they considered it a component that improved the user experience.

Two different groups pointed out that Configuration CV of phase C felt like a game because of the removal of orientational help and the task they had to achieve (find some elements in the video). This is, most likely, the explanation for the increase in the stimulation average.

The orientational scale difference was expected and is due to the lack of radar and target features in Configuration CV of phase C. Allowing us to conclude that these features significantly improve the orientation of the users.

The logs graph, illustrated in figure 5.18, exhibits that the "play" voice command usage was bigger in Configuration CS. We attribute these values to the fact that the users need to pause the video every time they wish to use the "target" functionality.

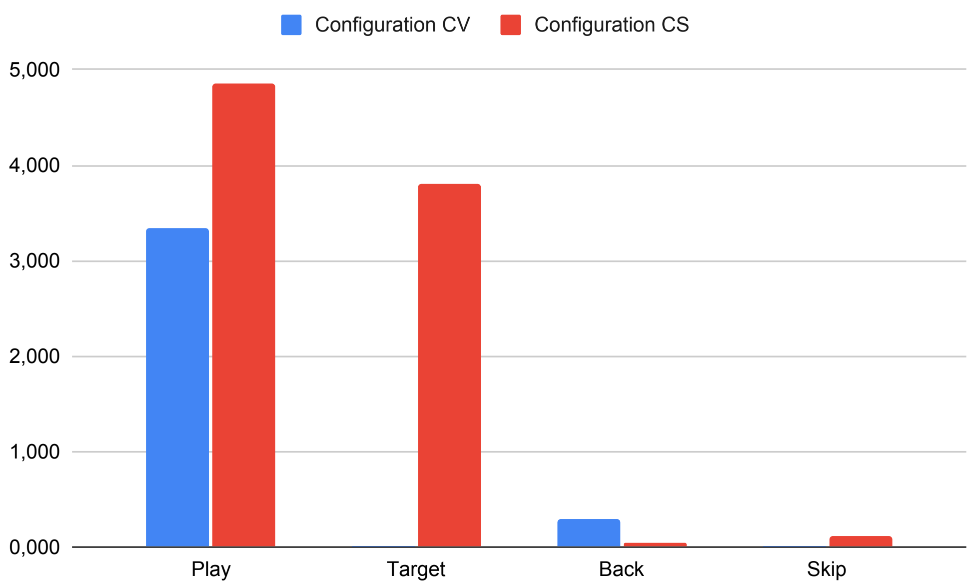


Figure 5.18: C - CV Vs C - CS. Logs comparison.

5.5 Discussion

In this section, we gather the main findings from the analysis of the data collected from the different phases. To facilitate the comparison between all the configurations, we illustrate Figure 5.19.

