Multimodal 3D User Interfaces for Augmented Reality and Omni-Directional Video

Gustavo Alberto Rovelo Ruiz

Departamento de Sistemas Informáticos y Computación

Universitat Politècnica de València

A Thesis submitted for the degree of

PhilosophiæDoctor in Computer Science

Under the Supervision of

Emilio Camahort, PhD and Francisco Abad, PhD

July 13rd, 2015

Members of the examining committee

Roberto Vivó (Universitat Politècnica de València, Spain)
Luis Matey (Universidad de Navarra, Spain)
Jean Vanderdonckt (Université catholique de Louvain, Belgium)

Members of the external evaluation committee

Géry Casiez (Université Lille 1, France)
Laurence Nigay (Université Joseph Fourier, France)
Jean Vanderdonckt (Université catholique de Louvain, Belgium)

To my family.

Acknowledgements

This document represents the end of one important step in my professional and personal life: achieving a goal that I set myself some years ago when I started my bachelor studies. During these years, full of enriching experiences working on my PhD project, I have had the privilege to meet and work with many people. People that have helped me to grow professionally, but most importantly as a person. I would like to express my gratitude to all of you in the following paragraphs.

First of all, my gratitude goes to Emilio Camahort and Francisco Abad, my PhD advisers. Thank you both for all the hours helping me to fill out all the reports, to prepare all the publications, to read and polish this text and in general, for helping me to put together the ideas of this research project.

Thank you very much to the friends that helped me proofreading the text of my PhD dissertation to improve it: Ryan Taylor, Marisela Gutiérrez, Kashyap Todi, Donald Degraen, Doug Stutzman, Mauricio Rodriguez, Patrik Goorts, Steven Maesen, and Emilio Granell. Thank you for the time you spent helping me.

I also thank Davy Vanacken and Kris Luyten for their guidance during my internship at Expertise Centre for Digital Media (EDM) of Hasselt University and during the last stage of the PhD. Thank you also to Wim Lamotte, Peter Quax and Maarten Wijnants for the opportunity to work with you on the AIVIE project. To Philippe Bekaert and Nick Michiels for their support using the facilities of the Omnidirectional CAVE of EDM. Thank you also to Ingrid Konings, Roger Claes and Peter Vandoren for the help they gave me making administrative work

easier. Thanks to all the good friends and colleagues I have made in Hasselt during the time I have been working at EDM. Especially, thank you Sabine Hubrechts.

Thank you to Géry Casiez, Laurence Nigay, Jean Vanderdonckt, members of the external group of reviewers of my dissertation, for their valuable feedback on the final draft of the thesis. Your comments helped me to improve the presentation of the PhD project. Thank you also to Luis Matey and Roberto Vivó who, together with Jan Vanderdonckt were part of my examining committee.

Thank you to all the good friends I have made in Valencia: Francisco, Armando, Vladimir, Iván, Emilio, Ihab, David, Antonio, Fernando, Carlos, Lupita, Axel, Malene, Roxane, Fernanda, Rime, Vivian, Carmen, Paulina, Adriana and German, Imelda and Ruben, Laura, Angela, Arantxa and Fernando, Isabel, Martha and José Luis, Chiara, Cinzia, Noora, Sabine, Elisabetta, César, Hugo, Ricardo, José, Mariela, Ismael, Emmy and Pablo. To Silvia and Alberto for their friendship, and the hours of good talk and fun. To Vicente, Santi, Johanna, Lorena, Lucian, David and Juan Fernando, my colleagues at the laboratory in Valencia. Also to Mari-Carmen Juan for her help during my years at the UPV. I apologise if I forgot to include some name(s). Fortunately you are a big group and it is hard to remember everyone that was an important part of my life during these years.

My deepest greetings to my loving parents and my sister. To all my family and friends in México that even when being far away (geographically speaking) they were always supporting and cheering me up.

Thank you also to the Instituto Tecnológico de Tuxtla Gutiérrez, México, and especially to professors Hector Guerra, Ariosto Rios, Franciso Suarez and Raul Paredes for their help to set up part of the experiments. Which remind me also to thank everyone of the participants that took part of my experiments.

Finally, last but not least, thank you to the National Council of Science and Technology of México (CONACyT) for funding my PhD studies through the 207587 scholarship.

Gustavo Rovelo
Valencia, Spain
July 2015.

Abstract

Human-Computer Interaction is a multidisciplinary research field that combines, amongst others, Computer Science and Psychology. It studies human-computer interfaces from the point of view of both, technology and the user experience.

Researchers in this area have now a great opportunity, mostly because the technology required to develop 3D user interfaces for computer applications (e.g. visualization, tracking or portable devices) is now more affordable than a few years ago.

Augmented Reality and Omni-Directional Video are two promising examples of this type of interfaces where the user is able to interact with the application in the three-dimensional space beyond the 2D screen.

The work described in this thesis is focused on the evaluation of interaction aspects in both types of applications. The main goal is contributing to increase the knowledge about this new type of interfaces to improve their design. We evaluate how computer interfaces can convey information to the user in Augmented Reality applications exploiting human multisensory capabilities. Furthermore, we evaluate how the user can give commands to the system using more than one type of input modality, studying Omnidirectional Video gesture-based interaction. We describe the experiments we performed, outline the results for each particular scenario and discuss the general implications of our findings.

Resumen

El campo de la Interacción Persona-Computadora es un área multidisciplinaria que combina, entre otras a las Ciencias de la Computación y Psicología. Estudia la interacción entre los sistemas computacionales y las personas considerando tanto el desarrollo tecnológico, como la experiencia del usuario.

Los dispositivos necesarios para crear interfaces de usuario 3D son ahora más asequibles que nunca (v.gr. dispositivos de visualización, de seguimiento o móviles) abriendo así un area de oportunidad para los investigadores de esta disciplina. La Realidad Aumentada y el Video Omnidireccional son dos ejemplos de este tipo de interfaces en donde el usuario es capaz de interactuar en el espacio tridimensional más allá de la pantalla de la computadora.

El trabajo presentado en esta tesis se centra en la evaluación de la interacción del usuario con estos dos tipos de aplicaciones. El objetivo principal es contribuir a incrementar la base de conocimiento sobre este tipo de interfaces y así, mejorar su diseño. En este trabajo investigamos de qué manera se pueden emplear de forma eficiente las interfaces multimodales para proporcionar información relevante en aplicaciones de Realidad Aumentada. Además, evaluamos de qué forma el usuario puede usar interfaces 3D usando más de un tipo de interacción; para ello evaluamos la interacción basada en gestos para Video Omnidireccional.

A lo largo de este documento se describen los experimentos realizados y los resultados obtenidos para cada caso en particular. Se presenta además una discusión general de los resultados.

Resum

El camp de la Interacció Persona-Ordinador és una àrea d'investigació multidisciplinar que combina, entre d'altres, les Ciències de la Informàtica i de la Psicologia. Estudia la interacció entre els sistemes computacionals i les persones considerant tant el desenvolupament tecnològic, com l'experiència de l'usuari.

Els dispositius necessaris per a crear interfícies d'usuari 3D són ara més assequibles que mai (v.gr. dispositius de visualització, de seguiment o mòbils) obrint així una àrea d'oportunitat per als investigadors d'aquesta disciplina. La Realitat Augmentada i el Vídeo Omnidireccional són dos exemples d'aquest tipus d'interfícies on l'usuari és capaç d'interactuar en l'espai tridimensional més enllà de la pantalla de l'ordinador.

El treball presentat en aquesta tesi se centra en l'avaluació de la interacció de l'usuari amb aquests dos tipus d'aplicacions. L'objectiu principal és contribuir a augmentar el coneixement sobre aquest nou tipus d'interfícies i així, millorar el seu disseny. En aquest treball investiguem de quina manera es poden utilitzar de forma eficient les interfícies multimodals per a proporcionar informació rellevant en aplicacions de Realitat Augmentada. A més, avaluem com l'usuari pot utilitzar interfícies 3D utilitzant més d'un tipus d'interacció; per aquesta raó, avaluem la interacció basada en gest per a Vídeo Omnidireccional.

