FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Natural Interaction in Smartphone Virtual Reality Experiences

Leonor Martins de Sousa

Mestrado em Engenharia Informática e Computação

Supervisor: Rui Rodrigues

Co-Supervisor: Teresa Matos

July, 2023

Natural Interaction in Smartphone Virtual Reality Experiences

Leonor Martins de Sousa

Mestrado em Engenharia Informática e Computação

July, 2023

Abstract

Virtual Reality (VR) has been evolving over the last few years, but most of the hardware still remains inaccessible to the common person. Smartphone VR, a type of VR where very simple and cheap headsets are combined with the everyday smartphone, surged to make VR more widespread and easily accessible. However, smartphone VR still presents a big shortcoming: due to it being constrained by smartphone hardware, the interactivity levels in smartphone VR applications are minimal. This poses a significant issue given that the lack of interaction, and in specific, natural interaction, can lead to a lesser sense of presence, perspective-taking and enjoyment of the user, which are crucial in particular contexts, as is the case of immersive journalism and social experiments.

Taking into account the smartphone's available inputs, we propose to assess what different kinds of natural interaction can be applied in the context of smartphone VR. We also aim to study how these interactions can impact the user experience, namely in terms of immersion, presence, empathy and enjoyment. Considering previous research about interaction in virtual environments, we focus on three different methods of natural interaction that show potential: speech, which can be detected by smartphone-embedded microphones; gestures that include the head, whose detection can be done using motion sensors present on most smartphones; and finally, simple hand and/or arm gestures that can be captured by the smartphone's back camera.

We designed and implemented a browser-based application, that allows users to go through a virtual scene, composed of different 360º video sections. At certain points of the application, the users are asked to complete interaction tasks that can be of one of the three different pre-defined methods. The recognition of different types of interaction is done in a simple and straightforward manner. We developed as well one use case, represented by two different scenarios that envelop the user in a real-life situation from a main character's point of view. We finally carried out tests with users, with 22 participants, to evaluate our solution, both in terms of the impact of the natural interaction on the user's experience and the efficiency of the produced software.

By analysing the results obtained during the user tests, we finally drew the necessary conclusions and answered our guiding questions. In terms of usability, the outcomes varied according to the interaction methods. Although some methods were more successful than others in terms of robustness in detection, in general the perceived intuitiveness, easiness of use, naturalness and clarity demonstrated good results. As for the impact on the user's experience, the results are quite positive for all the methods of interaction studied. We hence show that, with some future improvements in mind, there is potential to elevate smartphone VR experiences with the use of natural interaction.

Keywords: Smartphone Virtual Reality, Natural Interaction.

Acknowledgements

This endeavour would not have been possible without my supervisors, Rui Rodrigues and Teresa Matos. Thank you for putting your faith in me; for supporting me; for the kind words but also for all the criticism that made me grow; for being flexible to the roller-coaster that I know I sometimes am. You are, with no doubt, two of the best educators that I've had the pleasure of crossing paths with, and I cannot thank you enough for giving me the opportunity to go on this journey with you.

My sincere gratefulness to everyone that dedicated some minutes of their time to this dissertation. To my colleagues at GIG, with whom I shared ideas, tips and, sometimes even more important, frustrations. To my four "actors" that made my scenarios come to life. To the wonderful people at DEI, for assisting me with all the logistic obstacles during the filming and user tests. And to all the user tests participants, without whom my research questions would remain unanswered.

Lastly, words cannot express my deepest gratitude to the people that have been my biggest pillars for my whole life, but in particular for this last year: my family and friends. In specific, I need to thank my partner, Rodrigo. You came into my life at the beginning of this journey, and yet I cannot imagine how this all would have been possible without you. Thank you for being my biggest cheerleader, my proofreader, my source of inspiration and smiles, even on the days when nothing went right. And last but not least, to my mother. When, one year ago, I was unsure of what path to follow and afraid to fail you again, you gave me your trust, and, like you have been doing for all my life, you sacrificed a little bit of yourself so I could go on to live my dreams. Thanks to you, I stand here today: proud and accomplished!

Leonor Martins de Sousa

"You never really understand a person until you consider things from his point of view... Until you climb inside of his skin and walk around in it."

Harper Lee, To Kill a Mockingbird

"People have stars that are not the same. For some, who travel, the stars are guides. For others, they are no more than little lights. For others, who are scholars, they are problems. For my businessman, they were of gold. But all these stars are silent. You, you alone, will have the stars as no one else has them... (...) When you look at the sky, at night, because I'll be living in one of them, because I'll be laughing in one of them, it will be to you as if all the stars are laughing. You, only you, will have stars that know how to laugh!"

Antoine de Saint-Exupéry, The Little Prince

Contents

List of Figures

List of Tables

Abbreviations

1D	One-Dimensional
2D	Two-Dimensional
3D	Three-Dimensional
API	Application Programming Interface
CSV	Comma-Separated Values
DSRM	Design Science Research Methodology
GPS	Global Positioning System
HMD	Head-Mounted Display
HW	Hardware
ITQ	Immersive Tendencies Questionnaire
JS	JavaScript
JSON	JavaScript Object Notation
LDTM	Look Down To Move
PI	Place Illusion
POV	Point Of View
PQ	Presence Questionnaire
Psi	Plausibility Illusion
SW	Software
URL	Uniform Resource Locator
VE	Virtual Environment
VR	Virtual Reality
WGK	Word-level Gesture Keyboard

Chapter 1

Introduction

Although VR technology is evolving rapidly, it is still far from accessible to the general public. Smartphone VR surged as a more accessible alternative to the existing complex and expensive VR systems, including desktop VR and standalone VR. In smartphone VR, the user's smartphones are placed inside the headsets, and within this category, we can consider two different types of systems. Firstly, not so relevant to our study, the type of system where the headsets include hardware that is used by the system, like cameras and sensors, as is the example of the Samsung Gear VR. Secondly, and what we are referring to when mentioning smartphone VR further in this study, systems where a regular smartphone is inserted into a headset without any active components, that serves as a case or holder, as is the case of the commonly known Google Cardboard.

1.1 Motivation

Despite the introduction of smartphone VR a few years ago, the interaction within it is still quite limited, especially in cases where there is no use of external hardware besides the smartphone HMD (head-mounted display). On one hand, the user is not able to use the mobile screen directly to interact with the content. The motion sensors, although promising, are currently mostly used for navigation purposes (and sometimes text input). Finally, while there are some scientific studies done using the mobile's intrinsic hardware (for example the back camera), other everyday extra hardware (like smartwatches or headphones) or even the cardboard itself; they appear to be still in the research phase and without many concrete applications.

Moreover, some of the existent means of interaction are not natural; this is, they are not easy to learn or intuitive to the normal user. For example, in the simple case of grabbing an object, it will be natural to reach out with your arm and then do a grabbing motion with your hand. For the same case, it will feel quite unnatural to use a controller to point to the object and then to press a button to grab it. This characteristic of interactions, their naturalness, can have an impact in the senses of immersion and presence, which might be crucial in some applications and/or contexts.

An example of such is the use of VR in social experiments and immersive journalism, among other similar applications. In these cases, it is expected that the users develop empathy towards a

particular group of people by experiencing their realities (which differ from their own), in a space that is safe but that elicits an emotional connection.

Due to the complexity of the tasks these applications bring, there seems to be a compromise between the simplicity of the system and the extent of the interaction. On one hand, there are systems that have a high level of interactivity but that are complex, requiring extra hardware and being less accessible. This is the case of systems with interaction using the entire body through body tracking. On the other hand, systems with simpler technologies lack in terms of the types of interaction available for the user. Most times, these systems only allow for interaction using the movement of the head for navigation, in a passive way.

Considering all of this, there seems to be an opportunity to explore how we can have a natural and active interaction in VR while maintaining high accessibility and simplicity of the system by using no other hardware than the one included in current smartphones.

1.2 Problem

Although smartphone VR allows for VR to become more accessible, it presents some limitations in terms of the possible types of interaction. The usual smartphone VR system relies on the smartphone's input and output methods. The headsets are commonly just a simple plastic or cardboard casing for the mobile. Other types of external hardware, like hand controllers, gloves or full-body suits, are also not used since the accessibility of the system is a priority.

Hence, and considering the common smartphones of today, in terms of input, we seem then to be limited to the use of the motion sensors, the (rear and front) cameras, the microphone, the GPS receiver, the touchscreen, the buttons and the barometer.

It is worth noting that not all of these input methods might be relevant or even usable in the context of natural interaction and so it is necessary to evaluate each one of them, something that will be done in a further section of this document.

An important problem arises from this scenario: how can we use these available input methods to emulate natural interaction in smartphone VR?

1.3 Research Questions and Goals

With this problem in mind, we propose to answer two main research questions:

R1: What methods of natural interaction can be used in smartphone VR?

R2: Does natural interaction in smartphone VR contribute to the user's experience, in terms of immersion, presence, empathy or enjoyment?

To answer these questions, we define two main goals for this study:

O1: Explore different natural interaction methods, taking into consideration the smartphone input limitations.

O2: Evaluate and compare the effects of the implemented natural interaction methods on the user's experience, in terms of immersion, presence, empathy and enjoyment.

1.4 Outline

This first chapter (chapter 1) introduces the scope of the research, pointing out the relevant context and correspondent motivation. It also presents the problem we try to solve and the respective research questions and goals.

Chapter 2 then puts forward an analysis of the literature related to the defined problem, starting with a brief introduction to virtual reality and related concepts which are more relevant to this research (immersion and presence). Following an investigation into natural interaction and its impact on presence, we proceed with a more technical investigation, exploring past implementations that can be of use in the scope of our project.

We continue to chapter 3 with the proposal of a solution to our problem, including the methodologies adopted, the analysis and choice of the methods of interaction to be further explored, the requirements elicitation, the system architecture and the roadmap for the execution of the solution, from the beginning of the development phase to the delivery of the results.

In chapter 4, we describe the development of the proposed solution. We start by conveying the technologies and tools that are used for the development. After we give an overview of the system's workflow, we describe the implementation details for each of the three chosen types of interaction. Then follows a description of the input parameters and files of the implemented application as well as the produced output (data logs).

Chapter 5 goes on to detail the evaluation process of the implemented solution, delineating the design and implementation of a use case. It also describes the user tests, including their design, structure, data gathering and execution. Lastly, this chapter presents the results of the evaluation phase.

Finally, chapter 6 sets forward some conclusions drawn during the entire research, as well as some notes on its limitations and future work possibilities.

Chapter 2

Literature Review

In this chapter, we present an overview of previous studies in areas related to our research. We briefly introduce virtual reality, its story, evolution, and main limitations. After, we study two concepts deeply connected to VR: immersion and presence, including some of its subconcepts and ways of achieving them. We then focus on interaction within virtual environments, particularly the concept of natural interaction and its importance in specific contexts. Finally, considering the smartphone's input limitations, we look into works that implement interaction, natural or unnatural, but with the potential to be reused in different contexts.

2.1 A Glance at Virtual Reality

The concepts behind Virtual Reality can be traced all the way back to the 60s with Sutherland's introduction of "The Ultimate Display" [62] and, later on, of the first head-mounted display [63], a complex display system that allowed the user to be surrounded with 3D information through the use of a pair of special spectacles. However, it was not until 1988 that the term Virtual Reality was first coined by Lanier. He defined it as "a technology that uses computerized clothing to synthesize shared reality. It recreates our relationship with the physical world in a new plane (...) It only has to do with what your sense organs perceive." [31]

Fast forward two decades, the Oculus Rift project, which intended to create an affordable VR system for the masses, got funding in 2012. This marks the beginning of the second wave of VR that brought the emersion of new VR products and technologies, accessible to the general public [5]. With the objective of making VR more accessible, smartphone VR surged shortly after with the release of the Google Cardboard [35] in 2014 and the Samsung Gear VR [52] in 2015.

Since then, VR has seen incredible growth and research predicts a compound annual growth rate (CAGR) of 15% between 2022 and 2030 [50]. With applications that go from entertainment and education to health and sales, for example, VR still has challenges in need of better solutions [67].

Some of the current limitations of VR have to do with the high cost of most sets, the graphic capabilities of the systems and even the risk to eye health, as well as associated headaches and neck

pain [67]. Accessibility for disabled people seems to be still lacking in most cases and there is still room for improvement in terms of making the devices more ergonomic [38, 21, 11]. Additionally, VR sickness remains as a problem to be fully solved [59]. For the specific case of smartphone VR, there is still a limitation in terms of the possible levels of interaction, due to the mobile hardware constraints [20].

2.2 Immersion and Presence in VR

Two of the concepts that are deeply interconnected with Virtual Reality are immersion and presence. In this section, we explore these two concepts and subconcepts. We also explore the ideas of embodiment and perspective taking, which are correlated.

2.2.1 Immersion

Murray defined immersion as "the sensation of being surrounded by a completely other reality (...) that takes over all of our attention, our whole perceptual apparatus" [39].

Slater and Wilbur defined immersion as "a description of a technology (...) [that] describes the extent to which the computer displays are capable of delivering an inclusive, extensive, surrounding and vivid illusion of reality to the senses of a human participant". They argue that immersion is objective and quantifiable, as it is a property of the system [58].

Bringing together the perspectives of different authors, Nilsson et al. created a taxonomy based on three dimensions, showed in figure 2.1 [40]:

- system immersion: the ability of the technology to deliver senses, independent from the user's reaction
- narrative immersion: the sensation of being absorbed by the story and its components
- challenge-based immersion: the sensation of being absorbed by the tasks/challenges presented

2.2.2 Presence, Place Illusion and Plausibility Illusion

Associated with the term immersion is many times the known concept of presence. Presence is often simply defined as a sense of "being there".

For the specific case of virtual environments, Witmer and Singer have said it to be the experience of being in a computer-generated environment rather than the actual physical location. They also add that there are two necessary conditions for the existence of presence: immersion and involvement, which is related to the focus of the attention and energy of the user. Furthermore, they introduce several hypothesized factors that influence presence, as well as two questionnaires (the Presence Questionnaire (PQ) and the Immersive Tendencies Questionnaire (ITQ)) that can be

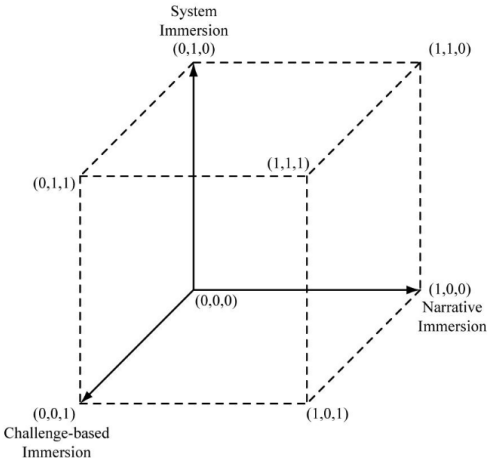

Figure 2.1: Dimensions of immersion according to Nilsson et al. - "Illustration of the proposed taxonomy of existing conceptualizations of immersion. The three axes represent the extent to which interaction with a system involves system immersion (vertical), narrative immersion (horizontal), and challenge-based immersion (depth). The degree to which each type of immersion is presented is represented on a scale from 0 to 1, where 0 represent absence and 1 represents the highest possible level of immersion" [40].

used to evaluate the sensed presence in a VE by an individual and the capacity of an individual to be involved and immersed, respectively [71].

Slater introduced the concept of Place Illusion (PI) as a more specific term for presence: “the strong illusion of being in a place in spite of the sure knowledge that you are not there”. Additionally, he presented the term “Plausibility Illusion” (Psi) and differentiated it from PI: it is “the illusion that what is apparently happening is really happening (even though you know for sure that it is not)” [57].

Presence can be achieved through different dimensions, including spatial sound, realism (which can relate to system immersion), interaction (under which we can consider natural interaction) and proprioceptive matching (also referred later on as sensorimotor correspondences), which is deeply interconnected with the concepts of embodiment [58].

2.2.3 Embodiment and Perspective Taking

Although immersion, presence and plausibility illusion are a component of any virtual environment, they can be more crucial in some specific application contexts than others. More specifically, we can talk about cases where an embodiment of the avatar and/or the perspective-taking of the participant is the main goal.

In the context of VR, embodiment can be described as the sense that the avatar’s body is the user’s body. Embodiment can be achieved through two different methods. The first, sensory signal correspondences, has to do with the correlation with what the user feels and sees (for example, when the user sees a fake hand being touched synchronously as their own hand is, that creates a sense that the fake hand is their own). On the other hand, sensorimotor correspondences relate to the synchronization between the avatar’s and the user’s own physical movement (for example,

when the user sees the avatar's hand moving synchronously as his own is in real-time). In both cases, synchronization is key to the illusion of ownership of the virtual body [6].

Furthermore, perspective-taking is about taking a different point of view (in relation to a person's own). It can take two different forms: "you can imagine how the other perceives the situation and how that person feels as a result (imagine other), or you can imagine how you would perceive the situation were you in the other's position and how you would feel as a result (imagine self)" [9]. Additionally, in the context of VR experiences, you can experience two different types of perspective: first-person POV (from the actor who performs a certain action) or third-person POV (from a passive observer, who witnesses the actor performing a certain action) [3].

These two concepts can be applied to different expected (psychological) results.

Firstly, we might want to allow the user to perceive themselves in a different way. Osimo et al., for example, used a virtual environment for a self-counselling application where the study participants discussed a problem with a counsellor. The counsellor would vary between a lookalike avatar of the participant and a representation of Freud. It was found that seeing oneself from a different perspective/body tended to improve the participant's mood and happiness, which may indicate an improvement in self-compassion [42]. Additionally, Porras-Garcia et al. work showed that body exposure in VR can help improve emotional, cognitive, and behavioural responses to body image in patients with eating disorders [47].

Secondly, we may wish the user to experience a situation that they already experienced but from a different perspective, which can lead to a change in perceptions, attitudes and behaviours in those kinds of situations [8, 1]. In specific, we can apply this to the reduction of biases, namely gender [34, 3] and race [8, 30] biases. We can also use these concepts in the context of the prevention of aggressive behaviours, such as the case of rehabilitating and improving emotion recognition in domestic violence perpetrators [53, 54].

Thirdly and lastly, the user might want to experience a situation they would not be able to experience (safely) another way. Immersive journalism (a form of interactive journalism where the participant is able to experience the news story in a virtual environment) is a clear use case of this. Some already existing news stories cover topics like hunger [18], displacement due to war [56], solitary confinement [12] and even glacier melting [17]. In 2010, de la Peña made the hypothesis that immersive journalism could create a bigger emotional and empathic response in the audience, making use of PI, Psi and perspective-taking [13]. Nowadays, some of these news stories already show us that that is indeed true [64].

Furthermore, a study in immersive journalism, by Sundar et al., showed that presence is correlated with credibility, memory and story-sharing intention. In fact, "when stories are emotionally powerful and also richly narrated, they may override the capacity of the technological factors", although these last can play a bigger role in the absence of the first. Besides that, realism plays a big role in trust and perceived source expertise, so it must be balanced with the sense of presence [61]. However, it is worth noting that, considering that the user does not have a decision in the direction of the narrative, since the story told must always correspond to the facts, the interaction levels in immersive journalism might be limited [29]. The same limitations in interactivity also

seem to be observed in the other applications mentioned above and so, it is important to explore how we can increase the interactivity levels in these situations where we wish to elicit not only presence but also an emotional response and empathy from the participants.

2.3 Eliciting Presence through Natural Interaction

Natural interaction can be defined as the "exploitation of natural (i.e. intuitive, familiar, innate, universal, cross-cultural, etc.) skills or abilities for controlling, either implicitly or explicitly, a computer system" [25]. Although natural interaction can be achieved in different ways, what really characterizes it is that its design is adapted to normal human behaviour and not the other way around [51]. Some commonly seen forms of natural interaction are gestures, speech, facial expressions, eye movement and even touch, among others [55]. However, since we are focusing on interaction in smartphone VR, we will limit ourselves to referring to the ones that are relevant within that context.