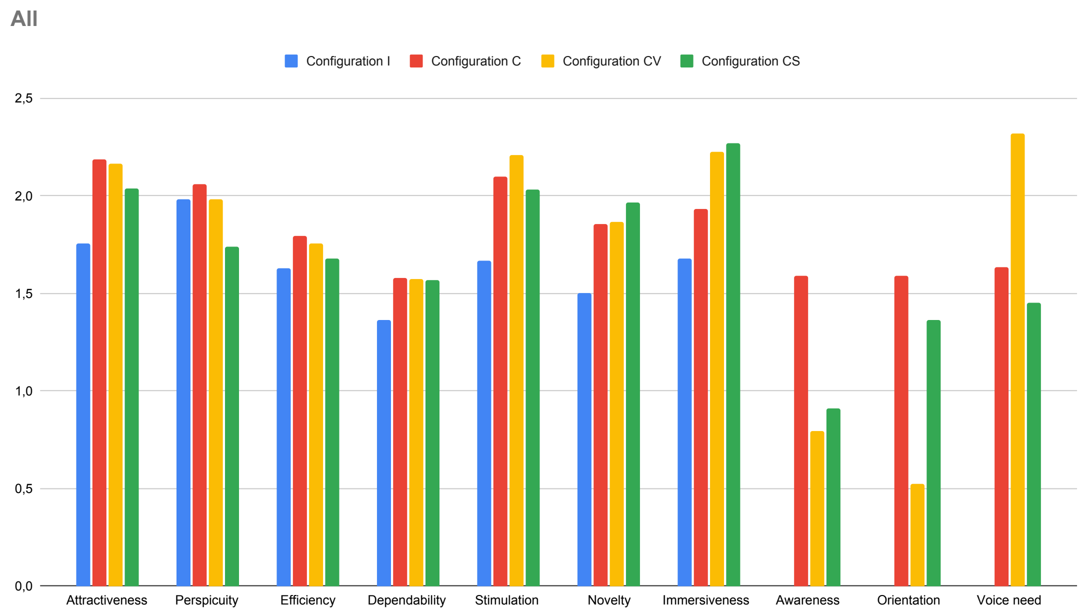


Figure 5.19: All configurations. Scales comparison.

The first main finding is regarding RQ1, where we try to understand if collaboration improves the quality of the user experience while visualizing 360-degree videos. When comparing Configuration I and Configuration C, both part of Phase B, we can notice that the scores from all the scales in the collaborative experience got higher values. This points to and increase in the overall user experience when immersive videos are visualized with others. This is consistent with what we investigated in chapter 2: Emotions can be intensified when shared with other humans and user experience is directly related with the user emotions. In configuration I we analysed that the overall individual experience was good, and we presuppose that the share of this experience with other people might have "intensified" the results.

The second result is regarding RQ2, where we try to understand if voice communication improves the quality of the user experience while visualizing 360-degree videos collaboratively. The original hypothesis was that the voice communication feature is very important to improve the user experience. The data analysed and the participants behavior when the voice communication feature was removed lead us to believe that it improves the user experience, however, it is not essential as long as they have collaborative features available as an alternative. This could mean that the implemented spatial orientational features provide some support for the collaborative experience, while not fully replacing the voice communication.

The third result is regarding RQ3, where we try to understand if the spatial orientational features (Radar and Point of Interest Share) help to improve the quality of the user experience while visualizing 360-degree videos collaboratively. With the data gathered we hypothesise that these features can improve the user's quality of experience depending on the system's goal. Suppose the system has the objective of stimulating the users, like a game. In that case, it should be pondered if those types of features would not decrease the difficulty level to a point where it would be too easy and not stimulating. However, if spatial awareness is vital for the experience, these features might be advantageous, without affecting negatively the user experience.

One extra finding that considers the averages of the immersion scale in each of the configurations, is that even though collaboration can increase the sense of immersion, too many features can create distraction for the user and decrease this sensation. We hypothesise this by comparing the immersiveness scale of configuration C - Phase B with the configurations from Phase C. The average on immersion was lower in the configuration where the users had access to all features. We consider this comparison acceptable because the videos distribution (video 1, 2 and 3) for Phase B and Phase C is almost equal.

Chapter 6

Conclusions

Immersive videos or 360-degree videos provide users with a unique experience of immersion where they have the freedom to explore all the angles around them. These videos increased their popularity in the last few years, and the market became more receptive to their commercialisation and usage. They gained value in several areas like storytelling, education, marketing, journalism, and clinical training.

Studies support that audience engagement is higher in 360-degree videos when the proper factors are considered. These elements vary from physical conditions like the display device and screen resolution to the video itself, how it is filmed, viewport dynamic, the bitrate, stalling effect. Another reason for the high audience engagement is that immersive videos have high impact on the emotions felt during their visualization. Besides the research on 360º videos, we also approached collaborative systems and how interacting with people improves the quality of experience. When trying to research regarding both, 360-degree videos and collaborative systems, merged, we noticed the scarcity of information.

This dissertation focused on understanding the impact of collaboration when visualising immersive videos. We wanted to understand if we could augment the quality of experience when watching 360-degree videos collaboratively and consequently increase the audience engagement with this technology. For that, we planned to develop a collaborative 360-degree video visualiser. A design was proposed, in chapter 3, with several collaborative features such as voice communication, the ability to share points of interest and a radar to inform about the field of view of all the participants.

The prototype developed is described in chapter 4. It was specially developed for the VR mode of mobile devices, and it of easy access because it is web base. This system allows users to watch immersive videos collaboratively in the browser. Besides synchronising the video to all the viewers present in the same "web room", the users have access to some features to help increase their orientation and awareness, like radar and the possibility to share points of interest in the viewport.