Al llarg d'aquest document es descriuen els experiments realitzats i els resultats obtinguts per a cada cas particular. A més a més, es presenta una discussió general dels resultats.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The ACM SIGCHI Curricula for Human-Computer Interaction (HCI) defines HCI as:

> *"a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them"* (Hewett et al., 1996).

It is a multidisciplinary research area that combines knowledge from different fields such as Computer Science, Industrial Design, Ergonomics or Psychology to study robust, effective and efficient communication techniques between users and computers. This communication process has two flows: commands from the user to the computer, and feedback from the computer to the user.

Recent technological advances have made possible the implementation of futuristic interfaces like those seen in Science-Fiction movies, such as "Minority Report" (20th Century Fox, 2002) or "Ironman" (Marvel Entertainment Inc., 2009). New hardware such as the Microsoft Kinect (Microsoft Inc., 2013), the Google Glass project (Google Inc., 2013) or the Leap Motion (Leap Motion, Inc., 2014), and powerful, affordable mobile devices open the door to wide adoption of a new generation of computer interfaces: 3D User Interfaces (3DUIs). The main feature of these interfaces is that they allow the user to work in a three-dimensional space (Bowman et al., 2004). Augmented Reality and Omni-Directional Video applications are two examples of this new generation of user interfaces.

Augmented Reality (AR) combines virtual objects within the view of a real scene. The user sees her surroundings through a device (typically eyeglasses, a smartphone or a head-mounted display) that overlays a virtual scene on top. The virtual objects are usually rendered with the same perspective as the real objects, and are used to provide extra information about the real world. AR is used in domains such as medicine, military, and entertainment. AR applications can use one or more feedback channels (modalities) to deliver information to the user. The most common feedback modalities in AR are visual, auditory and tactile because of the maturity of the technology to produce these kind of stimuli. We have high quality displays, or force feedback devices to mention a couple of examples. Taste and smell, on the other hand, represent an open research area mainly because of the challenges that synthesizing chemical particles represent. We refer the reader to Kortum (2008) for an extensive discussion of Human–Computer Interaction with non-traditional interfaces.

Omni-Directional Video (ODV) is an emerging media format that offers viewers a 360° panoramic video. To create an immersive experience, ODV is typically shown in a CAVE-like setup (Rovelo et al., 2014), or a personal display (e.g. a Head-Mounted Display) in combination with a tracking system to calculate the viewer's correct viewpoint (Bleumers et al., 2012).

The interaction with a 3DUI typically takes place in the mid-air space, and presents new challenges and opportunities for Human–Computer Interaction researchers. In this new type of applications, the virtual and real worlds can be mixed in one interface. For example, using AR it is possible to create a game where the player needs to chase virtual characters hiding in the real world (Sony Computer Entertainment, 2009). Thus, it is necessary to understand how to design applications that allow the users to naturally and efficiently perform the required tasks in the three-dimensional real space to interact with the virtual world. Bowman et al. (2008) highlights the importance of understanding how information can be efficiently delivered to the user taking into account the task domain, providing robust feedback by delivering information through different sensory channels simultaneously (*multimodal feedback*), and letting the user interact with these applications combining for example their voice, mid-air gestures and traditional input methods, such as a keyboard and mouse (*"multimodal input"*), as studied for example by Irawati et al. (2006a,b) or Lee and Billinghurst (2008).

Previous research has also studied how the human brain combines the information from our senses creating a robust and coherent interpretation of the environment, even when one or more senses might give noisy information (*e.g.* Blattner and Glinert, 1996; Ernst and Bülthoff, 2004; Ernst, 2006; Wozny et al., 2008). For example, imagine you are sitting on a static vehicle (e.g. a car or a train), in front of another vehicle. If the other vehicle starts moving while you are observing it through the window, your brain has to deal with an ambiguous situation: are you moving or is it the other vehicle? Your brain has to process the information received from the visual sensory channel and combine it with the vestibular system to disambiguate the situation and conclude that your vehicle is not moving (example adapted from Ernst and Bülthoff, 2004).

The work described in this thesis is focused on the evaluation of aspects of interaction in Augmented Reality and Omni-Directional Video applications. We evaluate how to convey information to the user in Augmented Reality applications exploiting human multisensory capabilities. Furthermore, we evaluate how the user can give commands to the system using more than one type of input modality, studying Omni-Directional Video gesture-based interaction. We describe the experiments we performed, outline the results for each particular scenario and discuss the general implications of our findings.

We focus on the user experience and the user performance when completing the tasks. User experience, when referring to a computer system, can be interpreted as the perception that the users have about the system after its use (see the international standard on ergonomics of human system interaction, ISO 9241-210 for a detailed description). Therefore, we assess how participants in our experiments perceive the feedback they receive and its benefits for the purpose of the task they performed. In that same regard, we consider user performance as the efficiency of participants to accomplish such tasks. We chose to characterize user performance in terms of the time to complete the tasks, but also in some cases, we consider the number of errors they make during the experiment.

1.1 Motivation

The increasing popularity of AR applications beyond research laboratories is due to several reasons: more affordable computer hardware—especially powerful mobile devices—more efficient tracking algorithms (we refer the reader to Van Krevelen and Poelman, 2010 for a detailed description of this topic) and the availability of multiple sensors, such as GPSs or digital compasses, integrated within many mobile devices. From the users' perspective, 3DUIs represent a new interaction paradigm where they can interact with computer systems—virtual objects—in the same way they interact with real life objects beyond the desktop and the computer's screen.

Gaming, marketing, searching assistants embedded in mobile AR browsers, medicine and military training simulators are some of the main application domains of AR. Layar (SPRX Mobile, 2010), Invizimals (Sony Computer Entertainment, 2009) and Wonderbook books of spells (Sony Computer Entertainment, 2012) are a few examples of commercial Augmented Reality applications.

Recent efforts such as Microsoft's Illumiroom (Jones et al., 2013) provide interesting possibilities for Omni-Directional Video, as they show how a living room environment could be turned into a small CAVE-like theatre. Benko and Wilson (2010a) show different scenarios in which ODV can be used. For example, they describe a portable dome setup in which users can interact with applications such as a 360° video conferencing system, a multi-user game or an astronomical data visualization system.

Some example of commercial Omni-Directional Video content available on the Internet are: the Reef sharks 360° experience from the BBC's "Oceans, exploring the secrets of the underwater world" series (BBC, 2014), or the AirPano project (AirPano project, 2014) with 360° video recordings of different tourist places around the world. These immersive experiences can be recorded with affordable off-the-shelf devices, such as the GoPro cameras mounted on special structures (e.g. Geerds, 2014).

However, even when capturing and rendering ODV have been widely investigated, little attention has been given to interaction with this type of content. Interaction with ODV includes triggering typical control operations we know from regular video (e.g. play, pause, fast forward and rewind), but also includes changing viewpoints by means of typical spatial interactions such as zooming and panning. These spatial interactions are, however, somewhat constrained, since spatial manipulations are always relative to the original camera position that was used while recording the ODV.

Our goal is to address this lack of formal user-centred experimentation in AR and ODV interaction. In this thesis we assess how visual, auditory and tactile feedback modalities and their combinations help users accomplish the goal of specific tasks. We designed experiments in three of the most important AR application areas: gaming, target finding and personal navigation systems. We developed one prototype for each area, and performed a number of experiments with users. Studying taste and smell modalities is out of the scope of this work because of the lack of robust, off-the-shelf devices to produce these two—complex—stimuli.

For ODV interaction, we carried out a gesture elicitation study, asking participants to come into an ODV CAVE. We also investigated the gesture variations and adaptations that users perform when they interact with the content on their own, and when they share the space with other users. Our goal was to find mid-air gestures and tried to identify the properties that result in a comfortable interaction for the users of ODV applications.

1.2 Objectives

The general objective of this research is to analyse the role of multimodal interfaces in the context of emerging 3D User Interfaces, such as Augmented Reality and Omni-Directional Video applications.