Georgiadis and Yousefi, for example, used natural gestures in the context of a smartphone VR game and showed that the natural interaction technique improved their presence and engagement in the game when compared to the use of buttons [16]. Additionally, an experiment conducted by Bailey et al. showed that, in comparison with arbitrary gestures, natural gestures lead to a higher sense of immersion, presence, usability, sense of control and interface quality [7].

Narrowing down the movement to the face, Ilves et al. compared the use of head movement and facial expressions (with computer vision techniques) to a joystick as input methods in a gameplay context. They found that, although the joystick was more effective, the first (and more natural) method enhanced the user's experience in terms of entertainment, interestingness, challenge and immersion [26].

On a different note, it is worth mentioning Osking & Doucette's work in establishing a connection between voice control systems and their effect on users' emotions. As a use case, they used a dialogue-based game where, at certain predefined breakpoints, the player had to choose from a selection of pre-written dialogue lines, either by reading them or selecting them with the help of hand controllers. By comparing the use of speech recognition with a traditional system (hand controllers), they discovered that the voice interface was more enjoyable and more emotionally impactful, enhancing the user's embodiment of the game's protagonist [43].

Breath can also be used as a natural interaction technique. Sra et al. created BreathVR, where the breathing of the user controls their actions inside a VR game. Different breathing patterns (like gale, waft, gust and calm breathing) are associated with different actions in the game (like stopping, slowing down, freezing, expelling fire, etc.). A user study showed a bigger sense of presence and better game experience with the breathing interface compared with the typical use of buttons in hand controllers. It was also concluded that the breath gestures need to be well contextualized with the right narrative and correspondent game effects [60].

These studies help us see not only the use of some different natural interaction methods but also how they can improve presence and even emotional response in virtual environments. It seems

now relevant to explore how methods like these (and others) can be implemented in the context of smartphone VR.

2.4 Interaction Methods in Smartphones

In comparison with other more "complex" types of VR, smartphone VR presents some limitations (in hardware) that can restrict immersion and, more concretely, system immersion, since narrative and challenge immersion do not depend on the technology used. Although the restraints in resolution, field of view, optics and positional tracking do not seem to elicit differences in terms of presence, usability, satisfaction and learning outcome [44], the lack of hand/body trackers and other specific hardware can limit the interaction between the users and the virtual environments.

Considering that among all the smartphone's input possibilities, the cameras (back and front), motion sensors and microphone seem to be the most explored and relevant in the context of smartphone VR, we will start by exploring some pertinent work focused on those three input methods and then move on to some other studied interaction types.

2.4.1 Using Cameras

Most smartphones nowadays include front and back cameras, with their quality (and the number of cameras) increasing with time [48]. These cameras can be used as input in smartphone VR for different types of applications and contexts.

The back camera can be used, for example, to detect the user's hand and/or fingers, which can be particularly useful for simple selection tasks or even more complex tasks with interaction via micro gestures.

Ishii et al. developed FistPointer, in which a smartphone's back camera detects the movement of the user's thumb. The user then can move the pointer by moving their hand and make their selection by folding their thumb (which is equivalent to a "click") [27]. On the same note, Luo & Teather compared three types of selection methods using the smartphone's back camera. Through a user study, they found that the ray techniques (head ray and finger ray) performed considerably better than the direct touch one (air touch) due to the imprecision in depth [37].

In the micro gestures area, Li et al designed a set of microgestures for system/video player control. They used, firstly, the knowledge of two professional ergonomists and, later on, a user study to come to 19 proposed microgestures for 20 commands, where comfort and naturalness were valued [32]. Similarly, Wu et al. designed a set of gestures for immersive VR shopping applications. They also performed a user study to compare the use of said gestures with other traditional VR interaction methods (virtual hand controller and ray-casting), with the gestures having outperformed the other methods in terms of user experience and presence without loss of performance [72]. Although both these studies used external cameras to capture the participants' movements, their conclusions can be equally applied to smartphone cameras.

Adding a new simple and affordable piece of "hardware" to a normal cardboard headset, Ahuja et al. developed a method to digitize the whole body of the user: MeCap [2]. Two mirrored half-spheres mirror the body and are then captured by the mobile's camera. This allows capturing not only the body pose of the user but also their hand gestures, mouth state and colours of skin and clothes. Although MeCap still presented a number of limitations, including latency, reliance on environment conditions and motion blur, among others, it presented an accuracy rate of about 80% for body pose estimation and 60% for mouth state, though the accuracy, in this case, increases to more than 95% when considering only two different mouth states.

Moving to the smartphone's front camera, its use is more limited, considering it will be facing the user's face in a normal use case and, in a very close range. However, this does not mean it can not be used for interaction. For example, Hakoda et al. used the front camera to develop a system for eye/gaze tracking. The tracking was tested in a selection task application where lower latency was essential. Even so, it is worth noting the several limitations of the system, namely the different positions of front cameras in different smartphone models, the dependency on the light of the environment and the fact that it is only possible to track one of the eyes [19].

2.4.2 Using Motion Sensors

Although less widely known hardware parts, the motion sensors are important components of any VR system, that are present in most smartphones nowadays. They can monitor "device movement, such as tilt, shake, rotation, or swing" [4]. The motion sensors are used in almost all systems to detect head movement and change the display accordingly.

Moreover, the smartphone's motion sensors are commonly used for navigation in virtual worlds. Tregillus & Folmer, for example, introduced VR-STEP, a method where the motion sensors are used for step detection and gaze detection (movement direction), respectively, allowing the user to walk in a virtual environment. This method proved to be more immersive and easy to learn than the baseline LDTM (look down to move) [66]. Tregillus et al. also presented a head tilt navigation method, that outperformed the more common joystick method both in presence and performance [66].

The motion sensors can also be used in the detection of head gaze, which can be applied in selection and text-entry tasks. Lu et al. compared three different text-entry methods for VR (DwellType, NeckType and BlinkType), all three using head gaze to "navigate" the keyboard. The study found BlinkType to perform better and to produce a better user experience, although it presents the limitation of needing an eye tracker [36]. Additionally, Yu et al. introduced a word-level gesture keyboard (WGK) for head-based text entry (GestureType) and compared it with more usual text-entry methods (TapType and DwellType). GestureType performed better, with an accuracy of 97%, and it was shown a learning effect, where users improve their speed of entry without sacrificing accuracy as they get more familiar with the method [74].

With what can be considered a more out-of-the-box idea, Yan et al. developed Cardboard-Sense, designed for the Google Cardboard headset. This interface allows the user to interact with the virtual environment by tapping in different spots of the cardboard surface. The taps produce

vibrations than can be detected by the gyroscope and identified with a deep learning model, with an accuracy of 98.9% [73].

2.4.3 Using the Microphone

Although external microphones can be used, all smartphones (and even older phone models) include embedded microphones that serve as input hardware and hence can be used in smartphone VR systems.

Pick et al. designed SWIFTER, a text input system that combines speech recognition with a point-and-click interface for handling input. Although there were no significant differences in terms of performance, it was noted that users preferred SWIFTER to a typical smartphone keyboard input [46].

On a different note, Hepperle et al. explored the use of speech recognition in the context of task solving, comparing it with other two interfaces: 2D and 3D. The tasks involved not only text input but also selection and object manipulation in a virtual environment. A user study revealed the choice of the "better" interface always depends on the prioritary needs of each application: for a bigger sense of presence and playful experience, 3D interfaces are recommended; for the manipulation of a large amount of objects, 2D interfaces allow for faster and more accurate results; if ease of learn is intended and/or the task needs a lot of text input, then speech is a better choice. Despite this, it is worth noting that the speech interface induced the greatest overall satisfaction, followed by 3D with 2D in last place [22].

Although the microphone is most commonly seen used in speech interfaces, it is not limited to voice interaction.

Such is the example of the work produced by Sra et al., previously mentioned. With their new natural interaction technique, BreathVR, the actions of a player inside a VR game are controlled by their breathing. The user's breathing is captured by the microphone, converted to a waveform and compared with baselines for categorisation into pre-established breathing "types". It was shown that the breathing interface was generally preferred by the users. However, no performance tests were made to better compare both methods [60].

On another hand, Chen et al. developed GestOnHMD, designed for the Google Cardboard. In this system, the users draw pre-defined gestures in the cardboard surfaces. The sounds produced by the touch are then captured by the microphone and processed by a deep learning model that classifies the gesture with an accuracy of 97,7%. The set of pre-defined gestures was designed with a user study, to ensure their naturalness. The biggest limitation of the system seems to lie on the sound variance within different headsets and error handling (especially for false positives) [10].

2.4.4 Using Other Types of Input

Although the magnetometer is only present in some smartphones, the initial version of Google Carboard included a magnet that allowed the user's to interact with the environment by simulating

a tap on the screen. This characteristic was further explored by Li et al. with ScratchVR. By slightly altering the Cardboard, they created 10 ridges in a circular shape so that the user can move the magnet from rigde to ridge. The magnet's position is then detected using a Support Vector Machine. It was shown that ScratchVR was more effective than gaze in large menu use [33].

Moreover, there are also some studies on the use of some common smartphone hardware like smartwatches and headphones.

WatchVR, designed by Hirzle et al. is one example. In here, the smartwatch's use in pointing/ selection tasks was tested taking into consideration two different variables: the interaction method and the wearing method. For the interaction method, it was found that the method using the smartwatch's motion sensors not only performed better, but was also preferred by the users, in comparison to the smartwatch's touchscreen. In terms of the wearing method, the hand-held controller-like method was preferred and performed better than the normal wrist-wearing method. However, it was noted that these results can be different when considering the duration of the task [23].

Similarly, PAWdio, introduced by Zayer et al. uses the smartphone headphones to track the hand position for pointing/selection tasks. Using acoustics, it is possible to calculate the distance from the hand-held earbud to the smartphone. Although it was proven than PAWdio offers a basic and affordable input with good immersion levels, it also presented several limitations. The 1D input seems to be the larger but there are also the limitations in playing audio while using the system or even the robustness of the system [75].

2.5 Summary

Virtual reality can provide alternative and appealing experiences to users that allow them to confront different realities and/or different perspectives and even change attitudes and behaviours. Natural interaction in specific can lead to a higher immersion, presence and emotional impact on the user. When the interaction is well contextualized, the use of hand, arm, or body gestures, facial expressions, speech or even breath, seems to enhance the user experience greatly. Although smartphone VR provides an accessible means to bring these experiences to the normal user, it also still presents some shortcomings in terms of interaction due to its input limitations.

Previous work indicates the use of the smartphone back camera to detect hand gestures/ movements and even facial expressions (mouth state), with the use of some extra simple hardware. The front camera, although considerably more limited and challenging, can also be used to detect eye gaze. Additionally, motion sensors allow for the use of head movements/gaze as well as steps as interaction techniques. On a different note, the microphone allows to introduce speech and breathing-based interaction, as well as gestures done on the headset, although this last option is considerably more constrained. Finally, extra hardware like smartwatches and headphones can also be used in hand movement/gesture recognition.

Although the existing studies show us the potential for interaction in smartphone VR, the correlation between their use in smartphone VR and results in terms of immersion, presence,

empathy and enjoyment is still to be further investigated. Furthermore, apart from the combination of speech and head gestures for text input tasks, the studies seem to shy away from exploring the combination of methods, only sometimes comparing them. With this in mind, there seems to be an opportunity to explore how natural interaction methods can be applied to smartphone VR and provide a means to improve the user's experience.

Chapter 3

Proposed Solution for Natural Interaction in Smartphone VR

Given the opportunities found in the literature review in the previous chapter and given the research questions presented in section 1.3, we propose the creation of a smartphone VR application where the user can interact with the environment using different natural interaction methods. This chapter provides insights into our proposed solution.

We start by offering a general description of what we intend to create in section 3.1, followed by the methodology we will follow in 3.2. We then proceed, in section 3.3, to elaborate on the reasoning behind the choice of methods of natural interaction to be further explored and implemented. After that, we define the system requirements in 3.4 and its architecture in 3.5. In section 3.6, we finally present a roadmap that will serve as a guideline for the implementation phase.

3.1 General Description

As established in sections 2.1 and 2.2, smartphone VR applications still seem to lack in terms of interaction due to input limitations, which, as established in section 2.3, can mean an unexplored potential for a higher sense of presence in such applications. Taking into consideration the use of smartphone inputs introduced in section 2.4, we propose to implement a simple smartphone VR application. This application should allow its user to visualize a virtual scene and interact with it via natural interaction methods. Given the research questions previously presented, the scene must be designed in order to allow its normal flow with no interaction at all (passive visualization) or mandatory interaction with one or more types of natural interaction. This allows for easier withdrawal of conclusions in the evaluation phase. As for the choice of which methods of natural interaction are further explored, it is elaborated upon in section 3.3 of this chapter.

3.2 Methodology

This study follows the Design Science Research Methodology (DSRM) defined by Peffers et al. [45].

We first identify the problem at hand and the motivation to solve it by exploring the existing literature and corresponding gaps in the research (chapter 2). We then move on to define the goals of the investigation, setting the solution requirements and architecture, always using the literature review as a basis for what technology to explore, in particular, which methods of natural interaction (chapter 3). The study proceeds with the design and implementation of a prototype (chapter 4) and further demonstration of the use of the solution with the design and implementation of a use case (section 5.1). Furthermore, we evaluate the solution with the execution of user tests, evaluating not only the usability of the solution (usability testing) but also the its technical performance (section 5.2). The communication of the results of this research is achieved through this document, as well as other possible future publications. Finally, in chapter 6, we discuss the need for iteration of the process as a way to improve the current solution.

3.3 Choice of the Methods of Interaction

Before moving on to the definition of the requirements of the solution, it is crucial to make a choice regarding which methods of natural interaction are worth exploring, given the constraints of the project. To analyse the possible methods in the context of smartphone VR, we consider the input methods offered by the technology, as presented in section 1.2. It is important to note that, for this analysis, the accessibility/availability of the technology is a deciding factor.

Starting with the barometer, we immediately exclude it from our list, since it is only present on some smartphone models nowadays. Although not all motion sensors (gyroscope, accelerometer and magnetometer) are present in all smartphone models, most of them use sensor fusion to obtain all the necessary data. As so, we consider that the motion sensor needs are met in nowadays smartphones. Taking into account that the smartphone's GPS receiver precision is in order of the meters [24], it also seems inadequate to use its input in VR, where the users are not expected to displace themselves (more than a couple of meters).

For the remaining input methods, we benefit from the information obtained in sections 2.3 and 2.4 to cross the interaction techniques with the smartphone's available inputs, as seen in table 3.1. It is worth reinforcing that the combination of the referred natural interaction methods is not observed in the literature.

Within our possible choices, it is worth noting that although facial expression detection is possible with the smartphone's rear camera, it takes some extra hardware (like the use of mirrored spheres [2]), which can bring complications in terms of the availability of the technology. As for the smartphone's front camera, although this is an input available in almost all smartphones nowadays, it suffers from being too close to the user's face, and possibly even obstructed, given the lack of uniformity in smartphone models.

Table 3.1: Use of smartphone inputs for natural interaction in the literature review. * indicates the need of simple extra hardware.

	Motion Sensors	Rear Camera	Front Camera	Microphone	Touchscreen	Buttons
Head Gestures	√					
Hand/Arm Gestures		√				
Facial Expression		√*				
Eye Movement			√			
Speech				√		
Breathing or Other User-Produced Sounds				√		
Touch(haptic)						

Given the time constraints of the project, we also exclude breathing or other user-produced sounds from the study. Although it has some similarities with speech interaction, the large availability of speech recognition software makes the second option a more viable one in comparison.

Finally, we reach a choice of three methods of natural interaction (and respective input methods) that seem to be worth exploring, namely in terms of viability, within the scope of this project:

- head (or full body including head) natural gestures, with detection using the motion sensors;
- hands and/or arms natural gestures, with detection using the rear camera;
- speech, with detection using the microphone.

3.4 Requirements

To guide the development of the proposed solution, as well as the design of the use case, we define a set of requirements for the system. For this phase, we consider as a use case an experience where the user goes through a narrative scenario composed of 360º videos, where there are opportunities to interact with the virtual environment without changing the course of the narrative. The requirements, based on the research questions set in chapter 1, as well as in the literature review in chapter 2, are the following:

- The environment and narrative should feel realistic to the user;
- The user should feel immersed in the environment;
- The user should have a sense of presence during the experience;
- The experience should elicit in the user empathy towards the main character;
- The experience should be enjoyable to the user;
- The user's interaction with the scenario should feel natural and intuitive to them;
- The user's interaction with the scenario should be easy to understand and execute by the user;

- The user's interaction with the scenario should feel contextualized within the narrative;
- The user should be able to advance in the narrative, even if not able to perform the interaction;
- The experience should not cause nausea to the user;

3.5 System Architecture

With all the constraints presented up to now, we design a simple architecture for the system.

We consider three main components within the architecture: the software implemented by us within the context of this study; the smartphone hardware components that will be necessary for this project; and the software developed by third parties, that will be used to obtain data from the input hardware and/or aid in the recognition of certain types of interaction. Figure 3.1 illustrates the architecture of the system, with these three components and the way they interconnect.

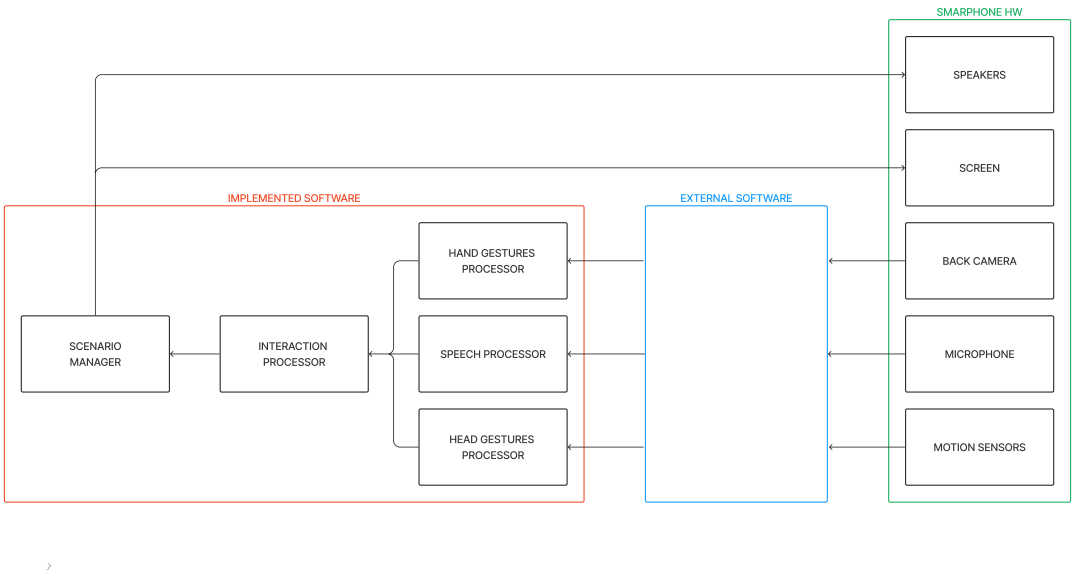

Figure 3.1: System architecture.

Within the "Implemented Software" component, we have a main subcomponent "Scenario Manager", which is responsible for starting the application and running the scenario, moving from across its several moments (with and without interaction). We also have a subcomponent for each of the three methods of interaction: "Hand Gestures Processor", "Speech Processor" and "Head Gestures Processor", which are responsible for detecting their type of interaction and acting accordingly. The subcomponent "Interaction Processor" does the connection between the Scenario and each of the three interaction types.

Regarding the “Smartphone Hardware”, as it was mentioned before, there are three different subcomponents used for input to the system: the “Back Camera”, the “Microphone” and the “Motion Sensors”. Additionally, the “Speakers” and the “Screen” work as the system outputs, as they are responsible for playing the scenario content.

3.6 Roadmap

With the requirements and system architecture defined, we plan the next steps and tasks of the research:

- Research the 3 interaction methods (speech, head gestures and hands gestures);
- Exploration of the technology to be used;
- Implementation of the Scenario Manager;
- Implementation of the Interaction Processors;
- Use Case design;
- Use Case implementation;
- Evaluation design and preparation;
- Testing and error fixing;
- User tests execution;
- Results analysis and conclusions;
- Document writting-up;

3.7 Summary

In this chapter, we proposed the creation of an application that allows a user to experience a 360º video-based smartphone VR experience and interact with it using three different methods of interaction: speech, head gestures and hand gestures.