After developing the prototype, user tests were performed on it. The tests had the intention of studding the impact of the collaboration features in the user of experience. Each participant viewed two videos, two times each. Each time the video was seen in different conditions to gather data and evaluate it. The participants saw the video individually and with others. When they saw it collaboratively, one time they add access to all the features, the other they could only use the voice to communicate without access to orientational features and the other they did not had access to orientational features but could speak to each other.

At the end of the experiments, the sample of 44 participants led us to some conclusions regarding the research questions described at the beginning of this document. We concluded that the implementation of collaboration in the visualisation of 360-degree videos could improve the user experience. When implementing collaboration in a 360-degree video player, one feature that improves significantly the user experience is the voice communication feature and some other features to orientate the users should be considered to be implemented, since the users usually feel disorientated without them. We also observed that the amount of collaborative features implemented might impact the feeling of immersion, if to many features are available it could become confusing for the user and reduce the immersiveness.

Some improvements were considered throughout the different phases of this dissertation. The following section details some possibilities for future work in the prototype and this area.

6.1 Future Work

From the user tests, some possible improvements were identified, such as:

- Implement a keyword for the system to recognize the beginning of a voice command. Example: "AV360 play" and "AV360 target". This way, external noise and parallel interactions could happen without confusing the voice recognition system.
- Implement the possibility of giving voice commands while the video is playing.

Besides these topics, we also recognize two improvements. The voice communication feature should be implemented appropriately for all devices to avoid the need for external systems for the users to communicate. It is important to note that giving voice commands to the video while it is playing and allowing voice communication for all devices are problems caused by limitations of the available tools. New versions of these tools or alternative ones is the solution for these problems. And finally, the last improvement is that minor visual bugs in the radar rotation and the target arrow guides should be fixed.

In this dissertation, we established that the implementation of collaboration in the visualisation of immersive videos is an asset. However, besides the synchronisation, we focused only on three collaborative features (voice communication, radar and point of interest share). Other collaborative features, presented in chapter 3, like the "king of the room", "tunnel vision" or even desktop features like "Mini-view" and "Text chat" could be implemented and tested to see how useful can they be and what is their impact on the quality of the user experience.

References

[AA06] National Aeronautics and Space Administration. Head-mounted display and wired gloves, 2006.

[ABAF20] Asaad Alghamdi, Younes Balah, Mohammad Albejadi, and Muhamad Felemban. BeeCast. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pages 1–3, New York, NY, USA, sep 2020. ACM.

[AEB20] Lemonia Argyriou, Daphne Economou, and Vassiliki Bouki. Design methodology for 360° immersive video applications: the case study of a cultural heritage virtual tour. *Personal and Ubiquitous Computing*, 24(6):843–859, dec 2020.

[av3] Augmented video 360. https://av360.inesctec.pt/. Accessed: 2021-06-30.

[Awa21] Awareness. *Cambridge Advanced Learner's Dictionary & Thesaurus*. Cambridge University Press, 2021.

[AWU$^+$20] Muhammad Shahid Anwar, Jing Wang, Asad Ullah, Wahab Khan, Sadique Ahmad, and Zesong Fei. Measuring quality of experience for 360-degree videos in virtual reality. *Science China Information Sciences*, 63(10):15, 2020.

[Bar91] David B. Barker. The Behavioral Analysis of Interpersonal Intimacy in Group Development. *Small Group Research*, 22(1):76–91, feb 1991.

[BCBM18] Mark Billinghurst, Maxime Cordeil, Anastasia Bezerianos, and Todd Margolis. Collaborative immersive analytics. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11190 LNCS, pages 221–257. Springer Verlag, 2018.

[BGRP01] Steve Benford, Chris Greenhalgh, Tom Rodden, and James Pycock. Collaborative Virtual Environments. *Communications of the ACM*, 44(7):79–85, 2001.

[BM02] Georgia Bafoutsou and Gregoris Mentzas. Review and functional classification of collaborative systems. *International Journal of Information Management*, 22(4):281–305, 2002.