The following particular objectives define the scope of our research:

- Analyse how users complete a specific task using an AR application, considering both:
  - performance measures collected when a user completes the given task, and
  - subjective opinions about the user experience.

- Develop prototypes for studying user performance in three different scenarios:
  - Gaming.
  - Target acquisition.
  - Personal navigation devices.
- Expand the knowledge base relating to multimodal AR applications.
- Address the lack of formal user-centred studies in AR using a statistical analysis of the results obtained through several formal experiments.
- Present a user-defined gesture set for ODV interaction.
- Analyse the previous gestures when used in two different configurations: single and collocated settings.

1.3 Main Contributions

The main contributions of this thesis are summarised in the following paragraphs.

- **Expand the knowledge base on multimodal AR applications, from the perspective of user-centred design**. We perform a set of experiments with users in three different scenarios: desktop computer games, assisting technologies, and navigation technologies. The aim is to evaluate how the seven possible combinations of visual, auditory and tactile feedback channels can be used to efficiently convey task-related information to the user.
- **Present a quantitative and qualitative study for eliciting user-defined gestures for ODV**. We will also present an analysis and classification of these gestures and an analysis of the changes in gestures when used in two different configurations: single and collocated settings.

- **Summarize the findings of all the experiments to help improving the development of 3D User Interfaces**. As a result of the experiments we performed, we distill our experience, giving insight into the key aspects that should be taken into account when providing feedback through the visual, auditory and tactile senses, and implementing a gesture-based interaction as an input technique.

1.4 Thesis Outline

The content of this document is summarised as follows:

Chapter 2 gives an overview of the Augmented Reality and Omni-Directional Video fields. We present the definition and brief history of the fields discussing the different research approaches that define the state-of-the-art in both areas. The experienced reader can safely skip this chapter.

Chapter 3 describes the experiment we performed for the evaluation of user performance while completing a daily task: looking for a book in a bookshelf, with assistance of our AR prototype. We describe the outcome of the statistical analysis of the performance measures and the subjective opinions we collected during the experiment we carried out with students of the Universitàt Politècnica de València.

Chapter 4 describes the study of a multimodal interface for an Augmented Reality game. It describes the goal of the experiment, the task performed by the users, the apparatus we employed, the procedure we followed during the experiment and the results of the statistical analysis of the performance and user experience.

Chapter 5 describes the experiment we performed to test the impact of stereoscopy in user performance while playing the Augmented Wire Loop Game described in Chapter 4. We present the results of the analysis of the user performance and the user experience.

Chapter 6 presents the experiment to evaluate the impact of multimodal AR in our third scenario: personal navigation devices. We show the experimental setup and the outcome of the statistical analysis we carried out with the performance measures and subjective opinions we collected during the experiment.

Chapter 7 presents a gesture elicitation study in which we asked users to perform mid-air gestures that they consider to be appropriate for ODV interaction, both for individual as well as collocated settings. We describe the resulting gesture set and the variation in gestures we observed during the study.

Chapter 8 discusses the practical implications of the results we described in previous chapters. We also present the summary of our findings, which can contribute improving the development of 3D User Interfaces.

Chapter 9 summarizes the contributions of our work.

1.4.1 Overview of the Experiments

In order to provide the reader with a general overview of the work presented in this thesis, we describe the experiments according to the Design Space for Mixed Reality Systems, DeSMiR (Trevisan et al., 2004). This design space considers six axis to characterize Mixed Reality systems:

Transform type: refers to the relationship between actions and their effects, according to their occurrence in the real or the virtual world. It includes the following transformations: *Real action with real effect (RARE)*, *Real action virtual effect (RAVE)*, *Virtual action virtual effect (VAVE)*, *Virtual action real effect (VARE)*, *Real action shared effect (RASE)* and *Virtual action shared effect (VASE).*

Connection type: describes the relationship between real and virtual objects. Depending on the moment when the link between them is defined the connection can be: *Static* if the designer defines the link, *Dynamic* if the user can link real and virtual objects while using the applications, or *Mixed.*

Insertion context: defines the interaction space according to its position with respect to the user: *central zone* (from 0 to 45 cm), *personal zone* (from 46 cm to 1.2 m), *social zone* (from 1.3 m to 3.6 m), and *public zone* (greater than 3.6 m).

Media: describes the level of complexity and dimensionality of the presented data: *text (1D)*, *images (2D)* or *3D animations*.

Interaction focus: describes whether the user's focus of interaction should be on the *real world without shared attention*, *virtual world without shared attention*, *real world with shared attention*, *virtual world with shared attention* or *shared between worlds*.

Kind of augmentation: based on the purpose of the virtual content augmenting the real world. It defines three types: *interaction* (try to make computers as transparent as possible), *user's actions* (try to increase the number and/or quality of the tasks that the user performs), and *user's perception* (try to augment the user's perception with new information).

By considering this design space, we can characterize the prototypes we built to achieve the objectives of this thesis as follows. Table 1.1 and Figure 1.1 depict the classification of each of the systems we developed in terms of the same design space. It is worth noticing that, for clarity of the representation, we only show the type *other* in the Media axis as our systems include different types of media presentation.

Searching assistant: this system guides a user searching a book in a bookshelf. Users have to move their hand in the 2D plane defined by the bookshelf, in their *personal zone*. *Sharing their attention* between a wand to receive the guidance and the bookshelf to read the books' titles. The system provides feedback (guidance) to indicate when the user's hand is getting closer to the desired book. Feedback given by the system (the *media*) is in the form of vibration patterns, sounds, and LEDs that are lit on and off as needed. In this sense, the system belongs to the category *RASE* of the transform type. Users' actions change the virtual world (the state of the system) but its effect is only visible on the real world. The link between real and virtual world can be considered *mixed* because the relationship between real objects (the wand acting as the interface) and the virtual world (the system state and corresponding feedback) is static, defined at the design step of the system. At the same time, the user can ask for assistance for searching different books at run time, thus, the states of the system are linked dynamically according to user's actions. The goal is to increase the efficiency of the user finding books (*action* type of augmentation).

AR Wire Loop Game: This system is the AR version of the Wire Loop Game. Users move a wand to control a virtual ring in the three-dimensional space within their *central zone*. *They share their attention between the two worlds*, the virtual and the real. They have to learn how to move their hands to control the ring, and at the same time they observe the virtual elements on the screen. The system provides feedback to alert the user when virtual objects have collided. The *media* to accomplish this are visual elements (changes in colour of virtual objects), sounds, and vibrations. This game can be included in the *RASE* category, as the movement of the wand in the real world determine the movement of the virtual ring. When objects collide, the wand vibrates and the game plays a sound. The link between real and virtual objects, as in the previous case is *mixed*, because the virtual loop is attached to the wand all the time, but the marker on the table can be linked to the different levels of the game. The goal of the feedback is to improve the way users perceive alert cues that inform when the virtual objects have collided (*perception* type of augmentation).

Pedestrian Navigation Assistant: this system guides pedestrians in the same fashion as a traditional GPS guides car drivers. Users need to move their smartphone in their *central zone* to find the direction to follow. Users need to *share their attention between the real and virtual worlds* to observe the visual navigational cues that are shown through the smartphone's screen, but also to be aware of their surroundings while walking. The system uses 3D animations of arrows and paths to visually indicate the direction to follow, combined with spoken directions and vibration patterns (the *media*). This is also a system where the user's actions have a shared effect between worlds (*RASE*). When the user walks and changes the position and orientation of the smartphone, the state of the system changes, visual feedback is displayed as needed (arrows for turns or the path to continue straight ahead), and at the same time, the auditory and tactile feedback is given through the smartphone. The link between the virtual and the real world is *dynamic*, as the user's actions while navigating define the virtual elements (feedback) that will be displayed on each one of the control points (geographic markers) depending on the chosen route. The goal of the information provided by the system is to reduce the time required to navigate from point A to point B (*action* type of augmentation).

Table 1.1: Overview of the experiments.