In the next chapter, we now proceed to describe the implementation phase of said solution.

Chapter 4

Development of the Smartphone VR Application

The developed solution, according to the proposal in the previous chapter, is a browser-based application, that allows to user to go through a 360º video scenario, in VR mode. The scenarios can or not include interaction, which serves as a means of comparison during the evaluation phase. In the case where the scenarios include interaction, the application requests, at certain points, a specific interaction "task", which the user needs to perform in order to advance in the scenario.

In this chapter, the application and correspondent implementation are described. First, we specify the tools and technologies used for the development of the system, in section 4.1. We then proceed to give an overview of the system and define the high-level workflow of the application, in section 4.2. Afterwards, in section 4.3, we present the implementation details for each of the interaction methods, including limitations and development decisions. In section 4.4, we identify all the inputs the application takes, namely its URL parameters and JSON input files. Finally, in section 4.5, we point out the output of the application, namely the data logs that are later used in the evaluation phase.

4.1 Development Tools and Technologies

Given the requirements of the system, defined in section 3.4, we decided to make the application browser-based. Browser-based applications have the advantage of being easily accessible on different types of devices, as well as not requiring any type of installation process. Furthermore, they simplify the development phase, since browsers nowadays, namely Google Chrome, have a myriad of tools for debugging and port forwarding that allow for instant launch of web applications.

In terms of the technologies used for web development, we decided to go with React and node.js. The compatibility of these technologies was decisive since they are compatible with all modern browsers [28]. However, other aspects factored in: React and node.js are widely used, which means the support community is quite large; they are easy to learn and there is already some personal experience with the technology, which facilitates the development.

We also employ the JavaScript WebXR API, which is used to "support rendering 3D scenes to hardware designed for presenting virtual worlds" [69], such as the Google Cardboard. Although still in the experimental phase, this API simplifies the visualization of Virtual Reality scenes in mobile devices.

Finally, Three.js, a cross-browser JavaScript library that simplifies the use of WebGL to draw 3D objects [65], is also used. This library allows to very easily create a 3D inverted sphere where the scenarios are displayed to the user. It was chosen due to the large amount of examples available which facilitated the fast beginning of the development.

For the implementation of the three interaction methods, other technologies are used that are important to consider when defining the compatibility of the system, namely the Web Speech API, the deviceorientation and decicemotion (in the "mwilber/nod.js" plugin) events from the Window API, and the react-camera-pro, all of which are further explored in section 5.3. With all of this we can define the final compatibility of the system, in table 4.1.

Table 4.1: Compatibility of tools and technologies used in most common browsers.

Technology	**Browser Version**				
	Chrome Android	Firefox for Android	Opera Android	Safari on iOS	Samsung Internet
WebXR [69]	79	NO	57	NO	11.2x
Speech Recognition [68]	33	NO	20	14.5	2.0
deviceorientation [15]	18	6	12	4.2	1.0
devicemotion [14]	32	6	18	4.2	2.0
react-camera-pro [49]	114	113	73	11	6.2
Final	114	NO	73	NO	11.2

4.2 System Overview and Workflow

The system is implemented according to what is set in chapter 3, using the technologies and tools described in the previous section. The components of the system and respective relations are seen in figure 4.1. As for the workflow from the point of view of the user, it is presented in figure 4.2 and described next. We also include some aspects related to the instrumentation for the user tests phase.

The application is initialized, first showing a screen that allows the user to start the experience by clicking a button on the screen, as seen in image 4.3.a. The scene is then initialized and checks the validity of the URL input data, which contains parameters that allow to retrieve the scenario's details and will be later on clarified in more detail. If not, the screen shows a message saying that the URL props are not well-defined or valid, as shown in figure 4.3.b. If the URL data is indeed valid, the participant is shown a screen that says "wait till the experiment starts", as seen in figure 4.3.c. From here, the system requires a press on key "S" to enter VR mode and start the scenario itself. The application was implemented this way to facilitate our control during the user tests,

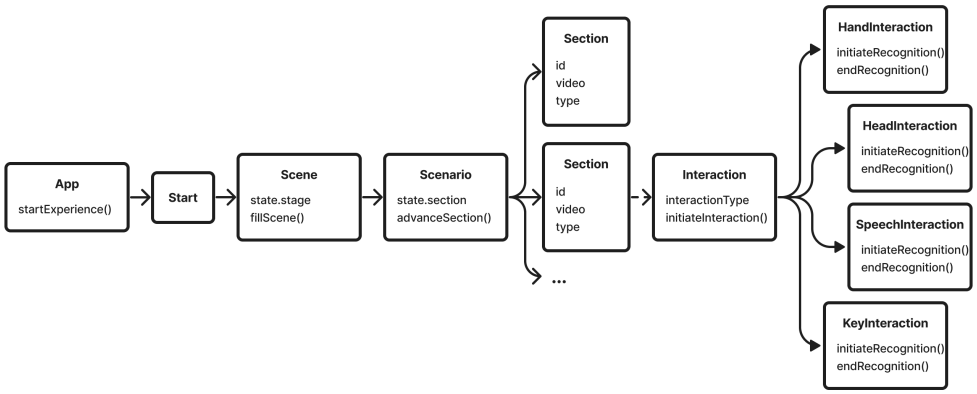

Figure 4.1: Components of the system and respective relations.

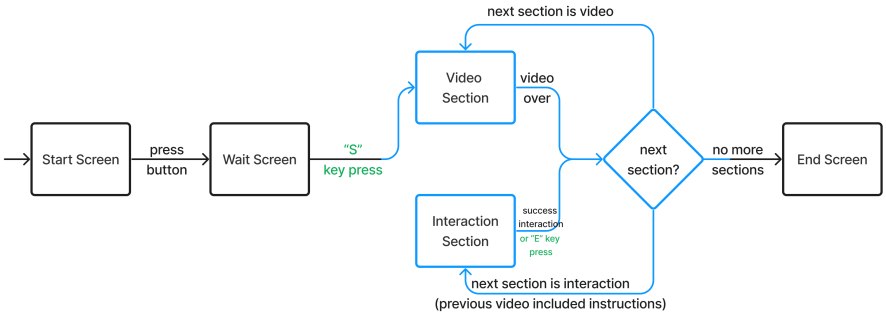

Figure 4.2: Workflow of the system from the user's perspective. The blue corresponds to when the system is in VR mode. The green text corresponds to instrumentation steps.

where the execution is in a computer environment but the visualization on a smartphone, as will be detailed in a future section of this document.

When the VR mode is initiated in *Scene.js*, an "inverted" sphere is created using *Three.js*, using the method seen in figure 4.4. This sphere is the display means for the 360º content, which is showned to the user like seen in figure 4.5. Each scenario is composed of several sections, which have a 360º video associated, and may be of two types: (just) video or (video with) interaction. *Scenario.js* has all of the scenario's corresponding sections and manages which section is currently active (being shown to the user), as well as the switch between different sections, once they have ended. *Section.js* represents the behaviour of the section which may or not have an *Interaction.js* component associated. Its in *Section.js*, that the video element in the *Three.js* sphere is updated to correspond to the new section. A section is considered to have ended:

- if the section is just video: when the video is over;
- if the section contains interaction:

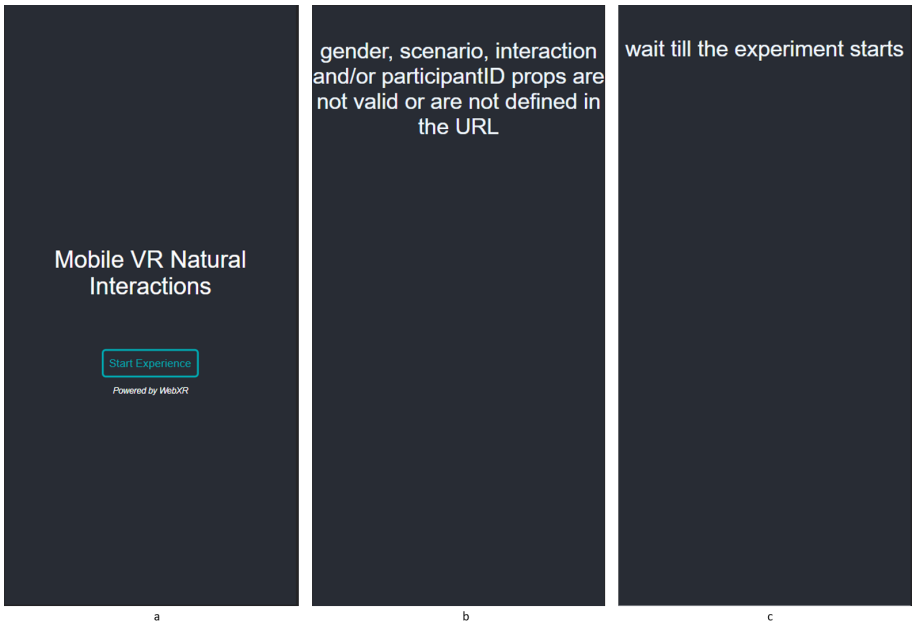

Figure 4.3: Screens of the initialization of the experience application. a. Screen asking user to start; b. Screen when the URL parameters are not valid; c. Screen when experiment is ready to start.

– once the interaction task is completed and detected with success;
– or, for instrumentation of the user tests, when the key “E” is pressed.

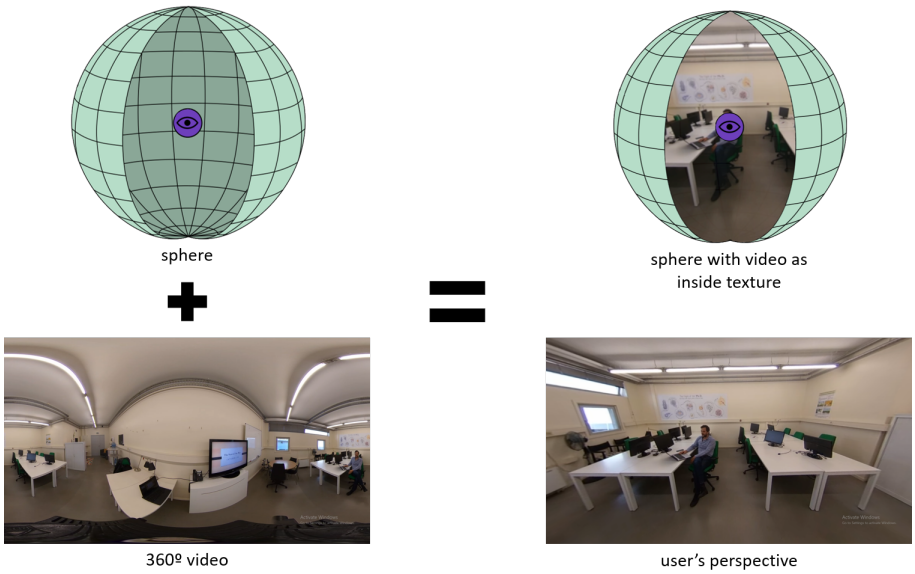

Figure 4.4: Application of video to "inverted" sphere.

Each interaction can be of four different types, representing the three interaction methods described before, plus an extra interaction, *KeyInteraction.js*, which was implemented to aid during the user tests and that comes to and end only when key “E” is pressed. The scenario is over when

Figure 4.5: Example of a video played in VR mode.

its last section has ended. Then, the VR mode is disabled and the user is shown a final screen, thanking them for their participation in the experience.

4.3 Implementation of the Interaction

Given that each method of interaction is very distinct from one another, each of them was implemented independently in its own module. In this section, we cover the implementation details of each of those methods: speech, head gestures and hands gestures.

4.3.1 Speech

When exploring voice and speech interaction, several options were considered in terms of what aspects can be detected and used as natural interaction. The first one that was identified with related work, and that became our choice for this phase of the implementation, is the detection of the content of the speech, this is, speech-to-text recognition. However, and although we decided not to cover it during this project, it is also relevant to mention different voice parameters, like volume or tone, which are inherent to any natural speech interaction and can be laid as possible future work.

For asynchronous speech recognition, the Web Speech API comes in handy, particularly the Speech Recognition interface. This API allows for a quick setup of the recognition software, which we then use to convert audio input from the smartphone's microphone to a string output, combined with the rate of confidence in the detected text. The main limitation of this API lies in the fact that it needs an internet connection, which does not seem significant considering that our application is web-based and would likely be disseminated using the internet as well. In terms of the parameters set for the API, which can be observed in listing 4.1, we used the (US) English language, since we wanted to make the application more globally accessible. The maximum number of alternative results is set to 1, which means the system only tries to match the audio input to one text output. Although in the future we may try to match several alternative results with the expected inputs, a

choice for simplicity was taken. The recognition is continuous, indicating that the system is always trying to "listen" to new voice inputs, which allows to continuously try for a match of the user's input. Finally, we do not receive interim results, since we wish to only process the input once the user has finished talking. To simplify, we do not set a specific grammar for the recognition.

```
1    this._recognition = new SpeechRecognition();
2    this._speechRecognitionList = new SpeechGrammarList();
3    this._recognition.grammars = this._speechRecognitionList;
4    this._recognition.continuous = true;
5    this._recognition.lang = 'en-US';
6    this._recognition.interimResults = false;
7    this._recognition.maxAlternatives = 1;
```

Listing 4.1: Setup of the speech recognition API.

Speech recognition is started as soon as a section with speech interaction is initiated. Whenever speech is detected and returned with success by the API (*onresult()*), we analyse the text input, by comparing it with the several possibilities for the expected input, set in the configuration JSON file, detailed on 4.4. Given our goal to have a simplified solution, the comparison is done by checking if, for each possibility, the input string is a subset of the expected string. If a match is found, the speech recognition is ended, as well as the current section. If not, then the speech recognition API continues its behaviour, continuously trying to detect and recognise new speech input. This is also the behaviour when the API is not able to recognise the detected speech or when an error occurs.

This implementation technique has several limitations worth mentioning. For once, when defining the expected inputs for each specific speech interaction, it is hard to cover all possible inputs from the user, considering all the possible synonyms and ways of expressing the same thing. This problem can escalate greatly with the size of the expected inputs. For the later development of the use cases, we mitigated this obstacle, by asking for very short and simple inputs.

Similarly, the solution can suffer from a large number of false positives, since the negative forms of expressions will likely be detected as a successful match. For example, if the expected input is "ready" and the user says "not ready", the system will match the two strings and consider a successful interaction, when, in fact, the user said the opposite of what was intended. Considering both these obstacles, we think that, for a possible next phase of the project, the solution should be heavily improved, with the use of, for example, Natural Language Processing techniques. However, to mitigate them during the user test phase, we made the task instruction clear in terms of the expected input. For example, in the case where the user was expected to say "ready" the instruction given was "let me tell (...) I'm ready to start".

Additionally, the speech interaction recognition can be limited in noisy environments. Although, in the future, noise reduction software can be used to remove this obstacle, during the user tests we avoided it by using a silent environment.

4.3.2 Head Gestures

In the context of this project, head gestures refer to every kind of movement that can be detected by the smartphone's motion sensors (that, in smartphone VR headsets, are attached to the user's head), including all body gestures that impact the position and/or rotation of the user's head. This of course can include a multitude of different movements, such is the case of:

- head only movements, like nodding yes or no and looking up, down, left or right
- upper body movements, like turning your body, dodging from an obstacle, leaning in and looking back or to any specific object in the scene;
- full body movements, like crouching, jumping, laying down, sitting, getting up and even shaking;

Considering this variety, and the challenges that it might bring during the implementation, our decision was to only implement two different types of head gestures, according to the use case needs: nodding (yes and no) and looking (to a specific location of the scene). Regarding what parameters of the interaction are recognised by the solution, we opted for only recognising if the gesture was performed or not by the user. However, it could be interesting in the future to explore how the speed and amplitude of the movement's execution might be used as a natural interaction parameter as well.

Regarding the implementation of the detection and recognition of the two gestures, it was, in general, done with a very simplistic and theoretical/non-empirical approach, as the motion data produced by the movements is interpreted independently and not compared with experimental data. Although an approach using an Artificial Intelligence model might be worth approaching in the future, we believe our simple approach to be more transparent and adequate given the focus of the project. Each of the two movements' recognition was implemented independently, given their distinctiveness.

To detect the gesture of looking at a specific location/object on the scene, we use the *device-orientation* event, which is a part of the Window API. This event is fired whenever the orientation of the device suffers a change, and it uses the motion sensors to gather and then return said current orientation, in its three main axes: alpha, beta and gamma, seen in figure 4.6. Our solution subscribes to the *deviceorientation* event listener as soon as the section with the look interaction begins. With every change of the orientation, it compares the current orientation input with the target set for that specific interaction/section. We set that the user must look at the target for at least 1 second straight for the interaction to be considered successful since we do not want to consider situations where the user's gaze goes through the target temporarily as shifting to another location. We also consider a margin of half a radian (approximately 28.6º) from the user's gaze to the target.

The biggest limitation of this implementation lies in the fact that the target given for each interaction is not absolute, but relative to a baseline orientation correspondent to when the user is looking straight forward. To normalize the final target value, according to the user's initial

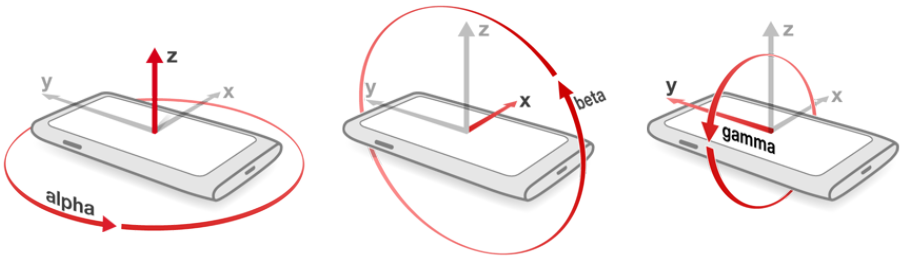

Figure 4.6: Axis considered for deviceorientation. [41]

gaze orientation, we subtract to that expected target location the initial orientation of the device, calculated when the VR mode is initiated. However, this only works when the user is looking straight forward (along the cardinal directions) when the VR mode begins. Otherwise, if the user is looking down or up, for example, the system will not be able to correctly calculate the target's location in relation to the device orientation data and the interaction will not be correctly recognised. And so, it becomes crucial that the user begins the experience looking straight forward (in any given direction).

For the detection and recognition of the nod gestures, we use an event plugin developed by Matthew Wilber, nod.js [70]. This plugin detects movement in four different directions, up, down, left and right, using the *devicemotion* event from the *Window* API. By subscribing to the event listeners of the plugin, we receive *nodEvent*s containing the direction of movement, at a fixed interval, set by us. For a "nod no", we consider that the user must shake their head right and then left or vice-versa, with a maximum interval in between of 2.5 seconds, as seen in listing 4.2. The implementation is similar for the "nod yes", with the exception that we consider the "up" and "down" directions instead of "left" and "right". For this implementation, the most relevant limitation seems to be that the input from *devicemotion* is not always accurate and there is an occurrence of false positives (for example, "right" or "left" being recognised when the user's head is only moving "up" and "down").

```
1    if (event.direction == right OR left){
2        if (last_direction == opposite of event.direction){
3            endRecognition();
4            clearTimeout();
5        }
6        else{
7            last_direction = event.direction;
8            setTimeout(cancelNod(), 2500);
9        }
10   }
```

Listing 4.2: Pseudocode for recognition of nod "no".

4.3.3 Hands Gestures

Similarly to head gestures, hands gestures can include a myriad of different movements: hand gestures that indicate like or dislike, like pointing the thumb up or down or doing a heart with both hands; shaking your hands in disagreement; reaching and/or grabbing an object in the scene; hugging or pushing someone, etc. Some of these movements however can require a complex Computer Vision model to be recognised. Because of this, and considering the focus and the restraints of the project, our solution covers one very simple movement that can be used in different contexts: covering your eyes with your hands/arms. This is a versatile movement, that is naturally done in varied situations, like to protect one's face (from an attack or from a strong light, for example), to hide from embarrassment or sadness when crying or to bring calmness in a stressful scenario.