[CBS$^+$00] J. J. Cadiz, A. Balachandran, E. Sanocki, A. Gupta, J. Grudin, and G. Jancke. Distance learning through distributed collaborative video viewing. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 135–144. Association for Computing Machinery (ACM), 2000.

[CL20] Samuel Cosgrove and Joseph J. Laviola. Visual Guidance Methods in Immersive and Interactive VR Environments with Connected 360° Videos. In *Proceedings - 2020 IEEE Conference on Virtual Reality and 3D User Interfaces, VRW 2020*, pages 653–654. IEEE, mar 2020.

[CMP12] P Le Callet, S Möller, and A Perkis. Qualinet white paper on definitions of quality of experience (2012). *European Network on Quality of Experience in ...*, 2012.

[CS98] E F Churchill and D. Snowdon. Collaborative virtual environments: An introductory review of issues and systems. *Virtual Reality*, 3(1):3–15, 1998.

[EB94] Mica R. Endsley and Cheryl A. Bolstad. Individual Differences in Pilot Situation Awareness. *The International Journal of Aviation Psychology*, 4(3):241–264, jul 1994.

[FBL02] Francesca Frassinetti, Nadia Bolognini, and Elisabetta Làdavas. Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental Brain Research*, 147(3):332–343, dec 2002.

[GPB+00] Chris Greenhalgh, Jim Purbrick, Steve Benford, Mike Craven, Adam Drozd, and Ian Taylor. Temporal links. In *Proceedings of the eighth ACM international conference on Multimedia - MULTIMEDIA '00*, pages 67–74, New York, New York, USA, 2000. ACM Press.

[GST05] Tom Gross, Chris Stary, and Alex Totter. User-centered awareness in computer-supported cooperative work-systems: Structured embedding of findings from social sciences. *International Journal of Human-Computer Interaction*, 18(3):323–360, 2005.

[HSA14] Abdelwahab Hamam, Abdulmotaleb El Saddik, and Jihad Alja'am. A Quality of Experience Model for Haptic Virtual Environments. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 10(3):1–23, apr 2014.

[ISI+17] Andrew Irlitti, Ross T. Smith, Stewart Von Itzstein, Mark Billinghurst, and Bruce H. Thomas. Challenges for Asynchronous Collaboration in Augmented Reality. In *Adjunct Proceedings of the 2016 IEEE International Symposium on Mixed and Augmented Reality, ISMAR-Adjunct 2016*, pages 31–35. Institute of Electrical and Electronics Engineers Inc., jan 2017.

[JBB+20] Manish Kumar Jha, Marwa Boukadida, Hamdi Ben Abdessalem, Alexie Byrns, Marc Cuesta, Marie Andrée Bruneau, Sylvie Belleville, and Claude Frasson. Improving cognitive and emotional state using 3D virtual reality orientation game. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12149 LNCS, pages 220–225. Springer, jun 2020.

[KA92] Kari Kuutti and Tuula Arvonen. Identifying potential CSCW applications by means of activity theory concepts: A case example. In *Proceedings of the Conference on Computer-Supported Cooperative Work*, pages 233–240, 1992.

[Lan88] Janor Lanier. A vintage virtual reality interview. *Whole Earth Review.*, 1988.

[LCH+17] Yen Chen Lin, Yung Ju Chang, Hou Ning Hu, Hsien Tzu Cheng, Chi Wen Huang, and Min Sun. Tell me where to look: Investigating ways for assisting focus in 360° video. In *Conference on Human Factors in Computing Systems - Proceedings*, volume 2017-May, pages 2535–2545, 2017.

[LHS08] Bettina Laugwitz, Theo Held, and Martin Schrepp. Construction and Evaluation of a User Experience Questionnaire. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5298 LNCS, pages 63–76. Springer Verlag, 2008.

[LKS+16] Anh Le, Lorenzo Keller, Hulya Seferoglu, Blerim Cici, Christina Fragouli, and Athina Markopoulou. MicroCast: Cooperative Video Streaming Using Cellular and Local Connections. *IEEE/ACM Transactions on Networking*, 24(5):2983–2999, oct 2016.