		AR Application		
		Searching Assistant	Augmented Wire Loop Game	Navigation Assistant
Transform type	RARE			
	RAVE			
	VAVE			
	VARE			
	RASE	✓	✓	✓
	VASE			
Connection type	Static			
	Mixed	✓	✓	
	Dynamic			✓
Insertion context	Central Zone		✓	✓
	Personal Zone	✓		
	Social Zone			
	Public Zone			
Media	Text		✓	✓
	Graphic		✓	✓
	Image			
	Video			
	3D Animation			✓
	Other	✓	✓	✓
Interaction focus	RW			
	Shared RW	✓		
	Shared Worlds		✓	✓
	Shared VW			
	VW			
Kind of augmentation	Action	✓		✓
	Perception		✓	
	Interaction			

The Appendix A provides a quick reference to the description of every experimental chapter in terms of the Goal/Question/Metric method Basili et al. (1994).

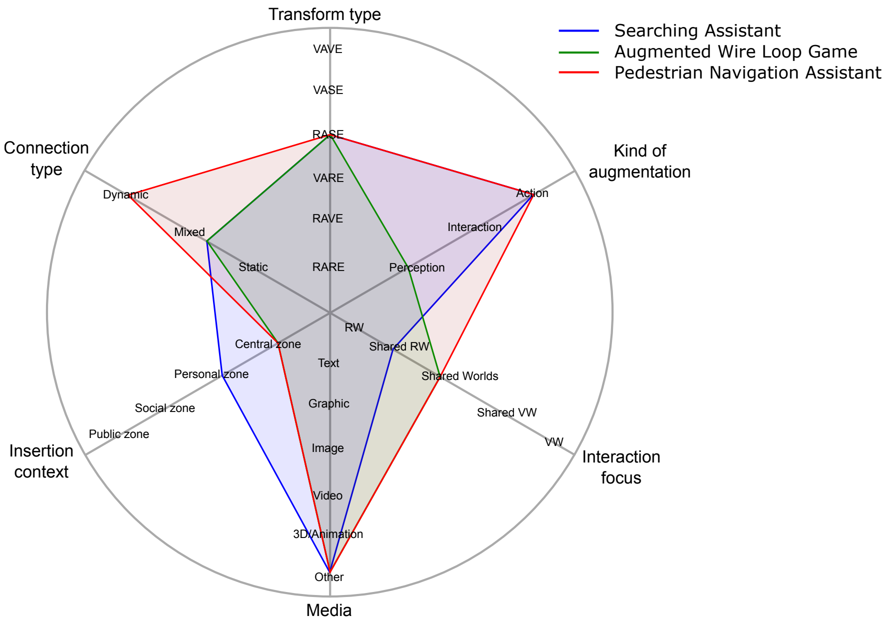


Figure 1.1: The experiments described in this thesis in the context of the Design Space for Mixed Reality Systems.

Considerations about the tasks

The tasks we evaluated to study multimodal feedback in Augmented Reality belong to three of the main application domains. We employed the main display technology used in AR (computer screens, smartphone displays and head-mounted-displays). We also considered three common tasks in 3DUIs: target finding, manipulation of virtual objects and pedestrian navigation. We evaluated tasks that require interaction in mid-air space, however, the task we describe in Chapter 3 is performed in a bi-dimensional space (the bookshelf front). We only evaluated one outdoor task for navigation using the smartphone screen, however, it would be interesting to further investigate the impact of wearable devices, such as Google glass-like displays on this type of applications.
Table 1.2 presents the summary of the tasks that we evaluated in each experiment. More details are given in the following chapters.

Regarding the study about mid-air gesture-based interaction with Omni-Directional Video, we decided to perform a gesture elicitation study, because we identified the lack of a more suited interaction method for time and space

Table 1.2: Summary of the tasks.

Experiment	Application domain	Interaction	Cognitive resource	Task space	Display technology
Searching assistant (Chapter 3)	Target finding	3D space	Visual Spatial orientation	Indoor	Hand-held device
Augmented Wire Loop Game (Chapter 4)	Gaming	3D Space	Visual Motor skills	Indoor (desktop)	Computer screen
3D Augmented Wire Loop Game (Chapter 5)	Gaming	3D Space	Visual Motor skills	Indoor (desktop)	Computer screen HMD
Pedestrian Navigation Assistant (Chapter 6)	Outdoor navigation	3D Space	Visual Motor skills Spatial orientation	Outdoor	Smartphone screen

control actions for this new media format. However, participants of the study did not have to accomplish any particular task besides the proper gesture elicitation. We did not go further than the definition and characterization of the gesture set and certainly, more studies are needed to validate its applicability.

1.5 Terminology

In this section we present the terminology and the conventions we follow throughout this document.

CAVE Abbreviation for Computer assisted virtual environment. It is an immersive virtual reality environment where projectors shows the computer generated content on the inside walls of a room-sized cube. More recent versions can use curved or spherical projection surfaces.

Mid-air space A region in the air not close to a surface. For example, that is not close to either a table or the ground.

Modality When referring to feedback types, it is any sense through which the user can receive the output of the computer (for example, visual, auditory or tactile modalities). When referring to input techniques, it is any sensor or device through which the computer can receive the input from the user, e.g. the mouse and keyboard or voice and mid-air gesture recognition systems.

Multimodal Refers to any combination of input or output modalities.

Pan In video display technology, panning refers to the horizontal scrolling of an image that is wider than the display.

Walk-up-and-use A property of self-explanatory computer systems that refers to the fact that can be used by users without previous training or experience with it.

Between subjects design An experimental design where participants are part of only one experimental condition. In other words, can test only one of the possible levels of the independent variable.

Within subjects design An experimental design where participants take part in all experiment conditions. In other words, one participant can be in more than one group, testing all the levels of the independent variable.

d.f. When describing the results of any statistical analysis, it refers to the degrees of freedom of the test. These are the number of independent observations in a sample minus the number of population parameters that must be estimated from sample data.

F Is the outcome of an F-test, to assess the equality (or homogeneity) of two variances. We used it when describing the results of an ANOVA test.

p When describing the results of a statistical analysis, the p-value represents the estimated probability of rejecting the null hypothesis (H_0) of a study question when that hypothesis is true.

CHAPTER 2

INTERACTION WITH 3D USER INTERFACES

In Human-Computer Interaction, 3D User Interfaces (3DUIs) are those that involve interaction directly in the 3D spatial context of the user, without using indirect metaphors like 2D widgets, choosing items from a menu or entering coordinates through a keyboard (Bowman et al., 2004). These systems use different tracking technologies to calculate the user's point of view of the scene and to capture user actions, for example, to control virtual characters or to navigate through a virtual world. On the other hand, feedback is typically presented using the visual, auditory or tactile sensory channels, depending on the context of the application.

Continuous innovation in computer hardware gives developers and researchers the opportunity to create more powerful 3DUIs, combining more than one input and/or output modality, in other words, developing multimodal interfaces. Portable devices have now multiple tracking sensors and enough computational power to execute complex simulations like those found in popular video games. For example, using the accelerometer in a tablet or smartphone, a GameLoft video game called Tom Clancy's Rainbow Six: Shadow Vanguard (GameLoft, 2013) lets the user look in different directions inside the game, moving the mobile device around. Another example is the development of more efficient tracking algorithms, linked in some cases to new portable devices such as Google's Project Tango (Google Inc., 2014), a 3D sensor similar to the Microsoft's Kinect (Microsoft Inc., 2013) for smartphones.

The goal of multimodal feedback in user interfaces is to assist users in completing tasks quicker and/or with fewer errors, by reducing their cognitive workload. This requires studying how visual, auditory and tactile feedback modalities complement each other helping users to accomplish their task efficiently.

The effect of combining different modalities on user performance has been previously studied in Psychology. Many of these studies focus on measuring how visual, auditory and tactile modalities can enhance the response of test subjects after receiving the stimuli, ie. reducing the reaction time or increasing the detection rate. Three examples of these studies are the work of Diederich and Colonius (2004), Rach and Diederich (2006) and Rach et al. (2011). They point out that multimodal stimuli improve user performance (faster reaction times) compared to unimodal feedback. Rach and Diederich (2006) evaluated the effect of visual and tactile stimuli duration on the reaction time, showing that short duration stimuli produce better performance. Rach et al. (2011) presented a study comparing the effect of visual and auditory stimuli. Their results show that combining both modalities produce faster reaction times, specially when the sources of the stimuli are in close proximity (temporally and spatially). Diederich and Colonius (2004) compared the effect of visual, auditory and tactile stimuli finding similar results: reaction times were shorter in multimodal conditions; being the shortest when the three modalities were combined. In this thesis we focus on how these findings can be applied to the design of multimodal 3DUIs.