The implementation of the recognition of this movement seems also to be quite straightforward. When people cover their eyes, and subsequently the smartphone's camera, the brightness of the visible image decreases, so all that is needed is to detect this decrease. To access the camera's input, we use the node.js package react-camera-pro [49] which allows to "take a photo" and get the corresponding camera's image. The implemented system continuously takes new photos, with 20 millisecond intervals, and calculates the associated brightness by averaging the RGB components of every pixel. The interaction is considered to have been successful if the average brightness is inferior to 50 (out of 255).

Although the simplicity of this solution is appealing, it presents some limitations worth mentioning. On one side, it needs a relatively bright environment to work successfully, since the camera will not be able to detect differences in brightness if the environment is too dark. However, this appears to be a limitation common to every solution that relies on vision for the detection of movements. Furthermore, with this solution, it is not possible to differentiate hands or arms from other different types of objects, so false positives can occur. On the other hand, some headsets allow some light in, which might make it much harder to detect a difference in brightness. All these limitations were taken in consideration during the execution of the user tests, in order to mitigate their consequences.

4.4 Input Parameters and Files

As mentioned before, the system takes in inputs from two different methods: the URL parameters and, based on them, a JSON file, containing the ordered sections of the scenario, with all relevant information about each of them. For the URL, 4 different parameters must be included:

- scenario name
- user's gender
- scenario with or without interaction
- participant's ID (needed for the evaluation phase)

An example of an URL can be:

```
http://localhost:3000/?gender=female&scenario=office&interaction=with&
    participantID=0
```

Composing a string with the first three parameters, we can create the name of a JSON file containing the scenario details, needed for its execution. An example of such a file can be found in appendix A. Each JSON is composed of an array of sections. Each section is a JavaScript object composed of at least two keys: the *type*, which can be *video* if the section does not contain interaction or *interaction* otherwise; and the *video_src* which is the path to the section's 360º video. The *video* sections only have those two parameters and one example can be seen in listing 4.3. For the case of sections with interaction, the object also contains a key *interaction* whose value is an object with the interaction type and other relevant parameters depending on the type of interaction.

```
{
    "type": "video",
    "video_src": "office/female/7_.mp4"
}
```

Listing 4.3: Video section example in JSON input file.

Speech interactions, for which an example can be observed in listing 4.4, contain the possible expected inputs from the user. The hands' interaction, seen in listing 4.5, does not need any particular parameters. For the head interaction, we can have two different types of head interaction types, which are described in the JSON file differently. The "look" movement, includes a *value* which defines the looking target, as exemplified in listing 4.6. As for the "nod" movement, the type of nod (yes or no) is already included in the *head_interaction_type* parameter, as shown in listing 4.7. Finally, the key interaction, used just to aid during the user tests, does not need any extra parameters, as seen in listing 4.8.

```
{
    "type": "interaction",
    "interaction":{
        "interaction_type": "speech",
        "expected_inputs": ["hello", "anyone", "hi", "hey"]
    },
    "video_src": "office/2_.mp4"
}
```

Listing 4.4: Speech interaction section example in JSON input file.

```
    {
        "type": "interaction",
        "interaction":{
            "interaction_type": "hands"
        },
        "video_src": "office/12_.mp4"
    }
```

Listing 4.5: Hands interaction section example in JSON input file.

```
    {
        "type": "interaction",
        "interaction":{
            "interaction_type": "head",
            "head_interaction_type": "look",
            "value": {
                "alpha": 45,
                "gamma": 0,
                "beta": 0
            }
        },
        "video_src": "office/4_.mp4"
    }
```

Listing 4.6: Look (head) interaction section example in JSON input file.

```
    {
        "type": "interaction",
        "interaction":{
            "interaction_type": "head",
            "head_interaction_type": "nod_no"
        },
        "video_src": "office/8_.mp4"
    }
```

Listing 4.7: Nod (head) interaction section example in JSON input file.

```
{
    "type": "interaction",
    "interaction":{
        "interaction_type": "key"
    },
    "video_src": "office/0_.mp4"
}
```

Listing 4.8: Key interaction section example in JSON input file.

4.5 Output Data Logs

When a scenario with interaction is completed, two files are generated and saved by the app. The two files are in formats CSV and JSON. The CSV file contains a subset of the information in the JSON file. Although, at first sight, this might seem like a redundant option, both files are kept because the CSV one is easier for results analysis but the information saved in the JSON file is more complete and might be useful in the future for a more in-depth analysis of certain cases.

The CSV file contains one line for each interaction section. Each column is described in detail in table 4.2 and an example can be found in figure 4.7.

Table 4.2: Parameters saved in the output CSV data log.

	Column		Content
Interaction Details	1	type	interaction method
	2	duration	time from when section started to when interaction is successfully recognised, in milliseconds
	3	normalTermination	set to *true* if interaction is successfully recognised or to *false* if, otherwise, the feature to finish the interaction by pressing the key "E" is used
Speech Interaction	4	numberOfTries	total number of recognised speech inputs
	5	finalInput	last user's input detected
	6	finalConfidence	confidence with which the finalInput was recognised
	7	inputMatch	expected input to which the finalInput matched
Head Interaction	8	head_type	type of head movement
Look Head Interaction	9	target	expected target of look movement
	10	finalQuaternion	final detected user's orientation
	11	finalAngle	angle between target and finalQuaternion
Hands Interaction	12	finalValue	last brightness value detected

The JSON file, contains an array of objects, each one representing a section with interaction. Each object contains the same information as described before for the CSV file. Additionally, it also contains a parameter register of type array, whose content varies depending on the type of interaction. For speech interaction, it contains all the recognised inputs with corresponding

type	duration	normalTermination	numberOfTries	finalInput	finalConfidence	inputMatch	head_type	target	finalQuaternion	finalAngle	finalValue
speech	3306	true	1	i'm ready to start	0.9722	ready					
head	1937	true					look	(0.424757	(0.304052	0.346638	
speech	5051	true	1	the date is october 1939	0.9514	october					
hands	1529	true									48.66667
head	11244	true					nod_yes				
hands	11255	false									89

Figure 4.7: Example of an output CSV data log.

confidences. In the case of hands interaction, it contains all measured average brightnesses. For the "look" movement of the head interaction, it contains all measured orientations of the device and corresponding angles with the target. And finally, for the "nod" movements, the register parameter contains all outputs registered from the nod plugin. In this particular case, there is also an extra parameter *registerMotion*, that contains all output read from the *devicemotion* event. An example of the JSON data log can be found in appendix B.

4.6 Summary

During this chapter, we presented how the solution was implemented and explained the reasons behind each implementation decision. We also showed the limitations inherent to the system and introduced some possible future work related to them. We detailed how the system operates, including its input needs and output formats. In the next chapter, we now get to see the system put to use, as well as evaluated in terms of interaction recognition success and impact on the user experience.

Chapter 5

Evaluation

In order to validate the solution implemented in the previous chapter, we designed and implemented a use case, detailed in section 5.1. Afterwards, and trying to give a response to the two research questions presented in chapter 1, we conducted user tests, with 22 participants, which are further described in section 5.2. From these, we collected both qualitative data regarding the user's experience and perception of the system, using user questionnaires, as well as quantitative data regarding the efficiency of the implemented application and, in particular, of the recognition of the different types of interaction. In section 5.3, we present the results of the user tests, comparing experiences with or without interaction, but also the three implemented interaction methods.

5.1 Design and Implementation of a Use Case

To demonstrate the use of the proposed and implement solution, we implemented two different scenarios, whose design considers the three types of interaction that need to be naturally contextualized. Although the scenarios present two very distinct situations or storylines, they have many similarities that allow them to be comparable. Both scenarios are relatively short, having a duration of 2 minutes and 20 seconds and 2 minutes and 29 seconds.

In the two scenarios, the user experiences the scenario from the point of view of a "main character", hearing a voice that represents their thoughts. Said voice is meant to guide the participant into the scene, giving insights into what the main character is feeling, experiencing and doing. In the case of the user experiencing the scenario with interaction, the voice/thoughts also serve as a guide to how and when the participant must perform the interaction tasks. Given the limitations of the system, the participants are able to look all around them, in both scenarios, but not move around the scene.

Both scenarios include exactly 6 moments designed for potential interaction. As so, the scenarios are each composed of 7 main videos (video sections) that compose the narrative presented. When the experience is run with the testing parameter "with interaction", the 7 video sections are intercalated by 6 interaction sections, 2 of each of the three types of interaction (speech, head and hands). The interaction sections contain "filler" videos (1/2 second videos where nothing happens)

that are played in loop until the interaction is completed with success. For the case of the scenario run without interaction, the sections corresponding to hands and head interaction are suppressed, while the ones corresponding to speech interactions are replaced by simple video sections, where one can hear the thought voice actor saying the expected input. This workflow can be observed in figure 5.1.

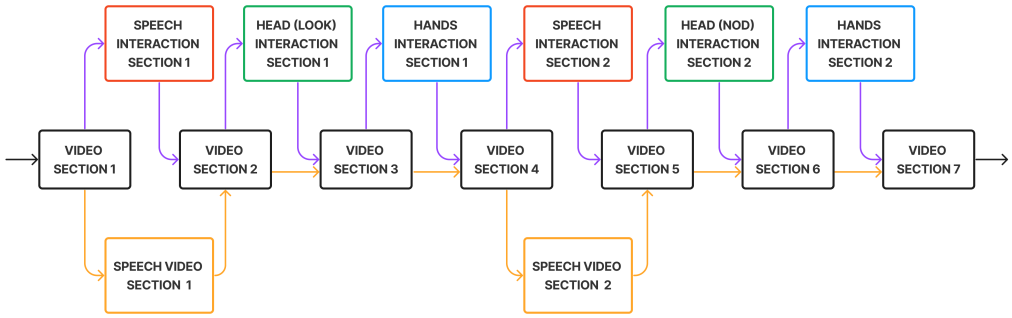

Figure 5.1: Workflow of the user test scenarios. The order of interactions can change according to the scenario. In purple is the workflow for the case with interaction and in orange for the case without interaction.

For the implementation of the scenarios, our decision was to use 360º footage, as opposed to 3D-generated environments/scenes, given our need for the realism of the scenarios. Given our specific requirements for the scenarios, we wrote two scripts, corresponding to the two scenarios. The video footage was filmed in Porto, using a Ricoh Theta V 4K 360º filming camera [1]. For the voice that emulates the thoughts of the main character, we recorded two sets of voice recordings, one for a male voice and another for a female one. We believe that the usage of a voice with which the participant can relate to, in terms of gender, is preferable for the prespective taking aspect of the scenario. All four participating actors, the two that feature in the video footage and the two whose voice was recorded, were volunteers. At the end of the recording process, we edited the material into the necessary video sections used to compose the two scenarios, using the software DaVinci Resolve [2].

In the next two subsections, we describe the two implemented scenarios, named for ease of communication: Presentation and Office.

5.1.1 Presentation Scenario

As the name might imply, in the Presentation scenario, the participant impersonates a student that is doing a presentation to a teacher. When the scenario starts, the student has just finished their presentation and is now preparing to reply to questions and hear comments from the teacher. The scene unfolds with the teacher being disappointed by the student's presentation and asking

[1] https://theta360.com/en/about/theta/v.html
[2] https://www.blackmagicdesign.com/products/davinciresolve

a question. The student is unsure of the answer. They recheck their presentation and answer incorrectly, which causes embarrassment. The teacher reiterates their disapproval but agrees to give the student a new opportunity. The student agrees nervously with a nod. As the teacher opens the door to the room to let the student out, a bright light comes in, causing the student to protect their eyes.

The Presentation scenario has a duration of 2 minutes and 29 seconds and includes 6 moments of possible interaction, described in table 5.1. Figure 5.2 shows three different perspectives of the Presentation scenario, as seen displayed to the user in the application. The full script can be found in appendix C.

Table 5.1: Interactions of the presentation scenario.

Order	Interaction Type	Description of Interaction
1	speech	saying they are ready
2	head (look)	looking at the presentation to see the answer
3	speech	answering the question asked by the teacher
4	hands	hiding their eyes in embarrassment
5	head (nod)	nodding yes to answer the teacher's proposal
6	hands	covering their eyes to protect them from the light

5.1.2 Office Scenario

In the Office scenario, the participant impersonates Sam, a character that works in an office and has decided to come to said office at night, alone, when no one else is supposed to be in the building, to prepare for an important presentation they have the next morning. The scene starts with a brief introduction and some steps being heard. Sam calls "hello" to see if someone else is in the building but no one answers. As the steps' sounds get louder and the lights go off, Sam gets scared and looks at the door to check if it is closed. It is not and a looming figure comes in. Scared to be attacked, Sam protects their face. The figure turns out to be Sam's colleague, Kyle, who asks if Sam is okay. Still shaking, Sam nods no. Kyle asks sorry, explains what happened and bids goodbye. Sam manages to answer out loud. As Kyle leaves, Sam tries to calm themself down by covering their eyes and taking deep breaths.

The office scenario has a duration of 2 minutes and 20 seconds and its 6 moments of possible interaction are described in table 5.2. Figure 5.3 shows three different perspectives of the Office scenario, as seen displayed to the user in the application. The full script can be found in appendix C.

5.2 User Tests

As mentioned before, we executed user tests to evaluate the impact of the interaction on the user's experience, as well as the performance of the software. In the next subsections, we describe the design process of the user tests, as well as the structure of a typical user test session. We also detail

Figure 5.2: Perspectives from the presentation scenario.

Table 5.2: Interactions of the office scenario.

Order	Interaction Type	Description of Interaction
1	speech	calling out “hello”
2	head (look)	looking at the door to check if it is closed
3	hands	protecting their face with their hands/arms
4	head (nod)	nodding no to answer Kyle
5	speech	saying goodbye to Kyle
6	hands	covering their eyes to calm down

how we gather data from the user tests and, finally, we specify how, when and where the user tests were executed.

5.2.1 User Tests Design

When first starting the design of the user tests, we defined and contemplated the variables that are worth considering during the tests: the presence of interaction (with vs without) and the scenarios (Presentation vs Office). We first defined that all participants will experience both scenarios as we

Figure 5.3: Perspectives from the office scenario.

want to establish the independence of our results with a specific scenario. Secondly, we agreed that for each of the scenarios, the study should be between-subjects, as each participant should only experience a specific scenario either with interaction or without it, and never both ways as the first could heavily influence the second viewing of the same scenario.

Pondering the possible order changes of the scenarios and the combinations of each scenario with the variable of interaction, we first obtained a user test schema, with eight different possible layouts of testing, as seen in table 5.3. However, given limitations of the project in terms of time and the minimal desirable number of participants for each layout, we regarded as the best option to opt for the 4 layouts highlighted in the table, where each user experiences one scenario with interaction and the other one without. We believe this alternative provides enough data to study all biases related to the order and combinations of the variable of presence of interaction, while simplifying the user test process.

Table 5.3: Layout of the user tests. The eight layouts constitute the ideal schema, while the highlighted layouts represent the schema used. "P" and "O" stand for Presentation and Office scenario, respectively. "w/" and "w/o" stand for with and withot interaction, respectively.

Layouts	1	2	3	4	5	6	7	8
Part 1	P - w/	P - w/o	P - w/	P - w/o	O - w/	O - w/o	O - w/	O - w/o
Part 2	O - w/	O - w/	O - w/o	O - w/o	P - w/	P - w/	P - w/o	P - w/o

5.2.2 User Test Session Structure

To better plan the user tests, we defined a simple first structure. Each test is planned for a duration of 30 minutes.

The first 5 minutes are reserved for the **set up**: explaining the project and test structure to the participant, as well as the signing of the informed consent document, which can be seen in detail in appendix D, and the filling of the first part of the questionnaire, dedicated to demographic and VR experience related questions.

Then follows two similar 10-minute sections, each of them corresponding to the **two scenario "parts"** of the experience, according to the layout presented in the previous subsection. Each 10-minute part is divided into: 3 minutes to give the context surrounding the scenario and explain, if that is the case, how the interaction works; 3 minutes to experience the scenario itself; and 4 minutes to fill the corresponding questionnaire section, with questions regarding that specific part. Although the part that is done with interaction is longer, we consider an average of 10 minutes for each part.

The final 5 minutes of the user test are dedicated to any **comments, observations or questions** the participant might have.

A longer script used for the execution of the user tests can be found in appendix E.

5.2.3 Data Gathering

To later evaluate the implemented solution, it is required to gather some data during the execution of the user tests. We used four different methods for that purpose:

1. one questionnaire to collect data related to the user and their personal experience, further exlained in this section;
2. data logs generated automatically by the application for the later evaluation of the performance of the software developed, previously clarified in section 4.5;
3. video recordings of the participant experiencing both scenarios;
4. hand-taken notes of anything unusual that happens during the test.

The first two were the main methods used for results withdrawal, while the last two were used mainly for the verification of the data collected by the data logs. The user was informed of all these data collection methods before giving their written formal consent to participate in the user test.

The questionnaire was built using Google Forms and is composed of 5 main sections which are filled throughout the user test. The participants received instructions from us to fill the sections at the right moments, with the exception of the first section which is filled directly by us. The full questionnaire can be seen in appendix F.

The first section of the questionnaire is not filled by the participant but by us and it serves only to register the participant's ID (for the future match with data logs) and the order in which the participant is gonna do both scenarios (Presentation and Office).

The second section of the questionnaire is dedicated to two demographic questions (age and gender), two VR experience-related questions (one for smartphone VR specific and another for other types of VR) and one control question regarding nausea (that serves to compare with similar questions after use of the system)

The third and fourth sections of the questionnaire are equal and are about the first and second parts of the experience. The first question in these sections is whether the experience was done with or without interaction. This question allows us to later redirect the participant to a subsection with interaction questions in the first case or advance to the next part otherwise. After, the user is required to rate their agreement with a few statements regarding the experience, using a 5-point Likert scale:

- The scenario I experienced felt realistic.
- I was aware of my real surroundings during the experience.
- I felt immersed in the environment.
- I felt like I was in the portrayed place.
- I felt like I was living the portrayed situation.
- I felt like I was in the body of the Main Character.
- I felt empathy towards the Main Character.
- I enjoyed the experience.

The section continues with another control question regarding nausea made to assess the effect of the experience. The subsection of the interaction-related questions starts by asking the user to order the types of interaction according to their preference. Then, the user is asked to rate their agreement with a few statements regarding the three types of interaction, using a 5-point Likert scale:

- It felt intuitive.
- It was easy to perform.
- It felt natural given the scenario.

- It was clear to me what I was expected to do.
- It contributed to me feeling immersed in the environment.
- It contributed to me feeling like I was in the portrayed place.
- It contributed to me feeling like I was living the portrayed situation.
- It contributed to me feeling like I was in the body of the Main Character.
- It contributed to me feeling empathy towards the Main Character.
- It contributed to me enjoying the experience.

Finally, the fifth and last section of the questionnaire is an open question and allows the participant to give some comments and suggestions about the experience/project.

5.2.4 Execution of the User Tests

The execution of user tests happened in the span of two weeks and two days. The location varied but all tests happened in a silent environment so that the participant would not be disturbed during the execution and the speech recognition software would not have noise interference. Moreover, the rooms used had natural daylight and/or artificial light so that the camera could detect a difference in the average brightness during the hand gestures detection.

In terms of hardware used, the smartphone used was a Xiaomi 9T (EAN: 6941059624936). The headset was the VR Box 2.0. The app was run locally on the computer (Lenovo Legion 5 15ARH05H-198 with CPU AMD Ryzen 7 4800H with Radeon Graphics 2.90 GHz, 16GB of RAM, GPU NVIDIA GeForce RTX™ 2060, running on Windows 10 Pro) with Google Chrome, which then forwarded it to the smartphone via USB debugging.

In figure 5.4, it is possible to observe a user during the execution of their user test.

Figure 5.4: User test example.

The results obtained from the execution of this user tests are presented in the next section of this chapter.