[Luc07] Annie Luciani. Virtual reality and virtual environment. In *Enaction and enactive interfaces : a handbook of terms*, pages 299–300. Enactive Systems Book, 2007.

[MGFS13] Mikhail Morozov, Alexey Gerasimov, Mikhail Fominykh, and Andrey Smorkalov. Asynchronous immersive classes in a 3D virtual world: Extended description of vAcademia. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7848, pages 81–100. Springer Verlag, 2013.

[MTUK95] Paul Milgram, Haruo Takemura, Akira Utsumi, and Fumio Kishino. Augmented Reality: A class of displays on the reality-virtuality continuum. In *Telemanipulator and Telepresence Technologies*, volume 2351, pages 282–292, 1995.

[NBM+19] David Narciso, Maximino Bessa, Miguel Melo, António Coelho, and José Vasconcelos-Raposo. Immersive 360 º video user experience: impact of different variables in the sense of presence and cybersickness. *Universal Access in the Information Society*, 18(1):77–87, mar 2019.

[OP07] Hannes Olivier and Niels Pinkwart. Collaborative virtual environments-hype or hope for CSCW? *IfI Technical Report Series*, page 13, 2007.

[OT10] Heather L. O'Brien and Elaine G. Toms. The development and evaluation of a survey to measure user engagement, jan 2010.

[PFPP21] Mariapaola Puggioni, Emanuele Frontoni, Marina Paolanti, and Roberto Pierdicca. ScoolAR: An Educational Platform to Improve Students' Learning through Virtual Reality. *IEEE Access*, 9:21059–21070, 2021.

[PK20] Gitte Pedersen and Konstantinos Koumaditis. Virtual Reality (VR) in the Computer Supported Cooperative Work (CSCW) Domain: A Mapping and a Pre-study on Functionality and Immersion. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12191 LNCS, pages 136–153. Springer, 2020.

[Pos80] Michael I. Posner. Orienting of attention. *The Quarterly journal of experimental psychology*, 32(1):3–25, feb 1980.

[Sch96] Eve M Schooler. Conferencing and collaborative computing. *Multimedia Systems*, 4(5):210–225, 1996.

[SCY+18] Maximilian Speicher, Jingchen Cao, Ao Yu, Haihua Zhang, and Michael Nebeling. 360Anywhere. *Proceedings of the ACM on Human-Computer Interaction*, 2(EICS):1–20, jun 2018.

[SG96] Scott L Springer and Rajit Gadh. State-of-the-art virtual reality hardware for computer-aided design. *Journal of Intelligent Manufacturing*, 7(6):457–465, 1996.

[SHP19] Rébaï Soret, Christophe Hurter, and Vsevolod Peysakhovich. Attentional orienting in real and virtual 360-degree environments. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 1–3, New York, NY, USA, jun 2019. ACM.

[SSTG17] Raimund Schatz, Andreas Sackl, Christian Timmerer, and Bruno Gardlo. Towards subjective quality of experience assessment for omnidirectional video streaming. In *2017 9th International Conference on Quality of Multimedia Experience, QoMEX 2017*, 2017.

[Ste92] Jonathan Steuer. Defining Virtual Reality: Dimensions Determining Telepresence. *Journal of Communication*, 42(4):73–93, 1992.

[SWGW18] Ayoung Suh, Guan Wang, Wenying Gu, and Christian Wagner. Enhancing audience engagement through immersive 360-degree videos: An experimental study. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10915 LNAI, pages 425–443. Springer Verlag, jul 2018.

[TF17] Anthony Tang and Omid Fakourfar. Watching 360° videos together. In *Conference on Human Factors in Computing Systems - Proceedings*, volume 2017-May, pages 4501–4506, New York, NY, USA, may 2017. ACM.

[THB+20] Alexander Toet, Fabienne Heijn, Anne-Marie Brouwer, Tina Mioch, and Jan B. F. van Erp. An Immersive Self-Report Tool for the Affective Appraisal of 360° VR Videos. *Frontiers in Virtual Reality*, 1, sep 2020.