According to the Multiple Resource Theory (MRT) by Wickens (2002), a computer system that presents information using multiple sensory channels simultaneously is more effective than a system that uses only one modality. This is especially true when performing high workload tasks. As we use different senses to perceive complementary information, our brain uses multiple cognitive resources to process all the information, thereby dividing the cognitive workload and increasing user performance.

On the other hand, according to Wickens' MRT and results presented by other research works (Wickens et al., 2002; Maltz and Shinar, 2007), providing the user with extra information in an already saturated sensory channel is counterproductive.

Multimodal input lets users interact with the application employing different modalities, such as voice, hand gestures, keyboard or mouse, giving alternative methods to accomplish the same goal. However, as explained by Bowman et al. (2008), not all the challenges of combining multiple input techniques have been solved. It is still necessary to understand the benefits of, for example, the use of mid-air gesture interaction compared to traditional keyboard and mouse input in 3DUIs, such as virtual environments.

Understanding the role of multimodal input and output for 3DUIs is still an open research area. This chapter gives an overview of the two fields we cover in the present work: Augmented Reality (AR) and Omni-Directional Video (ODV). We start by defining an AR system. Then we discuss the state-of-the-art in the field that is relevant to our investigation. Finally, we define ODV and its different properties and challenges and describe the relevant research that has been performed in the field.

2.1 Augmented Reality

Virtual environments (VE) immerse users in a completely synthetic scene. These systems supersede the users' perception of reality in those senses that are involved in the VE. Milgram and Kishino (1994) classify these systems according to the users' immersion degree. The classification goes from systems that show to the users only the real environment, to systems that show to the users only a virtual environment. The space between these two ends is known as *"Mixed Reality"* (see Figure 2.1).

Augmented Reality is one type of Mixed Reality applications. It combines techniques from computer vision, pattern recognition and computer graphics to overlay virtual objects directly onto users' view of reality. The main objective is to enhance users' perception providing more information, helping them to perform their job better or faster. More recently, with the widespread usage of powerful smarthphones, AR has also been applied to increase the appeal of entertainment applications.

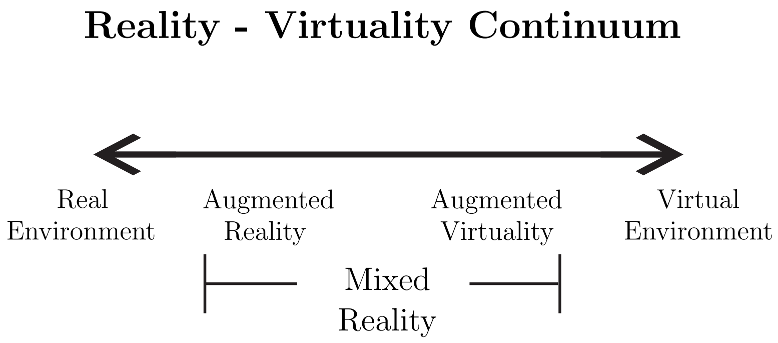


Figure 2.1: A classification of VE according to the user's immersion degree.

According to Azuma (1997), an AR system must fulfil these three features:

- Mix real and virtual content.
- Allow users real time interaction with virtual content.
- The registration process of virtual content must be in the 3D space.

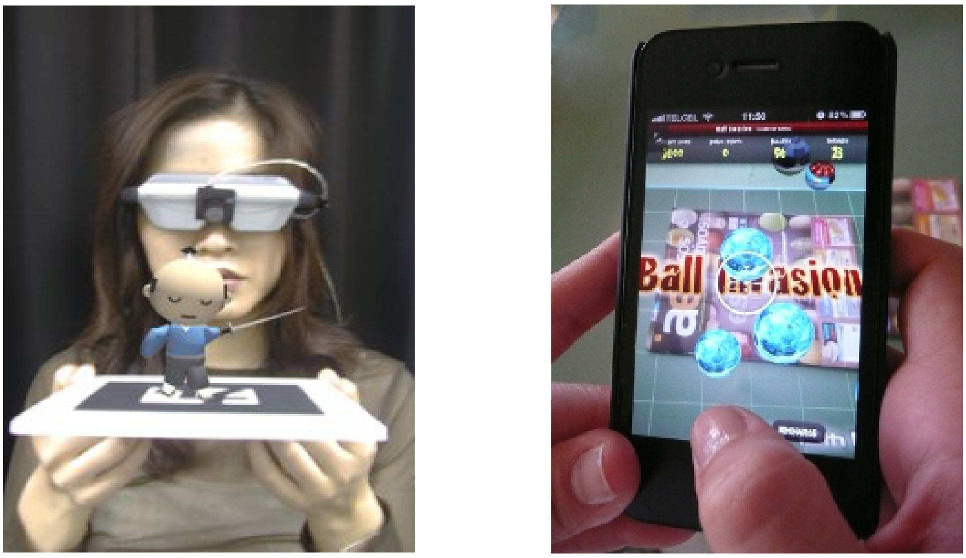
AR applications typically use markers to calculate the point of view from which the user observes the virtual objects as captured from a camera. These markers were traditionally black squares with different patterns in the centre (see Figure 2.2a[1]). However, the latest AR libraries allow developers to use more complex images as markers (see Figure 2.2b[2]), giving more flexibility to the applications.

Other AR research applications use special tracking systems, such as magnetic trackers or infrared (IR) tracking systems for following the user's point of view and the position and orientation of the interaction devices. Studierstube (Fuhrmann et al., 1997) is a good example of this type of AR applications. It is a collaborative environment where multiple users observe and interact with three-dimensional scientific visualizations, wearing head-mounted-displays (HMD) and using different types of input devices.

Mobile platforms, either smartphones or tablet PCs, have also become an important springboard for mobile AR. Mixare (Wikitude GmbH, 2011) and Layar (SPRX Mobile, 2010) are two examples of AR browsers where virtual content

[1] Image from `http://www.hitl.washington.edu/artoolkit/`

[2] PointCloud demonstrator 13th Lab AB (2012)



(a) Traditional AR marker (b) Complex image used as a marker

Figure 2.2: Two examples of the markers employed in AR applications.

is presented blended in the smartphone's camera view. In these applications, the user can vary the amount of information presented on the screen, for example, selecting a different channel or layer to filter information out, or using the proximity to the user as a filter to decide what to show and what needs to be hidden. This is a similar approach to the one described by Looser et al. (2007), where they adapted the concept of a Magic Lense (Bier et al., 1993) to a different desktop AR application.

Visual, auditory and tactile feedback modalities are the most common channels used to present virtual content to the user in AR applications. Visual feedback is typically shown using a computer screen, a HMD or a smartphone. Auditory feedback is provided through speakers or headphones. Finally, tactile feedback is commonly provided in the form of vibrotactile or force feedback, using a wide variety of devices, such as the Falcon haptic force feedback device (Novint Technologies Inc., 2012), or different vibration devices, e.g. the vibration features of a mobile phone or the Nintendo Wii remote.

However, most of the research on AR is focused on improving the visual quality of the applications (e.g. better shadow rendering or improving depth perception) and on how to use different input modalities to let the users interact with the augmented scene (for example, using hand gesture recognition, 3D tracking devices, fiducial markers, speech recognition, among others).

Dünser et al. (2008) and Dünser and Billinghurst (2011), discuss the challenges and efforts made on the evaluation of AR applications. They describe the difficulties of evaluating such systems due to several causes, including the lack of an evaluation framework for AR systems, the different senses involved (sight, hearing, touch, etc.), and the variety of input and output devices, among others.