5.3 Results

In this section, we put forward the results of the user tests. Firstly, we display the demographic data about the participants. After, a comparison of the results for the cases with and without interaction is made, focusing on the user experience. Then, we go in-depth into the interaction methods, comparing them in terms of user experience and success in their recognition by the implemented solution. We finish with an exposition of some other important results and observations.

5.3.1 Demographics

In total, 22 people participated, 16 identified as males and 6 as females, as seen in table 5.4.The age distribution is presented in the table 5.5. Nine out of the 22 participants (≈40%) had experienced Smartphone VR previously, with only five (≈23%) having experienced it 3 to 5 times, as seen in table 5.6. Regarding other types of VR (not including Smartphone VR), the level of experience of the participants was more dispersed, with only 6 of them (≈27%) having never experienced it previous to this user test.

Table 5.4: Gender distribution of participants.

Gender	Participants	%
Male	16	72,73%
Female	6	27,27%
Non binary	0	0,00%

Table 5.5: Age distribution of participants.

Age Interval	Participants	%
19 or less	2	9,09%
20-24	10	45,45%
25-29	4	18,18%
30-34	3	13,64%
35-39	2	9,09%
40-44	0	0,00%
45-49	1	4,55%
50-54	0	0,00%
55-59	0	0,00%
60 or more	0	0,00%

Table 5.6: Smartphone VR and other VR experience distribution of participants.

Number of Previous Contacts	Smartphone VR		Other VR	
	Participants	Percentage	Participants	Percentage
0	13	59,09%	6	27,27%
1-2	4	18,18%	5	22,73%
3-5	5	22,73%	4	18,18%
5-10	0	0,00%	3	13,64%
10+	0	0,00%	4	18,18%

5.3.2 With vs Without Interaction

For each of the statements that the participants were asked to rate their agreement with, we can compare the results from the cases with and without interaction. In figure 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, 5.11 and 5.12 we observe the distributions of the answers and, in tables 5.7, 5.8, 5.9, 5.10, 5.11, 5.12, 5.13 and 5.14, we analyse the respective median, quartiles and average values. In general, we can note better results for the cases with interaction, compared to the cases without interaction. The exception is in terms of immersion, where there does not seem to be a difference. The difference also appears to be more relevant in terms of the enjoyment of the participant, the empathy towards the main character, as well as its embodiment.

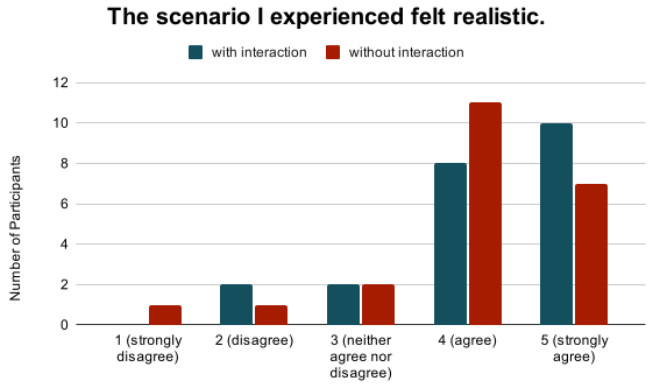

Figure 5.5: Distribution of answers for "The scenario I experienced felt realistic".

Table 5.7: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "The scenario I experienced felt realistic".

	With	Without
Q1	4	4
Mdn	4	4
Q3	5	5
M	4,18	4,00

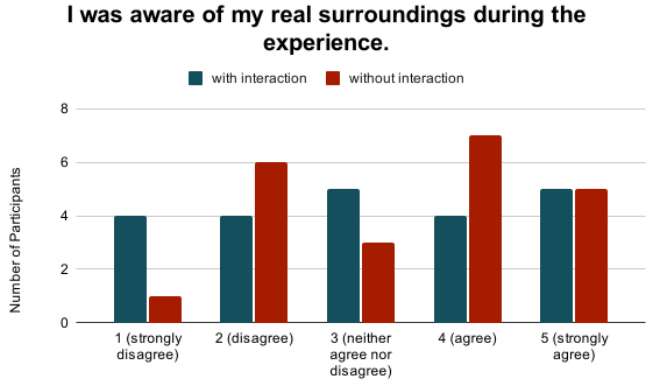

Figure 5.6: Distribution of answers for "I was aware of my real surroundings during the experience".

Table 5.8: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "I was aware of my real surroundings during the experience".

	With	Without
Q1	2	2
Mdn	3	4
Q3	4	4
M	3,09	3,41

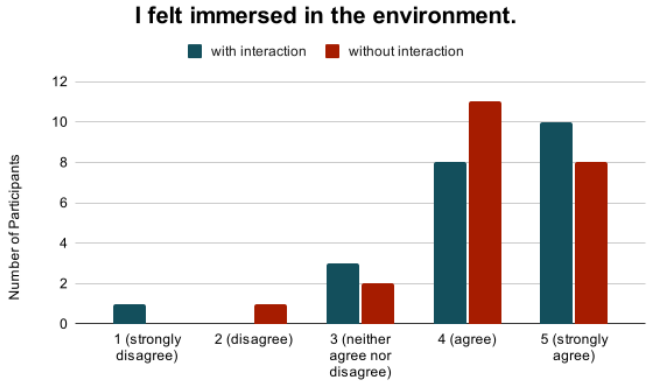

Figure 5.7: Distribution of answers for "I felt immersed in the environment".

Table 5.9: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "I felt immersed in the environment".

	With	Without
Q1	4	4
Mdn	4	4
Q3	5	5
M	4,18	4,18

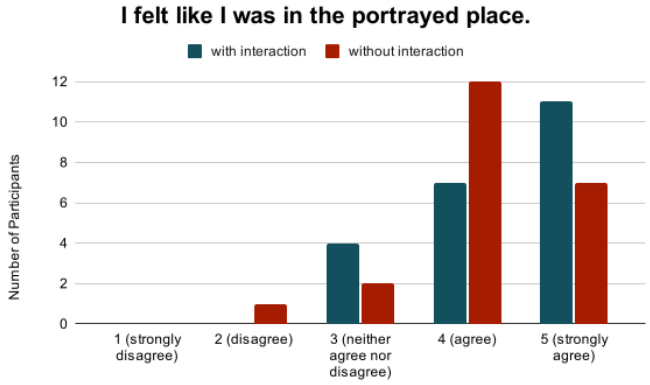

Figure 5.8: Distribution of answers for "I felt like I was in the portrayed place".

Table 5.10: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "I felt like I was in the portrayed place".

	With	Without
Q1	4	4
Mdn	4,5	4
Q3	5	5
M	4,32	4,14

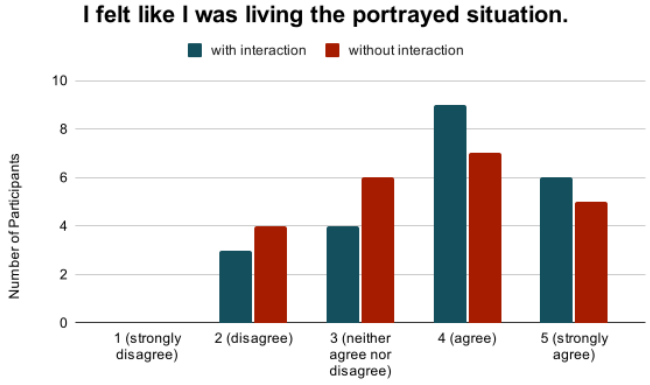

Figure 5.9: Distribution of answers for "I felt like I was living the portrayed situation".

Table 5.11: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "I felt like I was living the portrayed situation".

	With	Without
Q1	3	3
Mdn	4	4
Q3	4,75	4
M	3,82	3,59

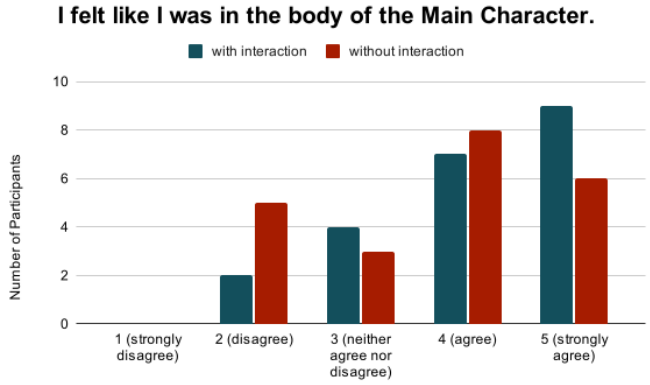

Figure 5.10: Distribution of answers for "I felt like I was in the body of the Main Character".

Table 5.12: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "I felt like I was in the body of the Main Character".

	With	Without
Q1	3,25	3
Mdn	4	4
Q3	5	4,75
M	4,05	3,68

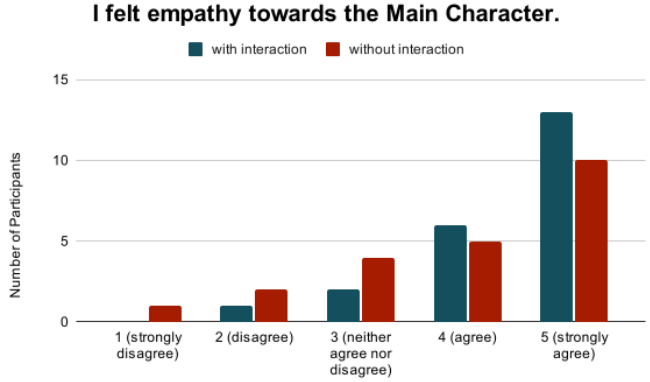

Figure 5.11: Distribution of answers for "I felt empathy towards the Main Character".

Table 5.13: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "I felt empathy towards the Main Character".

	With	Without
Q1	4	3
Mdn	5	4
Q3	5	5
M	4,41	3,95

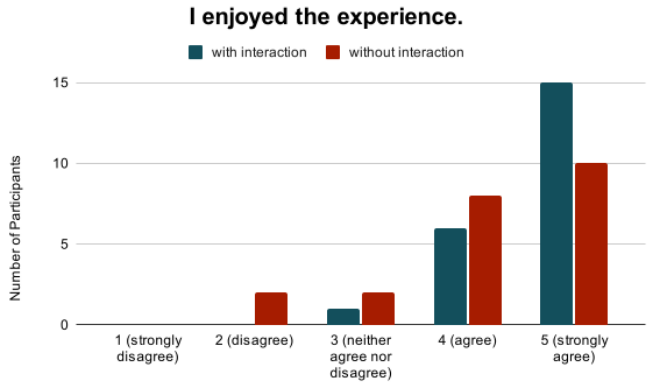

Figure 5.12: Distribution of answers for "I enjoyed the experience".

Table 5.14: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "I enjoyed the experience".

	With	Without
Q1	4	4
Mdn	5	4
Q3	5	5
M	4,64	4,18

We ran a Mann-Whitney test to try and prove a significant difference between the experience with and without interaction, for all of the questions/aspects studied, as seen in table 5.15, considering a critical value of 127. However, for no question did we obtain a significant value. It

is worth noting though that for both cases, with and without interaction, the results were positive for all answers (only medians of 4 or 5 were registered to the exception of the awareness of real surroundings). In fact, for the case "without interaction", which we can consider the baseline for comparison, the median was of 4 for all questions. Given that there was only one more point on the likert scale above that 4, this means that the possible improvement added by the interaction was of only one point in the likert scale, which heavily constraints the use of the Mann-Whitney test.

Table 5.15: Mann-Whitney accessing difference between with vs without interaction.

	U	Significance
The scenario I experienced felt realistic.	214	no
I was aware of my real surroundings during the experience.	212	no
I felt immersed in the environment.	230	no
I felt like I was in the portrayed place.	209,5	no
I felt like I was living the portrayed situation.	211,5	no
I felt like I was in the body of the Main Character.	198	no
I felt empathy towards the Main Character.	194	no
I enjoyed the experience.	178	no

Given what was presented, we believe we can say that the results for the case with interaction were overall better but that further tests should be conducted and that a larger likert scale should be used in order to capture the particularities of the user experience.

5.3.3 Interaction Methods

When asked directly to rank the types of interaction, participants showed, in general, a clear preference for the speech interaction. Between the hands and head interaction, the results are less distinctive, but there seems to be a slight preference for hands interaction, as seen in figure 5.13.

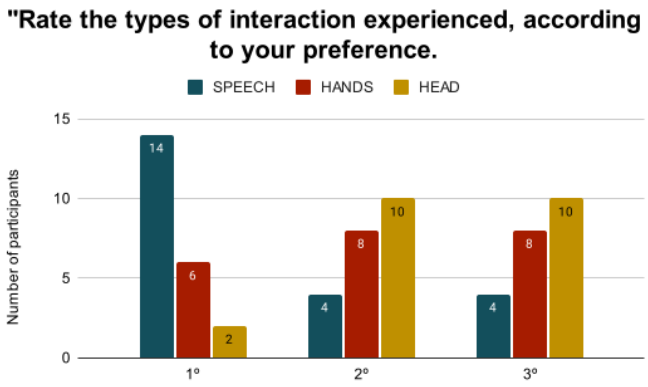

Figure 5.13: Participant's rating distribution of the 3 interaction methods.

We asked the participants to rate their agreement with a series of statements for each of the three interaction methods. We can now analyse the distribution of the answers in figures 5.14, 5.15, 5.16, 5.17, 5.18, 5.19, 5.20, 5.21, 5.22 and 5.23 and respective median, quartiles and average values on tables 5.16, 5.17, 5.18, 5.19, 5.20, 5.21, 5.22, 5.23, 5.24 and 5.25. Going question by question, we can also see clearly that the speech interaction performed better in all aspects, from intuitiveness, naturalness, ease and clarity of instruction to contribution to the senses of immersion, presence, empathy and enjoyment. As for the head and hands interaction, the results are still positive, although with not so much distinction. An exception is the case of the hands interaction, that participants have considered less natural given the scenario, as demonstrated in figure 5.16 and table 5.18. This comes to show that the interaction might not have been well contextualized in the narrative, which might explain the overall worse results for this method. It is also worth noting that, for hands and head interaction, the results are particularly positive for easiness and enjoyment.

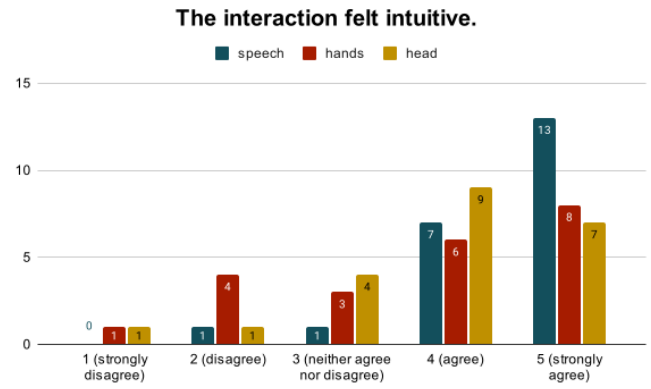

Figure 5.14: Distribution of answers for "The interaction felt intuitive" for speech, hands and head interaction.

Table 5.16: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "The interaction felt intuitive".

	Speech	Hands	Head
Q1	4	3	3,25
Mdn	5	4	4
Q3	5	5	5
M	4,45	3,73	3,91

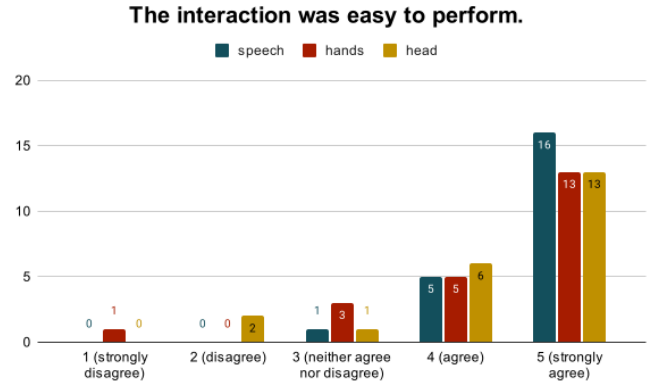

Figure 5.15: Distribution of answers for "The interaction was easy to perform" for speech, hands and head interaction.

Table 5.17: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "The interaction was easy to perform".

	Speech	Hands	Head
Q1	4,25	4	4
Mdn	5	5	5
Q3	5	5	5
M	4,68	4,32	4,36

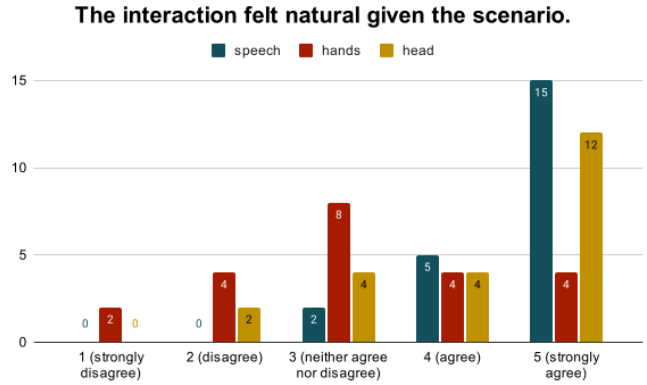

Figure 5.16: Distribution of answers for "The interaction felt natural given the scenario" for speech, hands and head interaction.

Table 5.18: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "The interaction felt natural given the scenario".

	Speech	Hands	Head
Q1	4	2,25	3,25
Mdn	5	3	5
Q3	5	4	5
M	4,59	3,18	4,18

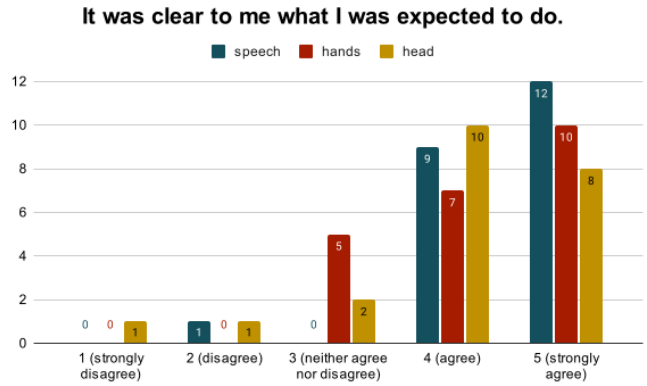

Figure 5.17: Distribution of answers for "It was clear to me what I was expected to do" for speech, hands and head interaction.

Table 5.19: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "It was clear to me what I was expected to do".

	Speech	Hands	Head
Q1	4	4	4
Mdn	5	4	4
Q3	5	5	5
M	4,45	4,23	4,05

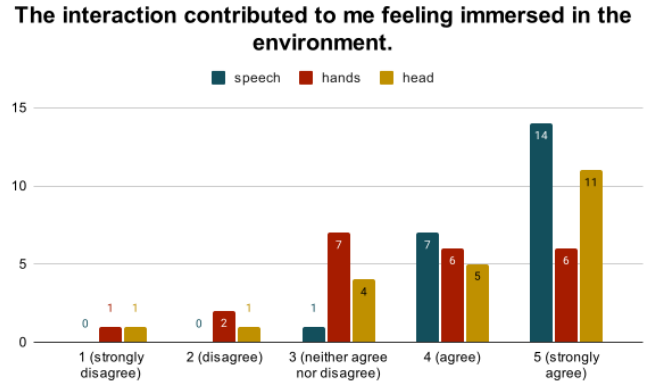

Figure 5.18: Distribution of answers for "The interaction contributed to me feeling immersed in the environment" for speech, hands and head interaction.

Table 5.20: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "The interaction contributed to me feeling immersed in the environment".

	Speech	Hands	Head
Q1	4	3	3,25
Mdn	5	4	4,5
Q3	5	4,75	5
M	4,59	3,64	4,09

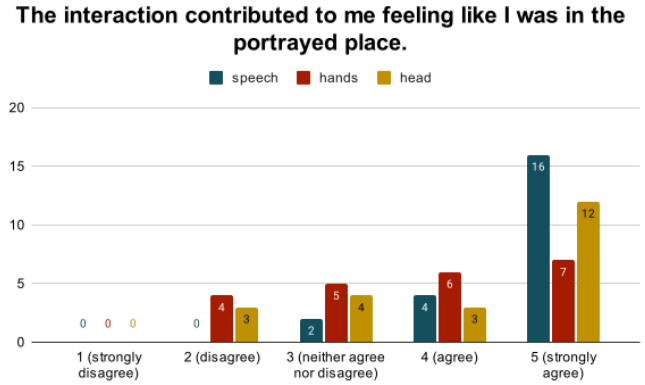

Figure 5.19: Distribution of answers for "The interaction contributed to me feeling like I was in the portrayed place" for speech, hands and head interaction.