[ZD09] Ning-Ning Zhou and Yu-Long Deng. Virtual Reality: A State-of-the-Art Survey. *International Journal of Automation and Computing*, 6(4):319–325, 2009.

[ZWWZ18] Cheng Zhan, Zhe Wen, Xiumin Wang, and Liyue Zhu. Device-to-Device assisted wireless video delivery with network coding. *Ad Hoc Networks*, 69:76–85, feb 2018.

[ZYY+20] Wenjie Zou, Lihui Yang, Fuzheng Yang, Zhibin Ma, and Qiyong Zhao. The impact of screen resolution of HMD on perceptual quality of immersive videos. In *2020 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2020*, 2020.

[@alpharefs]

Appendix A

User Scenarios

A.1 Non-hierarchical Scenarios

Entertainment

Scenario:

Six friends hang out remotely and decide to watch a 360-degree video together.

The motivation of the users:

The six friends just want to entertain themselves and hang out with each other.

Example:

The users will watch a 360-degree video related to skating. All of them are in their respective homes.

One of the friends create a room and share the room's link with the others.

When they agree, one of them starts the video.

Specific interactions:

- The users talk and laugh with each other.
- In the video timeline, one of the friends returns to point out something funny that he saw.
- One of the friends uses a ping with a reaction to react to a specific space in the video.
- One of the friends mutes himself due to a lot of background noise.

Exploration of the Mariana trench

Scenario:

Explorers send a 360-degree camera down in the Mariana trench and review the footage after returning.

The motivation of the users:

The explorers want to discover and be attentive to details.

Example:

The explorers will watch a 360-degree video that visits a part of the Mariana trench.
One of the explorers prepares the video and shares the room's link with the others.
When they agree, one of them starts the video.

Specific interactions:

- The explorers verbally communicate with each other reacting to what is being reproduced in the video.
- The explorers return in the timeline to study some detail in the video.
- An explorer uses a virtual beacon to direct the attention of others to a specific point.
- To have a clear video sound, every explorer mutes their microphone if they do not intend to talk.
- The explorers slow down the video to understand something moving in the back of the footage.

A.2 Hierarchical Scenarios

Medicine Class

Scenario:

Students take a medical class where a specific procedure is learned through a 360-degree video.

The motivation of the users:

Students want to learn and question what is displayed in the video. The teacher is present to explain the environment and critical points in the video.

Example:

Students will watch a 360-degree video of a successful heart transplant. Some students are together in the classroom, and others are in their respective homes.

The professor prepares the video and shares the room's link with all the students.

When the professor checks that everyone is ready, he starts the video and explains every step.

In the end, there is time to question the professor, and students can come back in the video to review specific parts.

Specific interactions:

- The professor verbally communicates with all the students to explain what is being reproduced in the video.
- A student with a question raises the hand emoji to have the professor's attention.
- The professor returns in the video timeline to explain the operation's details.
- While asking a question, a student uses a visual beacon to point to a specific place in the video and direct the other users' attention.
- A student has difficulty keeping track of what the professor is referring to in the 360-degree scenario. Therefore he changes the view of their screen to the professor's view.
- The professor notices that the group's attention is not in the right place by checking the radar with the vision cones of the students and corrects them by telling the cone that they need to follow.
- The professor uses the compass directions to orient the students to a specific area.
- The professor needs to mute a misbehaving student.
- The professor removes a student that is misbehaving from the room.
- The professor forces the student attention with the help of a vignetting effect created in the field of view of the students.
- The professor slows the video down to explain a detail happening in the operation.
- The professor pauses the video and draws in the paused image to explain a student's question adequately.

Field trips

Scenario:

Students from a school visit restricted companies where physical presence is not allowed to non-staff.

The motivation of the users:

Some students may be interested in the subject, and others may be forced to attend the visit. The teacher and/or guide is present to explain the environment and key points in the video.

Example:

Students will watch a 360-degree video that visits a contaminated room where specific suits are needed. Some students are together in the classroom, and others are in their respective homes.

The professor prepares the video with the company guide and shares the room's link with all the students.

When the professor checks that everyone is ready, the guide starts the video and explains every step.

In the end, there is time to question the guide, and students can come back in the video to review specific parts.