Billinghurst (2008) classifies AR user studies in three categories: perceptual, interaction and collaboration studies. The first category describes how users perceive virtual objects in the real scene. The second category compares different methods to help the user to interact with virtual objects. The third category studies the characteristics of effective AR tools that enhance user collaboration to accomplish one common task. Billinghurst also emphasized the need of more formal user evaluation, to find new AR *"interface metaphors"* for the *"unique relationship between the real and virtual worlds"*.

2.1.1 Comparing Visual and Tactile Feedback

Following the classification provided by Billinghurst, we are interested in perceptual and interaction studies; specially those that assess using alternative feedback modalities in AR applications. One example of such a study was presented by Ahmaniemi and Lantz (2009). They describe a mobile AR application that employs tactile feedback to lead users on a target finding task in an open space. The device vibrates when the user points it toward the objective, and changes in frequency of vibrations give a hint on the distance to the target. The results suggest that using tactile cues in AR applications helps guiding the user in open spaces. However, the authors also highlight the importance of a good feedback design, in this case, because some targets can be missed if they are too close to the user and the feedback frequency is too low.

Oron-Gilad et al. (2007) studied the impact of tactile feedback on the reaction time, also in a target finding task. They compare user performance while receiving vibrotactile or visual cues to indicate the direction of the initial movement to find the target. They also assess if vibrotactile cues can be used to inform the distance to the target efficiently. They report that no matter the location of vibration devices on the hand of the user (either in the palm or on the dorsum) they can indicate the direction of the initial movement successfully and more rapidly than using only visual cues. According to their results, using both modalities together reduces reaction times. In addition they report that tactile feedback can be used to successfully indicate distance to target. In spite of their results, Oron-Gilad et al. (2007) remark that it is still necessary to evaluate different vibration frequencies to indicate when the user has reached the target.

Lindeman et al. (2003) also studied how visual and tactile feedback channels complement each other in a searching task. Different visual signals and vibrations patterns were used to spatially indicate the area where the target was (a letter in one of three sets of letters organized in columns). Each subject received visual and tactile feedback for one second at the beginning of each trial. Vibrators were located on the back of a swivel chair in a 3×3 matrix, one column per set of letters. The results show that visual feedback outperforms tactile feedback, however when compared with no feedback at all, tactile cues improve the performance. Therefore, tactile feedback provides a potential replacement of visual feedback in those situations where the latter is not an option.

In a latter work, Lindeman et al. (2005) described their work about vibrotactile cues to guide a user inside of a building. This time, users wore a belt with 8 tactors (vibrators) that indicated the direction to follow (each tactor on every cardinal point). The idea was to reduce the time exposed to unexplored areas that can represent a threat in a military context. Their results showed again that using vibrotactile cues represents a significant improvement on user performance compared with no feedback at all: in their experiment users covered more space in less time.

Tactile feedback has also been used in driving assistance systems (Van Erp and Van Veen, 2004). Vibrotactile and visual feedback modalities were used to indicate distance and direction of the next turn. Drivers performed a set of circuits with one or both feedback modalities. The results of their experiment showed that multimodal feedback produced the fastest reaction times and that visual feedback alone produced poorer results than tactile feedback, as the latter reduces drivers' mental effort and workload.

Unger et al. (2002) presented a study that combines the visual and tactile feedback modalities. Users were asked to insert a squared peg in a hole using a haptic device, while seeing the scene in a computer monitor. The authors compared three scenarios, depending on the type of tactile feedback provided: no tactile feedback, force feedback provided by the haptic device, and "real feedback", by adding a real peg and a real hole to the haptic device. Users performed best completing the task when using the real peg and hole, while force feedback produced better results than visual feedback only. Unger et al. (2002) state that force feedback, in which the constraints are enforced as in the real word, enhances the realism of the application, but they remark that it is still necessary to assess to what extent this occurs.

Prewett et al. (2006) and Prewett et al. (2012) present two meta-analyses comparing visual and tactile feedback. They found that combining both feedback modalities reduces reaction times and improves user performance in high workload conditions, i.e. when multiple tasks are being performed. Their findings also show that combining visual and tactile feedback does not affect error rates, and that visual and tactile—bimodal—feedback have no positive effect on communication tasks. The authors state that combining these modalities produce the best results when providing alerting and orienting cues.

2.1.2 Comparing Visual and Auditory Feedback

Previous research has also studied the effect of sound on depth perception in AR environments. Zhou et al. (2004) assessed how auditory feedback complements the visual channel improving depth perception to locate characters in an AR computer game, thus reducing searching time. During the experiment, two participants tried to complete the game at the same time. Visual and auditory feedback modalities were used as cues to provide information to each player about the location of both the other player and the virtual objects. Their results showed that using spatial sound in this kind of applications enhance user depth perception of virtual objects, helping to complete the task faster and enhancing the collaboration between users.

Loeliger and Stockman (2013) presented another study on auditory feedback. Their system uses an audio map to help users navigate through a 3D virtual city. The application was distributed through a web page, and users with different degrees of vision problems tested it at home. Their results showed that using the audio map helped participants increase their spatial knowledge and their way finding performance during the test.

Pierno et al. (2005) presented another study that compares visual and auditory modalities, alone and combined, in a target acquisition task with different levels of visual workload. A high workload condition was simulated asking users to perform a second visual task during the target acquisition: counting the times a number between 4 and 9 appeared on the HMD. The studied feedback modalities indicated the direction in which users had to turn in order to locate the target. Their results showed that user performance improves when extra help is provided, no matter the type of cues (visual, auditory or both). They reported that visual feedback outperforms auditory feedback, even when users were performing both tasks at the same time in the high workload condition. The explanation is that counting the number of times a given number appears uses the same brain structures than listening to the location feedback. Using both visual and auditory modalities combined produced better results than using the auditory modality alone. As explained by Pierno et al. (2005), the visual channel helps in this case to overcome possible interferences between cognitive resources.

Visual and auditory feedback modalities were also studied in a collision avoidance system in the work by Maltz and Shinar (2007). Both feedback modalities were tested (individually and combined) to assess how they could help drivers to keep a safe distance from the car ahead. They found that using alerting messages improves user performance while driving. Auditory alerts obtained better results than visual and visual and auditory combined, although, users found combined alerts to be more helpful.

2.1.3 Comparing Visual, Auditory and Tactile Feedback

Visual, auditory and tactile feedback modalities have also been studied together, for example, to assess how these modalities can be used to represent kinesthetic properties of the everyday life objects (Herbst, 2005). Participants of the study had to evaluate the weight and the force necessary to move different cubes, putting them in order, according to their weight and resistance to movement. Changes in the colour of the cubes, in vibration intensity and sound tones represented both properties. Regarding the comparison among modalities, tactile feedback alone was rated less useful than auditory modality. However, Herbst points out that special attention must be given when selecting the tone to avoid confusing the user. Users became uncomfortable to a certain degree with vibrations, as these were felt on the top of the hand and not on the fingertips as one would expect. Herbst concluded that any of the three modalities, separated or combined, is enough to represent the evaluated physical properties. Using more than one modality decreased task completion times and reduced the number of movements while ordering the cubes.

Sun et al. (2010) studied how visual, tactile and auditory modalities (and their combinations) affect user performance in a steering task. Users were asked to follow a track with a stylus in the screen of a tablet computer. The system used different feedback channels to alert the user when the stylus was leaving the track. They found that tactile feedback made the best improvement in users' accuracy. However, the completion time of the task did not depend on the type of stimuli used. Users rated the combination of auditory and visual cues as the best option. They also pointed out that the vibrations on the stylus made them do unintended errors.

Smith et al. (2009) explored the differences between visual, tactile and auditory stimuli to help users in a task-switching situation. Visual and auditory were the preferred methods to let users know that they have to direct their attention to the other task. The problem is that these cues may go unnoticed in a visually or sound saturated ambient. Their results showed that tactile feedback is not only able to alert the user about a new task, but it can also help the user in locating the task. The tactile-orienting cues performed better than the tactile-alerting cues, but there were no differences with respect to auditory cues.