Table 5.21: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "The interaction contributed to me feeling like I was in the portrayed place".

	Speech	Hands	Head
Q1	4,25	3	3
Mdn	5	4	5
Q3	5	5	5
M	4,64	3,73	4,09

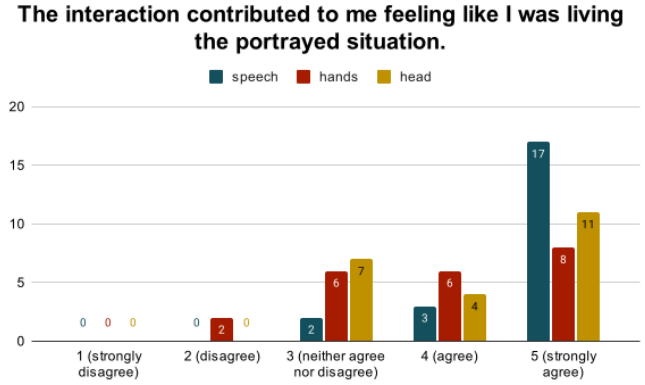

Figure 5.20: Distribution of answers for "The interaction contributed to me feeling like I was living the portrayed situation" for speech, hands and head interaction.

Table 5.22: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "The interaction contributed to me feeling like I was living the portrayed situation".

	Speech	Hands	Head
Q1	5	3	3
Mdn	5	4	4,5
Q3	5	5	45
M	4,68	3,91	4,18

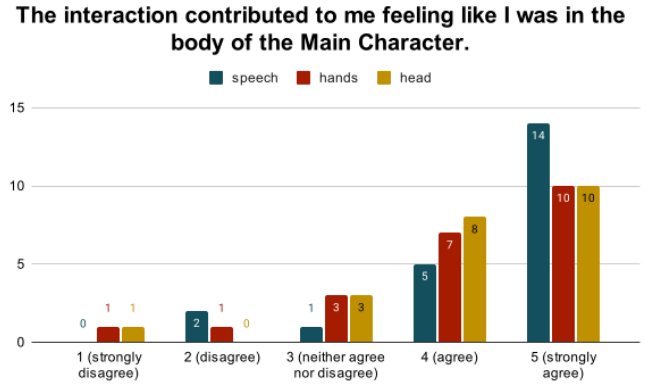

Figure 5.21: Distribution of answers for "The interaction contributed to me feeling like I was in the body of the Main Character" for speech, hands and head interaction.

Table 5.23: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "The interaction contributed to me feeling like I was in the body of the Main Character".

	Speech	Hands	Head
Q1	4	4	4
Mdn	5	4	4
Q3	5	5	5
M	4,41	4,09	4,18

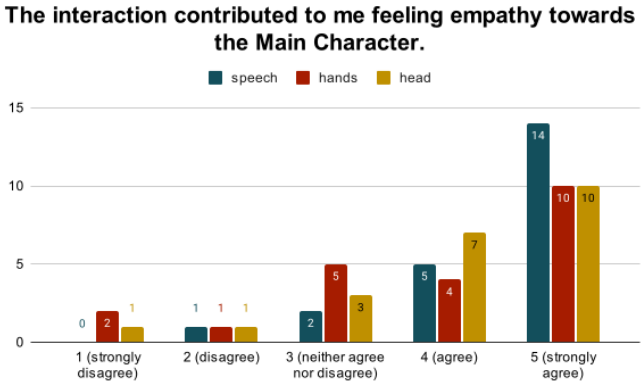

Figure 5.22: Distribution of Answers for "The interaction contributed to me feeling empathy towards the Main Character" for speech, hands and head interaction.

Table 5.24: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "The interaction contributed to me feeling empathy towards the Main Character".

	Speech	Hands	Head
Q1	4	3	4
Mdn	4	4	4
Q3	5	5	5
M	4,45	3,86	4,09

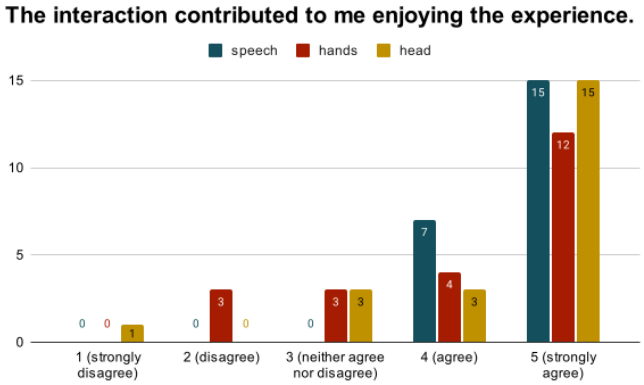

Figure 5.23: Distribution of answers for "The interaction contributed to me enjoying the experience" for speech, hands and head interaction.

Table 5.25: 1st quartile (Q1), median (Mdn), 3rd quartile (Q3) and mean (M) statistics for statement "The interaction contributed to me enjoying the experience".

	Speech	Hands	Head
Q1	4	3,25	4
Mdn	5	5	5
Q3	5	5	5
M	4,68	4,14	4,41

Aggregating the data from the data logs with the manually written notes taken during the user tests, tables 5.26 and 5.27 present the results of each participant for each of the interactions/tasks they had to perform in scenarios Presentation and Office, correspondingly. We consider a success (**S**), when the participant did the correct interaction without any intervention from us and the system correctly identifies that interaction. It also seemed relevant to consider an almost-success for the cases where the system correctly identified the interaction but we needed to have a minor voice intervention to help the user do the interaction correctly. This includes the times where we had to repeat or translate a part of the scenario (**RT**), where the participant first performed a different type of interaction and had to be warned (**WI**) or where we had to ask the participant to do a stronger nod (**SN**) or correct the nod's direction (**CN**). Finally, we consider failures all the times the system failed to correctly identify the interaction. This includes false positives (**FP**) and false negatives (**FN**), as well as trouble identifying hand interaction, either because the camera initialization failed (**CF**) or because the brightness did not lower enough with the interaction (**HF**).

Table 5.26: Results of the user tests interactions for the presentation scenario. Each row represents a participant and each column an interaction. On the title/first row, Ix means the xth interaction in the scenario. S stands for Speech interaction, H for Hands interaction, L for Look (head) interaction and N for Nod (head) interaction.

Participant	I1 S	I2 L	I3 S	I4 H	I5 N	I6 H
1	S	S	S	RT	FP	S
2	S	S	S	S	S	HF
5	FN	FN	S	HF	SN	S
6	FN	S	S	S	RT	S
9	S	FN	S	HF	S	HF
10	S	S	FN	HF	SN	S
13	RT	S	S	S	S	S
14	S	WI	S	S	FP	S
17	S	FN	S	HF	S	S
18	S	RT	S	S	RT	HF
21	FN	FN	S	HF	RT	HF

Table 5.27: Results of the user tests interactions for the office scenario. Each row represents a participant and each column an interaction. On the title/first row, Ix means the xth interaction in the scenario. S stands for Speech interaction, H for Hands interaction, L for Look (head) interaction and N for Nod (head) interaction.

Participant	I1 S	I2 L	I3 H	I4 N	I5 S	I6 H
3	S	S	S	S	S	HF
4	S	S	HF	S	S	HF
7	S	S	HF	CN	S	HF
8	S	S	CF	S	S	CF
11	FN	S	HF	S	S	HF
12	S	S	S	S	WI	HF
15	S	S	HF	S	S	HF
16	S	S	S	S	S	S
19	RT	S	RT	S	S	S
20	FN	S	S	S	FN	S
22	S	S	HF	CN	S	HF

Table 5.28 presents the success and almost-success rates of the system. In total, the system had a full success in 62.12% of the interactions and an almost-success 72.73%. Going into each type of interaction, we can see a clear problem in the detection of the hands interaction, where the system was only almost-successful 47.73% of the times. The almost-success rates of the speech and head interaction are, correspondingly, of 84.09% and 86.36%. Within the head interaction is worth differentiating between "nod" and "look" interactions, since their implementation is separate. The "look" interaction got an almost-success rate of 81.82%. On the other hand, the "nod" interaction had an almost-success rate of 90.91% but a success rate of only 59.09%. This difference corresponds to situations where participants were asked for stronger nods or to correct the

nod direction, which indicates the system might not be sensible enough.

Table 5.28: Distribution of successful, almost-succesful and failed recongition of interaction, in user test.

	Success	**Almost-Success**	**Failure**	**Total**	**Success Rate**	**Almost-Success Rate**
speech	34	3	7	44	77,27%	84,09%
hands	19	2	23	44	43,18%	47,73%
head	29	9	6	44	65,91%	86,36%
Total	82	14	36	132	62,12%	72,73%
look	16	2	4	22	72,73%	81,82%
nod	13	7	2	22	59,09%	90,91%

As seen in table 5.29, we applied the Pearson's correlation coefficient to the data obtained from the successful interactions (S), using a critical value of 0,195. It showed a correlation (r=-0,3989) between the "time until completion of interaction" and the type of interaction, which was expected since some types of interaction are faster to execute than others. It was also observed a correlation (r=0,3636) between "time until completion of interaction" and the scenario experienced (presentation or office). This might be related with the observation made by a large amount of participants that the sound quality in the presentation scenario was worse, making it harder to understand what was being said. Furthermore, a correlation was identified (r=-0,2151) between the "time until completion of interaction" and the number of the interaction (1st to 6th), where time decreases as the number of interaction increases. This can be justified by the fact that the participants get more familiar with the interactions as the scenario goes. We were expecting a correlation between the time and the order in which the participants experienced the scenarios (first with or without interaction) since we expected that experiencing the interaction in the second part would make the participants more familiarized with the scenario format and take less time to complete the tasks. However, such correlation was not found. No further correlations were found to be significant.

Table 5.29: Pearson's correlation of the time until completion of interaction tasks with different parameters.

	Pearson's Correlation
Interaction 1st vs 2nd	0,0502
Scenario	0,3636
Gender	-0,1003
Age	0,0789
Smartphone VR Experience	0,1301
Other VR Experience	0,0705
Interaction Number (1 to 6)	-0,2151
Type of Interaction	-0,3289

5.3.4 Other Results

During the user test, the state of nausea of the participants was also registered, with three checkpoints: before the start of the experience, after the first scenario and after the second scenario. No relevant connection was found between the existence of the interaction and participant's nausea.

The participants were also asked, by the end of the user test, to give any comments or suggestions they might feel relevant. One of the most pointed aspects was the quality of the sound, which was particularly lacking in the Presentation scenario. For the future, we suggest recording the sound of the videos using a separate microphone and not the one included in the 360º video camera. Some participants also brought up that, during the hands interaction, it felt unnatural not seeing their own hands in the virtual environment. This of course shows a limitation of this type of interaction in a smartphone VR context since a real-time full tracking of the hands might be too complex and computational consuming for a smartphone application. However, it is definitely something worth exploring in the future. When running the application in the smartphone, there was a small waiting period in the switch between sections. This was indicated by the participants as something that should be improved upon since it caused a break in the sense of presence. Furthermore, some participants showed their interest in the possibility of walking around the scene. Last but not least, some participants complained about the headset being slightly too heavy and not well adaptable to the head.

5.4 Summary

Analysing the results in a full picture, we see overall good results, which indicate that the use of interaction has a positive impact on the user experience. In the next chapter, we withdraw relevant conclusions from these results, as well as from the whole research. We also come to a full circle, as we connect these results and conclusions with the research questions that motivated the study.

Chapter 6

Conclusions

Although Smartphone VR presents itself as a more accessible alternative to other more complex types of VR, it still has great limitations in terms of the user's interaction with the virtual environment. This means that there is an under-explored potential in this technology that could potentially lead to a better user experience, and, more specifically, a higher sense of immersion, presence, empathy and enjoyment. Observing this gap in the literature and industry, we proposed to explore different types of natural interaction methods that can be applied to smartphone VR and assess if these have a positive impact on the user's experience.

Through chapter 2, we first investigated the concepts of immersion, presence and perspective-taking as a way to enhance user experience, but also to educate, foment self-improvement, change behaviours and biases, inform and create emotional responses. We then explored how we can achieve these concepts in practical terms with interaction methods and it has been shown that natural interaction outperforms unnatural interaction. Afterwards, we focused on how to apply natural interaction in smartphone VR considering the hardware's limitations. Regarding the smartphone's inputs as means for interaction, we mostly observed research focused on the use of cameras, motion sensors and microphones. Yet, not all of said solutions were implemented in smartphone VR, and some of them used unnatural interaction. Furthermore, the studies found focus mostly on the performance of the system on a technological level, and not so much on the user experience level. We also did not find in the literature studies compiling or comparing different types of natural interaction in the context of smartphone VR, which presented as an interesting gap to explore.

Considering the opportunities found in the literature review, we then proposed a solution in Chapter 3. Analysing all possible input devices and how they can be used for natural interaction, as found in previous research, we made a choice regarding which methods of interaction to explore for this project. We excluded several types of input hardware due to their great limitations and made strategic decisions based on our constraints, coming to the final decision of implementing three different methods: speech interaction with the microphone, head gestures interaction with the motion sensors and hand gestures with the back camera.

Proceeding with the development of the proposed solution, we developed an application that allows playing a 360º scenario, composed of several videos, interacting with the scene with the

three defined natural interaction methods. Although the final implemented system presents some limitations, it represents a simple prototype that can be further improved upon, as shown in the next sections of this chapter.

We finally proceeded to evaluate the solution, keeping in mind the goals set at the beginning of the project. Through a use case and the execution of user tests, we extracted and analysed results that help us to answer the research questions.

6.1 Goal Achievement and Research Answers

In Chapter 1, we have set ourselves to give answers to two main research questions that motivated the development of this project, as well as to fulfil two main goals. Now, we return to those goals and questions and analyse our achievements.

For our first goal, we wanted to "explore different natural interaction methods, taking into consideration the smartphone input limitations." On a theoretical level, we explored several natural interaction methods during the literature review, having later analysed them and their feasibility in the smartphone VR context. We later explored three methods on a practical and deeper level having implemented them and evaluated their performance.

With this goal accomplished and analysing our results, we are now able to give an answer to our first research question: "What methods of natural interaction can be used in smartphone VR?". Our research shows that, in theory, we can use head gestures (with the motion sensors), hands and/or arm gestures (with the back camera), facial expressions (with the back camera and very simple extra hardware), eye movement (with the front camera), as well as speech and breathing or other user-produced sounds (with the microphone). For a practical answer, given our results during the evaluation phase of our research, we consider that our almost-success rates of 84,09% and 86,36% for our simplistic approach for the speech and head interaction, respectively, show the feasibility of the use of these methods in smartphone VR. As for the hand gestures, given our low result of only 47,73% of almost-success, we were not able to prove the possibility of its successful use. Having assessed the possibility of the interaction's use on a technical level, it is also worth mentioning that the interactions were, in general, considered intuitive, natural, easy to perform and clear by the users, which shows their usability from the user's perspective.

Our second goal stated the will to "evaluate and compare the effects of the implemented natural interaction methods on the user's experience". Correlated with this goal was our second research question: "Does natural interaction in smartphone VR contribute to the user's experience, in terms of immersion, presence, empathy and enjoyment?" With the results obtained in the evaluation phase of our research, we fulfilled our goal and now try to provide the answers. In general, our results seem to indicate that yes, the natural interaction contributed to the user's experience in all four parameters. Although we were not able, by comparison, to fully prove a correlation between the use of natural interaction and the four parameters, we believe our results from the direct questions ("The interaction contributed to me...") are proof of the positive impact of the interaction in the user's experience. Comparing the three different types of interaction, we conclude that speech

interaction has a stronger impact, while hands interaction, although still positive, has a smaller impact. Contradictory results, however, show that the users tend to prefer hands interaction to head interaction, with speech interaction being more distinctively the top preference. We note nevertheless that these results, especially for the hand and head interaction, might be correlated with the specific gestures performed. Two gestures were implemented for the head interaction and only one for the hands interaction. This allied to the their particular use in the scenarios might be the cause of less positive results. As so, we state the necessity to further evaluate other types of gestures, in order to obtain a conclusive judgement.

6.2 Future Work

When first considering what further work can be done in terms of natural interaction in smartphone virtual reality experience, it is important to discuss our first main decision that impacted the research: the choice of which methods were to be explored. And so, we highlight that although only speech, hands gestures and head gestures were studied on a deeper level during this project, the other methods remain to be given the full attention: facial expressions, eye movement and breathing or other user-produced sounds.

In terms of implementation, as mentioned before, we opted for a simplified version of the interaction recognisers, due to the project constraints. For this reason, it would be relevant to further improve and develop all three methods.

For speech recognition, other voice parameters can possibly be used as natural interaction, like tone and/or volume. The system can also be further developed to better cover other languages and become more accessible. The input matching can also be improved, using, for example, Natural Language Processing techniques, to better solve cases of synonyms and false positives.

In regards to head gestures interaction, there is the potential to implement the recognition of other types of gestures that can be incorporated in different scenarios. It could also be interesting to use the amplitude or speed of movements for example as a natural interaction parameter in itself. Our solution for head gestures recognition also suffers from the limitations in reading the motion sensors input, which could be further explored and improved upon.

Similarly to head gestures, hand gestures could also be improved by implementing the recognition of other types of gestures (using, for example, computer vision techniques) as well as the recognition of amplitude and speed of movement. More important than that, the hand gestures recognition failed more than half the times during our user tests, which might indicate that an altogether different solution might be needed for a successful approach.

Furthermore, it was pointed out by some participants that not seeing their hands during the hands interaction felt unnatural. In a future version, a virtual (and real-time) representation of the hands should be shown to the user. Additionally regarding the implementation, it was noted that between different scenario sections, there was, sometimes, too big of a waiting period, that caused the user to disengage with the scenario. This problem should be further solved, possibly by improving the video loading process.

Regarding the use case, some further work should also be considered. For one, other varied scenarios, including different gestures and amounts of interaction, should be designed and implemented, in order to fully study their impact on the user experience. In specific to the videos, it was noted by the user tests' participants that the sound quality was not good and that that had a negative impact on the experience. For the further recording of 360º videos to be used in this context, we suggest the use of external recording microphones and later syncing with the video image.

Regarding the user tests execution, we believe that further tests should be executed. We think that studying the effects of each interaction independently could be beneficial and bring more insights into their impact on the user experience. As so, we could compare the different methods in a more impartial format but also understand if the amount of different types of interaction also has an impact on the user experience. For example, if using all three methods is more positive (and how much) than just using one or two methods. In addition, tests should be executed with longer scenarios, in order to fully study if the interaction has an impact on the feelings of nausea and/or fatigue of the user.

As our research comes to an end, we believe to have shown that there is still much unexplored potential in smartphone VR. Namely, the application of natural interaction can enhance the user's experience, without the need for extra hardware or complex headsets. With further developments, we trust that smartphone Virtual Reality can be elevated and bring interactive experiences to a much wider set of users.

References

[1] Sun Joo (Grace) Ahn, Amanda Minh Tran Le, and Jeremy Bailenson. The Effect of Embodied Experiences on Self-Other Merging, Attitude, and Helping Behavior. *Media Psychology*, 16(1):7–38, January 2013. Publisher: Routledge _eprint: https://doi.org/10.1080/15213269.2012.755877.

[2] Karan Ahuja, Chris Harrison, Mayank Goel, and Robert Xiao. MeCap: Whole-Body Digitization for Low-Cost VR/AR Headsets. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, pages 453–462, New York, NY, USA, October 2019. Association for Computing Machinery.

[3] Tanja Aitamurto, Shuo Zhou, Sukolsak Sakshuwong, Jorge Saldivar, Yasamin Sadeghi, and Amy Tran. Sense of Presence, Attitude Change, Perspective-Taking and Usability in First-Person Split-Sphere 360° Video. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–12, New York, NY, USA, April 2018. Association for Computing Machinery.