Specific interactions:

- The guide verbally communicates with all the students to explain what is being reproduced in the video.
- The students that are satisfied with the tour react with the clap emoji to the guide's presentation.
- A student with a question raises the hand emoji to have the attention of the professor and guide.
- The guide returns in the video timeline to explain some detail in the presentation.
- While asking a question, a student uses a visual beacon to point to a specific place in the video and direct the other users' attention.
- A student is having difficulty keeping track of what the guide refers to in the 360 scenarios. Therefore he changes the view of their screen to the guide's view.
- The guide notices that the group's attention is not in the right place by checking the radar with the vision cones of the students and corrects them by telling the cone colour that they need to follow.
- The guide uses the compass directions to orient the students to a specific area.
- The professor needs to mute a misbehaving student.
- The guide forces the student attention with the help of a vignetting effect created in the field of view of the students.

Touristic tours

Scenario:

Countries are in lockdown due to COVID-19. Several tourists worldwide have already bought tickets to the Louvre Museum in Paris, France. It is not allowed to get inside the museum because of the pandemic situation, so the Louvre creates a video tour carried with the help of a guide. This way, Louvre can offer the option of the video with a guide, instead of only accepting the return of the tickets.

The motivation of the users:

Tourists are interested in visiting the Louvre museum. Some may understand better about art, and others might be beginners starting their journey in exploring art. The guide is present to explain the environment and key points in the video.

Example:

The tourists will watch a 360-degree video that visits the Louvre museum.
The museum prepares the video and shares the room's link with all the tourists.
When the guide checks that everyone is ready, the tour starts.
In the end, there is time to question the guide.

Specific interactions:

- The tourists do not have permission to change the video time.
- The guide verbally communicates with all the tourists to explain what is being reproduced in the video.
- The tourists that are satisfied with the tour react with the clap emoji to the guide's presentation.
- The guide returns in the video timeline to explain some detail in the presentation.
- While asking a question, the tourist uses a visual beacon to point to a specific place in the video and direct the other users' attention.
- A tourist is having difficulty keeping track of what the guide refers to in the 360 scenarios. Therefore he changes the view of their screen to the guide's view.
- The guide notices that the group's attention is not in the right place by checking the radar with the vision cones of the tourists and corrects them by telling the cone colour that they need to follow.
- The guide uses compass directions to orient the tourists to a specific area.
- The guide needs to mute a tourist that has a broken microphone.

Football game revision

Scenario:

The football team players watch their last match in a 360-degree video.

The motivation of the users:

The players want to review their last game to acknowledge mistakes and improve their performance to the next game. The coach will be present with the team to help and analyse the game.

Example:

The coach prepares the video and shares the room's link with all the players.

When the team is ready, the coach starts the video and analyses every play.

Specific interactions:

- The team verbally communicates to analyse each play.
- The coach returns in the video timeline to explain some play that could be done better.
- While explaining something, the coach uses a visual beacon to point to a specific place in the video and direct the other users' attention.
- The coach uses the compass directions to orient the players to a specific area.
- The coach stops the video and draws in the frozen image to explain a play that can be made in that specific situation.

Appendix B

From requirements to features

Critical

Features	Requirements	VR	Desktop
Link room share	1	-	yes
Synchronized view	2	yes	yes
Voice recognition	3,4	yes	-

Table B.1: Correspondent critical requirements to features.

Important

Features	Requirements	VR	Desktop
Radar	5,6	yes	yes
Voice communication	7,8,9	yes	yes
Ping users	10	yes	yes

Table B.2: Correspondent important requirements to features.

Useful

Features	Requirements	VR	Desktop
King of the room	11,12	yes	yes
Mute and unmute users	13	yes	yes
Tunnel vision	14	yes	yes

Table B.3: Correspondent useful requirements to features.

Extra			
Features	Requirements	VR	Desktop
Personalize display and profile	15	yes	yes
Users volume	16	yes	yes
Individual mode	17	-	yes
Mini view	18	-	yes
Reactions	19	-	yes
Text chat	20	-	yes
Ask to pause	21	yes	yes
Voice commands menu	22	yes	-

Table B.4: Correspondent extra requirements to features.