Bresciani et al. (2008) studied how background signals in different sensory channels distract users from the activity they are performing. They reported that the visual modality is the most affected by auditory and tactile distractors, and conversely is the one that distracts users' attention the least. The auditory modality was the least susceptible to visual and tactile distractors and the modality that affected the perception of the other two modalities the most. They also found that a background bimodal signal distracted users' attention more than a single modal signal.

Charoenchaimonkon et al. (2010) studied different modalities to assist users in a target finding task. They evaluated the speed and accuracy of users while receiving visual, auditory and tactile feedback. They found that user performance improves more when receiving tactile feedback than when receiving auditory feedback, and this effect is more noticeable when the difficulty of the task increases. They also found that auditory and tactile channels alone work better than using visual feedback in this eminently visual task of target acquisition.

El-Shimy et al. (2009) propose using auditory and tactile modalities in addition to visual feedback for picking targets in a 3D scene rendered on a regular display. The added modalities are used to provide depth information. The authors evaluated two different ways of providing feedback: discrete feedback, which indicates whether or not the target has been acquired, and continuous feedback, that provides information on target distance. They conclude that discrete feedback improves accuracy of target acquisition. In addition, they observed that tactile feedback reduced both reaction time and task completion time but at the expense of a decrease in success rate.

Burke et al. (2006) present a meta-analysis comparing visual and auditory feedback with visual and tactile feedback. Their results showed that adding either aural or tactile feedback to visual feedback improved reaction time and user performance, but did not decrease error rate. Furthermore, is seems that tasks under high workload conditions do not benefit from additional aural feedback, since it appears to increase the workload and decrease effectiveness. The meta-analysis revealed that adding tactile feedback improved user performance under the same conditions.

2.2 Omni-Directional Video

Omni-Directional Video is an emerging media format designed to offer viewers a panoramic video-recording where they can observe content in all the 360° around them. This type of video is recorded with a special type of cameras, either using a curved mirror, or a special arrangement of cameras, such as the one shown in Figure 2.3.

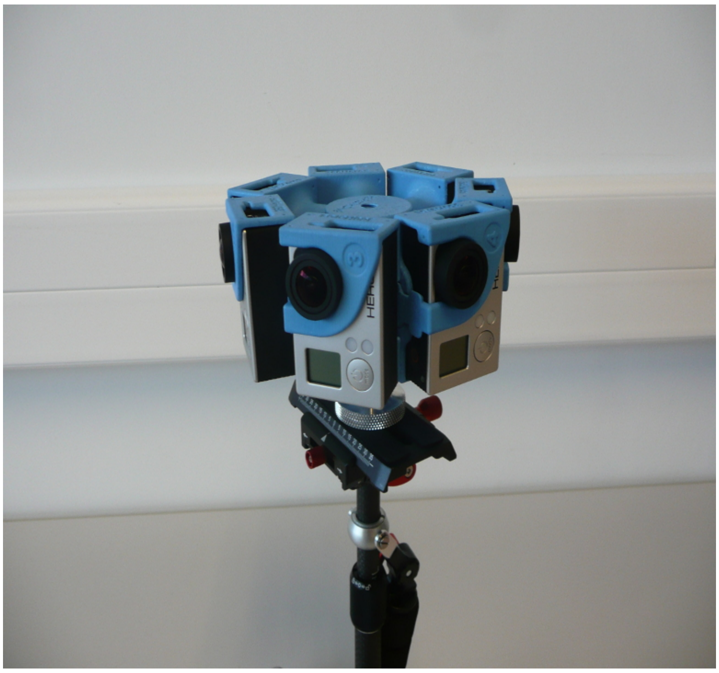


Figure 2.3: ODV Camera composed of 7 Go Pro Hero3

The goal of ODV is to immerse the viewer inside the pre-recorded experience, that shows content such as the frame depicted in Figure 2.4, typically using a CAVE-like setup (see Figure 7.1), or a head-mounted display, in combination with a tracking system to calculate the viewer's correct viewpoint.

ODV has different application domains where the viewer can benefit from the immersive experience. Some examples are video conferencing, interactive video recordings where the viewer can participate in games, such as a treasure hunt, or exploring a virtual tour in a museum. ODV can also extend the Google Street View metaphor, but instead of presenting static 360° frames, showing to the viewers a pre-recorded video path in which they can explore a remote location.

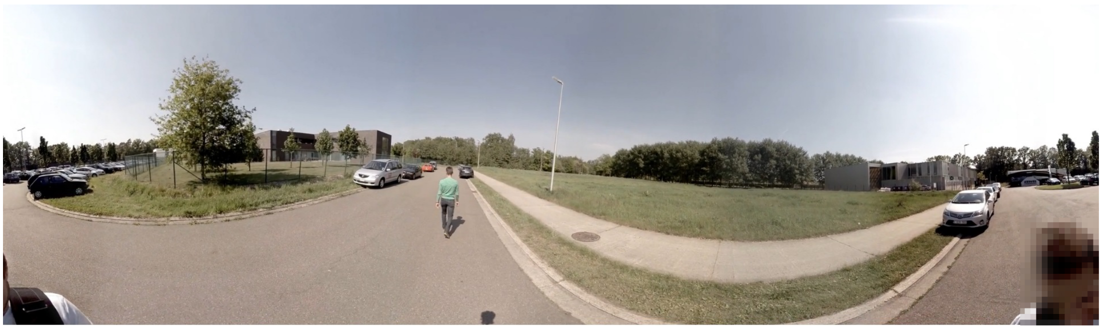


Figure 2.4: Example of a 360° video frame.

Because of the spatial properties of ODV, the interaction with this new media type can be classified into two groups: actions to control time, and actions to control space. Time-based commands include play, pause, skip a scene, rewind and fast forward. Space-based interactions include panning and zooming the video.

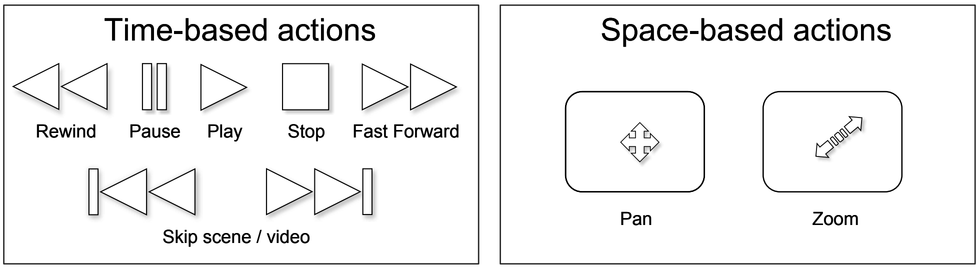


Figure 2.5: ODV Interaction can be classified in time and space based actions.

2.2.1 ODV Interaction

ODV CAVE setups can show the content to a single user or multiple, collocated users. Whereas within the single user scenario the interaction can be handled in a straightforward way, when multiple users are present, occlusion problems will likely occur, specially when there are more than three users. Users will partially block the view from each other, making the interaction with the content harder to perform.

ODV can be shown on public displays, for example as part of a museum exhibition. These setups require that ODV systems allow "walk-up-and-use" interactions, where non-trained users are able to interact with the content easily. In this context, informing the user about how to interact with the system is vital: providing feedback and feedforward (Vermeulen et al., 2013) about the actions being performed by themselves and by other users.

Another interesting aspect that deserves attention regarding ODV interaction is how to avoid the King Midas golden touch—according to the Greek mythology, the King Midas could transform everything he touched into gold—when designing a gesture recognition system. This problem occurs when the system is continuously scanning the environment searching for a performed gesture, hence, every movement in the tracking volume can be mistakenly interpreted as interaction with the system. It is important to define the appropriate mechanism to enable system's listening mode because human behaviour and human-to-human interaction usually involves hand movements that can cause undesired actions in the system.

Parameterization of the actions is another open research area in ODV, defining interaction techniques that are capable of adjusting the command's effect. It is important to define usable interaction techniques that allow transparent and easy control over the zooming level or the fast forward speed.