[4] Android. Motion sensors. https://developer.android.com/guide/topics/sensors/sensors_motion. Accessed: 2022-12-19.

[5] Christoph Anthes, Rubén Jesús García-Hernández, Markus Wiedemann, and Dieter Kranzlmüller. State of the art of virtual reality technology. In *2016 IEEE Aerospace Conference*, pages 1–19, March 2016.

[6] Jakki O. Bailey, Jeremy N. Bailenson, and Daniel Casasanto. When Does Virtual Embodiment Change Our Minds? *Presence: Teleoperators and Virtual Environments*, 25(3):222–233, December 2016.

[7] Shannon K. T. Bailey, Cheryl I. Johnson, and Valerie K. Sims. Using Natural Gesture Interactions Leads to Higher Usability and Presence in a Computer Lesson. In Sebastiano Bagnara, Riccardo Tartaglia, Sara Albolino, Thomas Alexander, and Yushi Fujita, editors, *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, Advances in Intelligent Systems and Computing, pages 663–671, Cham, 2019. Springer International Publishing.

[8] Domna Banakou, Parasuram D. Hanumanthu, and Mel Slater. Virtual Embodiment of White People in a Black Virtual Body Leads to a Sustained Reduction in Their Implicit Racial Bias. *Frontiers in Human Neuroscience*, 10, 2016.

[9] C. Daniel Batson, Shannon Early, and Giovanni Salvarani. Perspective Taking: Imagining How Another Feels Versus Imaging How You Would Feel. *Personality and Social Psychology Bulletin*, 23(7):751–758, July 1997. Publisher: SAGE Publications Inc.

[10] Taizhou Chen, Lantian Xu, Xianshan Xu, and Kening Zhu. GestOnHMD: Enabling Gesture-based Interaction on Low-cost VR Head-Mounted Display. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2597–2607, May 2021. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[11] B.A. Ciccone, S.K.T. Bailey, and J.E. Lewis. The Next Generation of Virtual Reality: Recommendations for Accessible and Ergonomic Design. *Ergonomics in Design*, 2021.

[12] Caroline Davies. Welcome to your virtual cell: could you survive solitary confinement? *The Guardian*, April 2016.

[13] Nonny de la Peña, Peggy Weil, Joan Llobera, Elias Giannopoulos, Ausiàs Pomés, Bernhard Spanlang, Doron Friedman, Maria V Sanchez-Vives, and Mel Slater. Immersive Journalism: Immersive Virtual Reality for the First-Person Experience of News. *Presence: Teleoperators and Virtual Environments*, 19(4):291–301, August 2010.

[14] Window: devicemotion event - Web APIs | MDN. https://developer.mozilla.org/en-US/docs/Web/API/Window/devicemotion_event, April 2023. Accessed: 2023-06-09.

[15] Window: deviceorientation event - Web APIs | MDN. https://developer.mozilla.org/en-US/docs/Web/API/Window/deviceorientation_event, April 2023. Accessed: 2023-06-09.

[16] Abraham Georgiadis and Shahrouz Yousefi. Analysis of the user experience in a 3D gesture-based supported mobile VR game. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, VRST '17, pages 1–2, New York, NY, USA, November 2017. Association for Computing Machinery.

[17] Emblematic Group. Greenland Melting. https://emblematicgroup.com/experiences/greenland-melting/. Accessed: 2023-01-10.

[18] Emblematic Group. Hunger in Los Angeles. https://docubase.mit.edu/project/hunger-in-los-angeles/. Accessed: 2023-01-10.

[19] Hiroyuki Hakoda, Wataru Yamada, and Hiroyuki Manabe. Eye Tracking Using Built-in Camera for Smartphone-based HMD. In *Adjunct Publication of the 30th Annual ACM Symposium on User Interface Software and Technology*, UIST '17, pages 15–16, New York, NY, USA, October 2017. Association for Computing Machinery.

[20] Daniel Harley, Aneesh P. Tarun, Sara Elsharawy, Alexander Verni, Tudor Tibu, Marko Bilic, Alexander Bakogeorge, and Ali Mazalek. Mobile Realities: Designing for the Medium of Smartphone-VR. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, DIS '19, pages 1131–1144, New York, NY, USA, June 2019. Association for Computing Machinery.

[21] F. Heilemann, G. Zimmermann, and P. Münster. Accessibility Guidelines for VR Games - A Comparison and Synthesis of a Comprehensive Set. *Frontiers in Virtual Reality*, 2, 2021.

[22] Daniel Hepperle, Yannick Weiß, Andreas Siess, and Matthias Wölfel. 2D, 3D or speech? A case study on which user interface is preferable for what kind of object interaction in immersive virtual reality. *Computers & Graphics*, 82:321–331, August 2019.

[23] Teresa Hirzle, Jan Rixen, Jan Gugenheimer, and Enrico Rukzio. WatchVR: Exploring the Usage of a Smartwatch for Interaction in Mobile Virtual Reality. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, pages 1–6, New York, NY, USA, April 2018. Association for Computing Machinery.

[24] Jiefan Huang, Yingyu Guo, Xuan Li, Ning Zhang, Jiang Jiang, and Guangyu Wang. Evaluation of Positioning Accuracy of Smartphones under Different Canopy Openness. *Forests*, 13:1591, September 2022.

[25] Samuel A. Iacolina, Alessandro Lai, Alessandro Soro, and Riccardo Scateni. Natural Interaction and Computer Graphics Applications. *Eurographics Italian Chapter Conference 2010*, page 6 pages, 2010. Artwork Size: 6 pages ISBN: 9783905673807 Publisher: The Eurographics Association.

[26] Mirja Ilves, Yulia Gizatdinova, Veikko Surakka, and Esko Vankka. Head movement and facial expressions as game input. *Entertainment Computing*, 5(3):147–156, August 2014.

[27] Akira Ishii, Takuya Adachi, Keigo Shima, Shuta Nakamae, Buntarou Shizuki, and Shin Takahashi. FistPointer: Target Selection Technique using Mid-air Interaction for Mobile VR Environment. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, page 474, New York, NY, USA, May 2017. Association for Computing Machinery.

[28] JavaScript Environment Requirements – React. https://legacy.reactjs.org/docs/javascript-environment-requirements.html. Accessed: 2023-06-09.

[29] Sarah Jones. Disrupting the narrative: immersive journalism in virtual reality. *Journal of Media Practice*, 18(2-3):171–185, September 2017. Publisher: Routledge _eprint: https://doi.org/10.1080/14682753.2017.1374677.

[30] Sameer Kishore, Bernhard Spanlang, Guillermo Iruretagoyena, Shivashankar Halan, Dalila Szostak, and Mel Slater. A Virtual Reality Embodiment Technique to Enhance Helping Behavior of Police Toward a Victim of Police Racial Aggression. *PRESENCE: Virtual and Augmented Reality*, 28:5–27, October 2021.

[31] Jaron Lanier. A Vintage Virtual Reality Interview. http://www.jaronlanier.com/vrint.html, 1988. Accessed: 2022-12-02.

[32] Guangchuan Li, David Rempel, Yue Liu, Weitao Song, and Carisa Harris Adamson. Design of 3D Microgestures for Commands in Virtual Reality or Augmented Reality. *Applied Sciences*, 11(14):6375, January 2021. Number: 14 Publisher: Multidisciplinary Digital Publishing Institute.

[33] Richard Li, Victor Chen, Gabriel Reyes, and Thad Starner. ScratchVR: low-cost, calibration-free sensing for tactile input on mobile virtual reality enclosures. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, ISWC '18, pages 176–179, New York, NY, USA, October 2018. Association for Computing Machinery.

[34] Sarah Lopez, Yi Yang, Kevin Beltran, Soo Jung Kim, Jennifer Cruz Hernandez, Chelsy Simran, Bingkun Yang, and Beste F. Yuksel. Investigating Implicit Gender Bias and Embodiment of White Males in Virtual Reality with Full Body Visuomotor Synchrony. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Glasgow Scotland Uk, May 2019. ACM.

[35] Josh Lowensohn. Google's Cardboard turns your Android device into a VR headset. The Verge, June 2014. https://www.theverge.com/2014/6/25/5842188/googles-cardboard-turns-your-android-device-into-a-vr-headset. Accessed: 2022-12-06.

[36] Xueshi Lu, Difeng Yu, Hai-Ning Liang, Wenge Xu, Yuzheng Chen, Xiang Li, and Khalad Hasan. Exploration of Hands-free Text Entry Techniques for Virtual Reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 344–349, 2020.

[37] Siqi Luo and Robert J. Teather. Camera-Based Selection with Cardboard HMDs. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 1066–1067, March 2019. ISSN: 2642-5254.

[38] Martez Mott, Edward Cutrell, Mar Gonzalez Franco, Christian Holz, Eyal Ofek, Richard Stoakley, and Meredith Ringel Morris. Accessible by Design: An Opportunity for Virtual Reality. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 451–454, October 2019.

[39] Janet Horowitz Murray. *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. Simon and Schuster, 1997. Google-Books-ID: bzmSLtnMZJsC.

[40] Niels Nilsson, Rolf Nordahl, and Stefania Serafin. Immersion Revisited: A Review of Existing Definitions of Immersion and Their Relation to Different Theories of Presence. *Human Technology*, 12:108–134, November 2016.

[41] Dev.Opera — The W3C Device Orientation API: Detecting Orientation and Acceleration. https://dev.opera.com/articles/w3c-device-orientation-api/. Accessed: 2023-06-20.

[42] Sofia Adelaide Osimo, Rodrigo Pizarro, Bernhard Spanlang, and Mel Slater. Conversations between self and self as Sigmund Freud—A virtual body ownership paradigm for self counselling. *Scientific Reports*, 5(1):13899, September 2015. Number: 1 Publisher: Nature Publishing Group.

[43] Hunter Osking and John A. Doucette. Enhancing Emotional Effectiveness of Virtual-Reality Experiences with Voice Control Interfaces. In Dennis Beck, Anasol Peña-Rios, Todd Ogle, Daphne Economou, Markos Mentzelopoulos, Leonel Morgado, Christian Eckhardt, Johanna Pirker, Roxane Koitz-Hristov, Jonathon Richter, Christian Gütl, and Michael Gardner, editors, *Immersive Learning Research Network*, Communications in Computer and Information Science, pages 199–209, Cham, 2019. Springer International Publishing.

[44] Nikiforos M. Papachristos, Ioannis Vrellis, and Tassos A. Mikropoulos. A Comparison between Oculus Rift and a Low-Cost Smartphone VR Headset: Immersive User Experience and Learning. In *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*, pages 477–481, July 2017. ISSN: 2161-377X.

[45] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3):45–77, December 2007. Publisher: Routledge _eprint: https://doi.org/10.2753/MIS0742-1222240302.

[46] Francisco José Rodrigues de Pinho. Framework for Developing Interactive 360-Degree Video Adventure Games. Master's thesis, Faculdade de Engenharia da Universidade do Porto, July 2019. Accepted: 2020-02-03T03:44:28Z.

[47] Bruno Porras-Garcia, Marta Ferrer-Garcia, Eduardo Serrano-Troncoso, Marta Carulla-Roig, Pau Soto-Usera, Helena Miquel-Nabau, Laura Fernández-Del Castillo Olivares, Rosa Marnet-Fiol, Isabel de la Montaña Santos-Carrasco, Bianca Borszewski, Marina Díaz-Marsá, Isabel Sánchez-Díaz, Fernando Fernández-Aranda, and José Gutiérrez-Maldonado. AN-VR-BE. A Randomized Controlled Trial for Reducing Fear of Gaining Weight and Other Eating Disorder Symptoms in Anorexia Nervosa through Virtual Reality-Based Body Exposure. *Journal of Clinical Medicine*, 10(4):682, February 2021.

[48] Marina Proske, Erik Poppe, and Melanie Jaeger-Erben. The smartphone evolution - an analysis of the design evolution and environmental impact of smartphones. In *Electronics Goes Green 2020+*, September 2020.

[49] react-camera-pro. https://www.npmjs.com/package/react-camera-pro, April 2023. Accessed: 2023-06-09.

[50] Grand View Research. Virtual Reality Market Size & Share Report, 2022-2030. Technical report, Grand View Research, June 2020. https://www.grandviewresearch.com/industry-analysis/virtual-reality-vr-market. Accessed: 2022-12-05.

[51] Dan Saffer. *Designing gestural interfaces*. O'Reilly, 1st edition, 2009. OCLC: ocn232976935.

[52] Samsung and Oculus Introduce the First Consumer Version of Gear VR. https://news.samsung.com/global/samsung-and-oculus-introduce-the-first-consumer-version-of-gear-vr. Accessed: 2022-12-06.

[53] S. Seinfeld, J. Arroyo-Palacios, G. Iruretagoyena, R. Hortensius, L. E. Zapata, D. Borland, B. de Gelder, M. Slater, and M. V. Sanchez-Vives. Offenders become the victim in virtual reality: impact of changing perspective in domestic violence. *Scientific Reports*, 8(1):2692, February 2018. Number: 1 Publisher: Nature Publishing Group.

[54] Sofia Seinfeld, Ruud Hortensius, Jorge Arroyo-Palacios, Guillermo Iruretagoyena, Luis E. Zapata, Beatrice de Gelder, Mel Slater, and Maria V. Sanchez-Vives. Domestic Violence From a Child Perspective: Impact of an Immersive Virtual Reality Experience on Men With a History of Intimate Partner Violent Behavior. *Journal of Interpersonal Violence*, page 08862605221106130, June 2022. Publisher: SAGE Publications Inc.

[55] Wei Shen. Natural Interaction Technology in Virtual Reality. In *2021 International Symposium on Artificial Intelligence and its Application on Media (ISAIAM)*, pages 1–4, May 2021.

[56] Jake Silverstein. The Displaced: Introduction. *The New York Times*, November 2015.

[57] Mel Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3549–3557, December 2009.

[58] Mel Slater and Sylvia Wilbur. A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 6(6):603–616, December 1997.

[59] Andrej Somrak, Iztok Humar, M. Shamim Hossain, Mohammed F. Alhamid, M. Anwar Hossain, and Jože Guna. Estimating VR Sickness and user experience using different HMD technologies: An evaluation study. *Future Generation Computer Systems*, 94:302–316, May 2019.

[60] Misha Sra, Xuhai Xu, and Pattie Maes. BreathVR: Leveraging Breathing as a Directly Controlled Interface for Virtual Reality Games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–12, New York, NY, USA, April 2018. Association for Computing Machinery.

[61] S. Shyam Sundar, Jin Kang, and Danielle Oprean. Being There in the Midst of the Story: How Immersive Journalism Affects Our Perceptions and Cognitions. *Cyberpsychology, Behavior, and Social Networking*, 20(11):672–682, November 2017.

[62] Ivan E Sutherland. The Ultimate Display. *Proceedings of the IFIP Congress*, pages 506–509, 1965.

[63] Ivan E Sutherland. A head-mounted three dimensional display. *AFIPS '68 (Fall, part I)*, page 8, 1968.

[64] Ana Luisa Sánchez Laws. Can Immersive Journalism Enhance Empathy? *Digital Journalism*, 8(2):213–228, February 2020. Publisher: Routledge _eprint: https://doi.org/10.1080/21670811.2017.1389286.

[65] Fundamentals - three.js manual. https://threejs.org/manual/#en/fundamentals. Accessed: 2023-06-09.

[66] Sam Tregillus and Eelke Folmer. VR-STEP: Walking-in-Place using Inertial Sensing for Hands Free Navigation in Mobile VR Environments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 1250–1255, New York, NY, USA, May 2016. Association for Computing Machinery.

[67] Pawan Verma, Rohit Kumar, Jai Tuteja, and Neha Gupta. Systematic Review Of Virtual Reality & Its Challenges. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pages 434–440, February 2021.

[68] Web Speech API - Web APIs | MDN. https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API, February 2023. Accessed: 2023-06-14.

[69] WebXR Device API - Web APIs | MDN. https://developer.mozilla.org/en-US/docs/Web/API/WebXR_Device_API, February 2023. Accessed: 2023-06-09.

[70] Matthew Wilber. nod.js. https://github.com/mwilber/nod.js, July 2017. Accessed: 2023-06-09.

[71] Bob G. Witmer and Michael J. Singer. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments*, 7(3):225–240, June 1998.

[72] Huiyue Wu, Weizhou Luo, Neng Pan, Shenghuan Nan, Yanyi Deng, Shengqian Fu, and Liuqingqing Yang. Understanding freehand gestures: a study of freehand gestural interaction for immersive VR shopping applications. *Human-centric Computing and Information Sciences*, 9(1):43, December 2019.

[73] Xiaoqi Yan, Chi-Wing Fu, Pallavi Mohan, and Wooi Boon Goh. CardboardSense: Interacting with DIY Cardboard VR Headset by Tapping. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, DIS '16, pages 229–233, New York, NY, USA, June 2016. Association for Computing Machinery.

[74] Chun Yu, Yizheng Gu, Zhican Yang, Xin Yi, Hengliang Luo, and Yuanchun Shi. Tap, Dwell or Gesture? Exploring Head-Based Text Entry Techniques for HMDs. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 4479–4488, New York, NY, USA, May 2017. Association for Computing Machinery.

[75] Majed Al Zayer, Sam Tregillus, and Eelke Folmer. PAWdio: Hand Input for Mobile VR using Acoustic Sensing. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '16, pages 154–158, New York, NY, USA, October 2016. Association for Computing Machinery.

Appendix A

Input JSON File Example

```
[
    {
        "type": "interaction",
        "interaction":{
            "interaction_type": "key"
        },
        "video_src": "office/0_.mp4"
    },
    {
        "type": "video",
        "video_src": "office/female/1_.mp4"
    },
    {
        "type": "interaction",
        "interaction":{
            "interaction_type": "speech",
            "expected_inputs": ["hello", "anyone", "hi", "hey"]
        },
        "video_src": "office/2_.mp4"
    },
    {
        "type": "video",
        "video_src": "office/female/3_.mp4"
    },
    {
        "type": "interaction",
        "interaction":{
            "interaction_type": "head",
            "head_interaction_type": "look",
            "value": {
                "alpha": 45,
                "gamma": 0,
                "beta": 0
```

```
                }
            },
            "video_src": "office/4_.mp4"
        },
        {
            "type": "video",
            "video_src": "office/female/5_.mp4"
        },
        {
            "type": "interaction",
            "interaction":{
                "interaction_type": "hands"
            },
            "video_src": "office/6_.mp4"
        },
        {
            "type": "video",
            "video_src": "office/female/7_.mp4"
        },
        {
            "type": "interaction",
            "interaction":{
                "interaction_type": "head",
                "head_interaction_type": "nod_no"
            },
            "video_src": "office/8_.mp4"
        },
        {
            "type": "video",
            "video_src": "office/female/9_.mp4"
        },
        {
            "type": "interaction",
            "interaction":{
                "interaction_type": "speech",
                "expected_inputs": ["goodbye", "bye", "tomorrow", "later", "soon"]
            },
            "video_src": "office/10_.mp4"

        },
        {
            "type": "video",
            "video_src": "office/female/11_.mp4"
        },
        {
            "type": "interaction",
            "interaction":{
                "interaction_type": "hands"
            },
```

```
            "video_src": "office/12_.mp4"
        },
        {
            "type": "video",
            "video_src": "office/13_.mp4"
        }
    ]
```

Appendix B

Output JSON File Example


```
[
    {
        "type": "speech",
        "duration": 4126,
        "normalTermination": true,
        "numberOfTries": 1,
        "register": [
            {
                "input": "hello",
                "confidence": 0.9680845141410828
            }
        ],
        "finalInput": "hello",
        "finalConfidence": 0.9680845141410828,
        "inputMatch": "hello"
    },
    {
        "type": "head",
        "head_type": "look",
        "duration": 1074,
        "normalTermination": true,
        "target": "(0.43128746351189523, 0.36652575112823826, 0.5587996217924666,
            0.6061295078297646)",
        "finalQuaternion": "(0.3388961670987477, 0.42685473360536696,
            0.4736585296116213, 0.691803455940071)",
        "finalAngle": 0.3275648257487108,
        "register": [
            {
                "quaternion": {
                    "isQuaternion": true,
                    "_x": 0.3557244312537605,
                    "_y": 0.43846905028051725,
                    "_z": 0.4718318406081871,
```

```
                    "_w": 0.6771851557318576
                },
                "angle": 0.3068815247312919
            },
            (...)
            {
                "quaternion": {
                    "isQuaternion": true,
                    "_x": 0.3388961670987477,
                    "_y": 0.42685473360536696,
                    "_z": 0.4736585296116213,
                    "_w": 0.691803455940071
                },
                "angle": 0.3275648257487108
            }
        ]
    },
    {
        "type": "hands",
        "duration": 2393,
        "normalTermination": true,
        "finalValue": 48.666666666666664,
        "register": [
            139.66666666666666,
            149.33333333333334,
            (\dots)
            96.33333333333333,
            48.666666666666664
        ]
    },
    {
        "type": "head",
        "head_type": "nod_no",
        "duration": 1519,
        "normalTermination": true,
        "register": [
            "up",
            "left",
            "up",
            "right"
        ],
        "registerMotion": [
            {
                "alpha": 5.500000000000001,
                "beta": -1,
                "gamma": -3
            },
            (...)
            {
```

```
                "alpha": -8.4,
                "beta": -0.9000000000000001,
                "gamma": -7.300000000000001
            }
        ]
    },
    {
        "type": "speech",
        "duration": 2482,
        "normalTermination": true,
        "numberOfTries": 1,
        "register": [
            {
                "input": "goodbye",
                "confidence": 0.9473445415496826
            }
        ],
        "finalInput": "goodbye",
        "finalConfidence": 0.9473445415496826,
        "inputMatch": "goodbye"
    },
    {
        "type": "hands",
        "duration": 4972,
        "normalTermination": false,
        "finalValue": 151,
        "register": [
            101.66666666666667,
            153.33333333333334,
            (...)
            171,
            151
        ]
    }
]
```

Appendix C

Scenarios Scripts

C.1 Presentation Scenario Script

```
Actors: STUDENT (POV) and TEACHER (Mr. White).
Surroundings: Classroom or presentation room. The STUDENT (user) is standing in
    front/side of a television on which a presentation is being shown. The TEACHER
     is sitting, facing the presentation and the STUDENT. THE STUDENT is also
    facing the TEACHER.
Context: The STUDENT just finished their presentation and they're now entering a Q&
    A with the TEACHER.
Note: The STUDENT's THOUGHTs will be heard with a voice-over as a way to guide
    their actions (interactions).

THOUGHT: Well... This could have gone better. But let's take a deep breath. Just
    the Q&A left... I can do this! Let me just tell Mr White I'm ready to start.
SPEECH INTERACTION: "(i'm/ i am) ready."
TEACHER: I must say I'm not impressed with your presentation. You don't seem like
    you prepared yourself that much for it. For example, tell me again, when did
    the war really start? What month precisely?
THOUGHT: What is the date again? I can't really remember... I had it in the
    presentation though, a peek couldn't hurt.
HEAD INTERACTION: look at the screen
THOUGHT: There it is. October '39. Now, let me answer it confidently.
SPEECH INTERACTION: "october"
TEACHER: Exactly, just as I thought. You should already know by now that the war
    started in September and not in October! These mistakes are unforgivable at
    your level!
THOUGHT: Oh no... I can't believe I messed up the dates! I'm so embarrassed, I wish
     I could disappear from here. I want to cover my eyes so I don't have to see
    this disaster happening.
HANDS INTERACTION: cover camera
VIDEO: cover the screen with hands
TEACHER: No point in hiding yourself now...
VIDEO: uncover screen
```

```
TEACHER: Right now, I cannot give you more than the bare minimum passing grade. But
     I'm willing to give you a second opportunity. I will let you review the
    presentation and tomorrow we can meet again for a new try. But as it is a
    second chance, I will not tolerate any mistakes at all. So... do you want to
    try again tomorrow?
THOUGHT: I know I have it in me to do better! I'll work hard and tomorrow I'll be
    more confident! Right now though, I'm just too shaky to even talk... I'll
    answer with a nod.
HEAD INTERACTION: nod yes
TEACHER: Very well then. I will email you later with the details. You may leave now
    .
TEACHER: *opens the door of the room and a very bright light shines in*
VIDEO: same view but with brightness almost to 100\% to imitate bright light coming
     in
THOUGHT: So much light! Got to protect my eyes!
HANDS INTERACTION: cover camera
VIDEO: cover the screen with arm
```

C.2 Office Scenario Script

```
Actors: SAM (POV) and KYLE (SAM's work colleague).
Surroundings: An office. There's a door on one of the sides of the room (from SAM's
     perspective).
Context: SAM went to their office at night to prepare for their big presentation
    the next day.
Note: SAM's THOUGHTs will be heard with a voice-over as a way to guide their
    actions (interactions).

THOUGHT: I think I'm feeling ready. Already have everything memorized. As long as I
     can manage my stress, I'm gonna do well in my presentation.
VIDEO: footsteps sounds
THOUGHT: What is this? There isn't supposed to be anyone else in the building.
    Maybe I should call out hello to see if someone replies.
SPEECH INTERACTION: "hello"
THOUGHT: Hmmm, nothing... I'm so tired I must be starting to hear things...
VIDEO: the lights go off and more footsteps are heard, closer
THOUGHT: Who could be there? Whatever it is, it seems to be heading this way... Did
     I close the door? I can't remember...
HEAD INTERACTION: look at the door
THOUGHT: Oh no! I left it open. And they seem to be getting closer and closer. No
    time to go close the door now.
VIDEO: A shadow enters through the door and approaches SAM.
THOUGHT: I'm so afraid. Are they gonna attack me? I need to protect myself. My face
     first!
HANDS INTERACTION: Cover camera
VIDEO: cover the screen with arm
```

```
KYLE: Sam... Is that you?
THOUGHT: This voice sounds familiar. But I can't quite place it...
VIDEO: uncover screen
KYLE: It's me, Kyle. You seem frightened, Sam. Are you okay?
THOUGHT: Of course I'm not okay, Kyle. I can't stop shaking. I can't even speak.
    All I can do is shake my head to answer you.
HEAD INTERACTION: nod no.
KYLE: I'm so sorry, I didn't mean to scare you. I just forgot my computer and came
    to pick it up. I wasn't expecting anybody else to be in the building. I saw the
     lights on but I just thought someone had forgotten to turn them off before
    going home earlier. Anyway, I'm so sorry again, Sam! I need to get going now.
    See you tomorrow!
THOUGHT: *takes deep breath* Am I calm enough to at least say some words? Let's
    just try to say goodbye to Kyle.
SPEECH INTERACTION: "goodbye" OR "bye" OR "(see you/until) tomorrow" OR "(see you/
    until/talk to you) later" OR "(see you/until) soon"
VIDEO: Kyle goes out
THOUGHT: I swear to god I cannot understand this guy... Only him to scare me like
    this and get me even more anxious than I already am. Ok, let's calm down now.
    Remember what the therapist told you: cover your eyes, take a deep breath and
    everything will be fine.
HANDS INTERACTION: cover camera
VIDEO: cover the screen with hands
```

Appendix D

Declarations of Informed Consent

INFORMED CONSENT STATEMENT

In the context of the development of the Master dissertation of the Master in Informatics and Computing Engineering at the Faculty of Engineering of the University of Porto, entitled **"Natural Interaction in Smartphone Virtual Reality Experiences"**, conducted by the student Leonor Martins de Sousa, supervised by Prof. Rui Rodrigues and under the co-supervision of Prof. Teresa Matos, I, the undersigned, declare that I have understood the explanation provided to me regarding the study in which I will participate, specifically the voluntary nature of this participation and that I have been given the opportunity to ask any necessary questions.

I have been informed that the information or explanation provided to me covered the objectives, methods, potential discomfort, and the absence of risks to my health and that the utmost confidentiality of the data will be ensured.

Furthermore, I have been explained that I may withdraw from the study at any time without any disadvantages resulting from it.

Therefore, I consent to participate in the study and the collection of necessary data, answering all proposed questions.

Porto, __ of _____ of 2023

_____

(participant or their representative)

Figure D.1: Informed consent statement.

Declaração de Consentimento de Direitos de Imagem

No âmbito da realização da tese de estrado do Mestrado de Engenharia Informática e Computação da Faculdade de Engenharia da Universidade do Porto, intitulada **"Natural Interaction in Smartphone Virtual Reality Experiences"**, realizada pelo estudante Leonor Martins de Sousa, orientada pelo Prof. Rui Rodrigues e sob a co-orientação da Prof. Teresa Matos, eu abaixo assinado declaro que autorizo à filmagem da minha imagem, bem como a difundi-la no contexto de investigação acima mencionado.

A presente autorização é concedida a **título gratuito.**

Porto, __ de _____ de 20__

_____

(Participante ou seu representante)

Figure D.2: Informed consent statement for image rights.

Appendix E

User Test Script

- Welcome
- I'm doing this experiment for my thesis which explores Natural Interaction in Smartphone VR Experiences
- During this user test, you are going to experience two different scenarios, from the point of view of a main character. In one of those scenarios you will only be able to observe what's happening, **while** in the other one, you will have to interact with the scene, using 3 different types of interaction (which I will explain later).
- Before we begin, I will ask you to sign this informed consent form and then to fill a **short** questionnaire with demographic questions, as well as about your experience with VR. After each one of the scenarios, I will also ask you to fill out a questionnaire about your experience. All of this data will only be used **for** statistical purposes, and you will not be identified.
- During the experiences, I will also record you. These recordings will not be published. They will only be used **for** me to recheck some data I've collected automatically during the experience.
- In regard to the experiment, it is also possible, although not likely, that you experience some nausea.
- Is this okay with you?
- Let's start then. Here is the informed consent.
- Here is the first part of the questionnaire.
- Before we begin, **do** you identify most with a female or male voice?
- Now, let's start with the first scenario.

(1ST SCENARIO PART)

- I'm going to start recording and give you the headset now.
- Warning: there is some lagging between some videos.
- *start recording*
- *give headset to participant*
- *warn **for** camera location and adjust lensis distance*
- Before the scenario itself starts I will give you some time to look around the scene and get used to it and once you're ready to begin, just say the word
- *begin*
- *once scenario is over: end recording*

```
- *fill questionnaire*
- let's go on to the second scenario
(2ST SCENARIO PART)
- Warning: there is some lagging between some videos.
- I'm going to start recording and give you the headset now.
- *start recording*
- *give headset to participant*
- Before the scenario itself starts I will give you some time to look around the
    scene and get used to it and once you're ready to begin, just say the word
- *begin*
- *once scenario is over: end recording*
- *fill questionnaire*
- Thank you so much for participating!

OFFICE SCENARIO:
- In this scenario, you will impersonate Sam, which works in an office. Sam decided
     to come to the office at night when nobody else is in the building. They have
    an important presentation the next morning, and they wanted to prepare.
- During the scene, you will be able to look all around you, but you will not be
    able to move.
- You will be able to hear your own thoughts.
(INTERACTION PART)

PRESENTATION SCENARIO
- In this scenario, you will impersonate a student that is doing a presentation to
    their teacher. You just finished doing the presentation and the teacher is
    going to ask you some questions and do some comments on your presentation.
- During the scene, you will be able to look all around you, but you will not be
    able to move.
- You will be able to hear your own thoughts.
(INTERACTION PART)

WITH INTERACTION:
- for this part of the experiment, you will be able to interact with the scene.
- the interaction will be done in 6 specific moments.
- for each of the interaction moments, you will have to use 1 of 3 types of
    interaction
    - you can use your voice to say some words
    - you can use your head to do some movements
    - or you can use your hands or arms to do some very simple movements
- you should be able to understand when and how to interact based on the thoughts,
    so it's very important that you pay close attention to them
- if after a few tries you are not able to perform the interaction correctly, I
    will advance the scene myself so you don't get stuck on a part of the
    experiment.

WITHOUT INTERACTION:
- for this part of the experiment, you will not have the possibility to interact
    with the scene. You will only observe what's happening and hear your thoughts.
```

Appendix F

User Experience Questionnaire

User Study Questionnaire

* Indica uma pergunta obrigatória

1. Participant's ID *

2. Scenario Order *

Marcar apenas uma oval por linha.

	Presentation	Office
1st Scenario	◯	◯
2nd Scenario	◯	◯

User Information

3. What gender do you identify yourself as? *

Marcar apenas uma oval.

- ◯ Female
- ◯ Male
- ◯ Non binary
- ◯ Outra: _____

4. What is your age? *

Marcar apenas uma oval.

- ◯ 19 years or less
- ◯ 20-24 years
- ◯ 25-29 years
- ◯ 30-34 years
- ◯ 35-39 years
- ◯ 40-44 years
- ◯ 45-49 years
- ◯ 50-54 years
- ◯ 55-59 years
- ◯ 60 years or more

5. How many times have you experienced smartphone Virtual Reality in the past? *

Marcar apenas uma oval.

- 0
- 1-2
- 3-5
- 5-10
- 10+

6. How many times have you experienced other types of Virtual Reality in the past? *

Marcar apenas uma oval.

- 0
- 1-2
- 3-5
- 5-10
- 10+

7. On a scale of 1 to 5, how nauseated do you feel now (before initiating the experiment)? *

Marcar apenas uma oval.

	1	2	3	4	5	
not	◯	◯	◯	◯	◯	very nauseated

PART 1

8. Did you experience this part of the experiment: *

Marcar apenas uma oval.

- with interaction. *Avançar para a pergunta 11*
- without interaction. *Avançar para a pergunta 22*

9. On a scale of 1 to 5, rate your agreement with the following statements: *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
The scenario I experienced felt realistic.	◯	◯	◯	◯	◯
I was aware of my real surroundings during the experience.	◯	◯	◯	◯	◯
I felt immersed in the environment.	◯	◯	◯	◯	◯
I felt like I was in the portrayed place.	◯	◯	◯	◯	◯
I felt like I was living the portrayed situation.	◯	◯	◯	◯	◯
I felt like I was in the body of the Main Character.	◯	◯	◯	◯	◯
I felt empathy towards the Main Character.	◯	◯	◯	◯	◯
I enjoyed the experience.	◯	◯	◯	◯	◯

10. On a scale of 1 to 5, how nauseated do you feel after completing this part of the experiment? *

Marcar apenas uma oval.

	1	2	3	4	5	
not	◯	◯	◯	◯	◯	very nauseated

Part 1 - Interaction Questions

Speech interaction corresponds to the moments where you had to say aloud one or more words.

Hands interaction corresponds to the moments where you had to cover your eyes with your hands and/or arms.
Head interaction corresponds to the moments where you had to move your head to look somewhere specific or to nod (yes or no).

11. Rate the types of interaction experienced, according to your preference. *

Marcar apenas uma oval por linha.

	1º	2º	3º
Speech	◯	◯	◯
Hands	◯	◯	◯
Head	◯	◯	◯

For the rest of this section, rate on a scale of 1 to 5 your agreement with each of the statements, for each type of interaction.

12. The interaction felt intuitive. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech	◯	◯	◯	◯	◯
Hands	◯	◯	◯	◯	◯
Head	◯	◯	◯	◯	◯

13. The interaction was easy to perform. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech	◯	◯	◯	◯	◯
Hands	◯	◯	◯	◯	◯
Head	◯	◯	◯	◯	◯

14. The interaction felt natural given the scenario. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech	◯	◯	◯	◯	◯
Hands	◯	◯	◯	◯	◯
Head	◯	◯	◯	◯	◯

15. It was clear to me what I was expected to do. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech	◯	◯	◯	◯	◯
Hands	◯	◯	◯	◯	◯
Head	◯	◯	◯	◯	◯

16. The interaction contributed to me feeling immersed in the environment. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech	◯	◯	◯	◯	◯
Hands	◯	◯	◯	◯	◯
Head	◯	◯	◯	◯	◯

17. The interaction contributed to me feeling like I was in the portrayed place. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech	◯	◯	◯	◯	◯
Hands	◯	◯	◯	◯	◯
Head	◯	◯	◯	◯	◯

18. The interaction contributed to me feeling like I was living the portrayed situation. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech	◯	◯	◯	◯	◯
Hands	◯	◯	◯	◯	◯
Head	◯	◯	◯	◯	◯

19. The interaction contributed to me feeling like I was in the body of the Main Character. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech					
Hands					
Head					

20. The interaction contributed to me feeling empathy towards the Main Character. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech					
Hands					
Head					

21. The interaction contributed to me enjoying the experience. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech					
Hands					
Head					

Avançar para a pergunta 22

PART 2

22. Did you experience this part of the experiment: *

Marcar apenas uma oval.

- ◯ with interaction. *Avançar para a pergunta 25*
- ◯ without interaction. *Avançar para a pergunta 36*

23. On a scale of 1 to 5, rate your agreement with the following statements: *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
The scenario I experienced felt realistic.	◯	◯	◯	◯	◯
I was aware of my real surroundings during the experience.	◯	◯	◯	◯	◯
I felt immersed in the environment.	◯	◯	◯	◯	◯
I felt like I was in the portrayed place.	◯	◯	◯	◯	◯
I felt like I was living the portrayed situation.	◯	◯	◯	◯	◯
I felt like I was in the body of the Main Character.	◯	◯	◯	◯	◯
I felt empathy towards the Main Character.	◯	◯	◯	◯	◯
I enjoyed the experience.	◯	◯	◯	◯	◯

24. On a scale of 1 to 5, how nauseated do you feel after completing this part of the experiment? *

Marcar apenas uma oval.

	1	2	3	4	5	
not	◯	◯	◯	◯	◯	very nauseated

Part 2 - Interaction Questions

Speech interaction corresponds to the moments where you had to say aloud one or more words.

Hands interaction corresponds to the moments where you had to cover your eyes with your hands and/or arms.
Head interaction corresponds to the moments where you had to move your head to look somewhere specific or to nod (yes or no).

25. Rate the types of interaction experienced, according to your preference. *

Marcar apenas uma oval por linha.

	1º	2º	3º
Speech	◯	◯	◯
Hands	◯	◯	◯
Head	◯	◯	◯

For the rest of this section, rate on a scale of 1 to 5 your agreement with each of the statements, for each type of interaction.

26. The interaction felt intuitive. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech	◯	◯	◯	◯	◯
Hands	◯	◯	◯	◯	◯
Head	◯	◯	◯	◯	◯

27. The interaction was easy to perform. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech	◯	◯	◯	◯	◯
Hands	◯	◯	◯	◯	◯
Head	◯	◯	◯	◯	◯

28. The interaction felt natural given the scenario. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech	◯	◯	◯	◯	◯
Hands	◯	◯	◯	◯	◯
Head	◯	◯	◯	◯	◯

29. It was clear to me what I was expected to do. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech	◯	◯	◯	◯	◯
Hands	◯	◯	◯	◯	◯
Head	◯	◯	◯	◯	◯

30. The interaction contributed to me feeling immersed in the environment. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech					
Hands					
Head					

31. The interaction contributed to me feeling like I was in the portrayed place. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech					
Hands					
Head					

32. The interaction contributed to me feeling like I was living the portrayed situation. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech					
Hands					
Head					

33. The interaction contributed to me feeling like I was in the body of the Main Character. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech	◯	◯	◯	◯	◯
Hands	◯	◯	◯	◯	◯
Head	◯	◯	◯	◯	◯

34. The interaction contributed to me feeling empathy towards the Main Character. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech	◯	◯	◯	◯	◯
Hands	◯	◯	◯	◯	◯
Head	◯	◯	◯	◯	◯

35. The interaction contributed to me enjoying the experience. *

Marcar apenas uma oval por linha.

	1 (strongly disagree)	2 (disagree)	3 (neither agree nor disagree)	4 (agree)	5 (strongly agree)
Speech	◯	◯	◯	◯	◯
Hands	◯	◯	◯	◯	◯
Head	◯	◯	◯	◯	◯

Thank You for Participating!

36. Any observations, comments or suggestions?