ODV has large dimensions, both, regarding the space required to store the files and the visualization of such content. This situation opens more research questions, e.g. how to render such content in an appropriate format to provide a good experience to the user, helping them to acquire a complete mental model of the available content and interaction possibilities. Large dimension of the ODV files represent also a technical challenge for data transmission over a network in order to achieve a good quality of experience. However, the discussion of such techniques is out of the scope of this document.

Bleumers et al. (2012) describe user's expectations about ODV, highlighting the uncertainty among users about how to interact with ODV. They put forward mid-air gestural interfaces as a possible solution , although they did not explore such interfaces in their work. Mid-air gesturing has been used since the early nineties for controlling television sets (Baudel and Beaudouin-Lafon, 1993; Freeman and Weissman, 1995), and nowadays television sets with a built-in camera and simple gestural interface are commercially available, such as Samsung Smart

TV (Samsumg Corporation, 2013).

Regarding the study of user-defined gestures and users' preferences, researchers often focus on finding the best set of gestures for specific tasks, such as grabbing and rearranging a set of objects (Hespanhol et al., 2012), or pan-and-zoom operations (Nancel et al., 2011). Others analyse the gestures for a very specific action such as rotation (Hoggan et al., 2013), evaluate users' behaviour when interacting with zoomable video (Axel et al., 2010), study how users rate the appropriateness of the gestures they observe (Fikkert et al., 2010), or compare the acceptance level of different gesture sets, i.e. one set created by HCI experts and another by "inexperienced" users (Morris et al., 2010).

To generate a set of user-defined gestures, Nielsen et al. (2004) and Wobbrock et al. (2009) proposed similar elicitation approaches: define what operations have to be executed through gestures, ask participants to perform gestures for those operations, and finally extract the gesture set from the collected data. Wobbrock et al. also use Likert scales to gather qualitative feedback, while Nielsen et al. benchmark the set in a second round of trials. The methodology of Nielsen et al. has for instance been applied to find gestures that can be used to interact with music players (Henze et al., 2010). Another approach is proposed by Grandhi et al. (2011), who asked a group of users to describe and mimic different daily tasks that can be extrapolated to human-computer interactions.

Also of potential interest when looking into user-defined gestures is how users give and receive instructions to and from other users using only hand gestures. Aigner et al. (2012) studied how users create their own gesture sets to successfully communicate instructions to other participants through a video chat.

Benko (2009) and Benko and Wilson (2010b) described the challenges of interactive curved and spherical displays, which we believe to be representative for CAVE-like ODV setups. These challenges include developing walk-up-and-use interaction techniques, creating a transparent environment where users can interact with the appropriate device for each task, and devising compelling applications for this type of device. Our focus lies on the challenge of designing appropriate interaction in the context of ODV.

Researchers already investigated several aspects of ODV interaction. Macq et al. (2011) implemented ODV navigation using the camera of a tablet PC as the orientation tracker and the screen as the display device (i.e. a peephole display). Neng and Chambel (2010), on the other hand, described the use of 360° "hypervideos", which provide extra information through embedded navigational links. These videos are watched over the Internet, on a regular computer screen.

Zoric et al. (2013) presented a user study in which they observed pairs of participants interacting with high definition panoramic TV through gestures. Their observations suggested that the design of multi-user gesture systems should allow for socially adapted gestures for controlling and navigating video content. However, Zoric et al. (2013) considered this study to be merely a first step in exploring how users interact with such content using a gesture-based system. We investigate this topic in-depth.

2.3 Chapter Summary

The goal of multimodal interfaces is to help users accomplishing their task efficiently and effectively. Using for example visual, auditory and tactile feedback modalities to provide information to the user, and different input modalities such as speech recognition, gesture-based interaction or traditional keyboard and mouse.

This chapter summarizes the most relevant research in the fields of Augmented Reality and Omni-Directional Video related to the work described in this thesis.

Having provided to the reader the appropriate context to understand the work presented in this thesis, we now proceed to describe each one of the experiments we performed.

CHAPTER 3

A MULTIMODAL SEARCHING ASSISTANT

The advantage of multimodal feedback, in a computer system, to assist people in daily tasks has been studied before in different professional and non-professional domains. For example, for designing driving assistants (Van Erp and Van Veen, 2004), for improving the way pilots receive vital in-flight information (Wickens, 2002), or for helping soldiers covering large territories on foot, in unknown places, during reconnaissance tasks (Lindeman et al., 2005).

Systems that use different feedback channels at the same time improve users' performance when multitasking, especially in highly demanding tasks. However, when the same sensory channels or cognitive resources are used in different simultaneous tasks, user performance decreases (we refer the reader to Wickens, 2002, for a detailed explanation about this topic). A good example of this problem is the action of setting a new destination on a GPS device. This task requires exclusive attention of the visual sense. Whereas this could be done while driving, user performance decreases and the probability of an accident increases. On the other hand, providing directional information using the auditory channel, a GPS can guide a driver without distracting her visual attention from the road.

The experiment described in this chapter focuses on a multimodal user interface to assist humans in everyday target acquisition task. We use visual, tactile, and auditory feedback (and their combinations) to inform the users about the location of a book in a unorganized bookshelf. This is a predominantly visual task, in which the user is focused on reading the books' titles while searching for a given book.

This experiment is the starting point of our research on the interaction aspects of multimodal 3DUIs.

Key contribution - This experiment lets us study the effect of multimodal feedback in 3DUIs when the user is required to perform a basic task. These findings can be extrapolated to other applications with similar purposes. Some examples of these applications are: assistive technologies for visually or hearing impaired persons, pedestrian navigation assistants, or complex machinery maintenance assistance.

3.1 Experiment Description

Searching for a specific object in a two-dimensional arrangement of objects is a common everyday task. Looking for the milk jar in the refrigerator's door or locating our preferred juice brand in the supermarket's shelves are examples of this task. The difficulty (cognitive load) of the task depends on the number of shelves, the number of items to choose from, how different they look, if they are organized in a predetermined, known pattern, etc. We are used to perform this searching task in our daily lives, however, it is especially challenging in unknown environments, even when the items are ordered following some code-based distribution.

Finding a book in a bookshelf is a task that every university student does. Thus, the task difficulty does not represent a confounding factor when we evaluate if the multimodal feedback is helpful to them. In other words, the task has no extra cognitive overhead. We use a hand-held device that guides the user's hand to quickly locate the desired book in the bookcase. This wireless, compact device is able to generate the different types of feedback under evaluation. Our prototype does not need to change the shelf structure or the use of cumbersome devices.

3.1.1 Methodology

Participants started their session by filling out a questionnaire to record their demographics. Next, they watched an introductory video where they received information about the experiment and instructions on how to use the application.

They were informed that their task was to find 24 books in total: three for each one of the seven combinations of visual, auditory and tactile feedback and one extra set for the baseline condition. We also told them that the bookshelf distribution would be changed every third book (as it will be explained later in this chapter). We also informed the participants about the combination they were going to receive every time we changed the bookshelf.

Before participants started the experiment, they searched for three books to get familiar with the searching strategy implemented in the assistant. This searching strategy is explained in the Section 3.1.2. To avoid any possible bias we randomized the feedback modality participants received when completing the training session.

After the training stage, every participant started from the same bookshelf completing the search task without any help. We used these trials as the baseline condition. Then, participants received all the combinations of the three feedback modalities in a different order. We counterbalance the effect of books being located in an especially easy or difficult position across different bookshelves, and thus, avoiding any possible bias in results.

Because participants had to keep their arm up while performing the searching task, they were allowed to rest between each book search. They controlled the starting of each individual search, requesting the system for the next book title by pressing a key on the device. They also had to confirm when they found each book. All participants performed the experiment standing up in front of the bookshelf model at a distance of an arm.

On average, they needed between 12 and 15 minutes to finish the whole experiment. After completing all the trials, participants answered a questionnaire about their experience.

3.1.2 Searching Strategies

We implemented three search strategies for our system. In order to reduce the number of independent variables, we performed a pilot study to select the best one, using the same apparatus described in Section 3.1.4 in this chapter. The description of each one of the strategies is as follows:

Find Row First - FRF

This strategy has two steps. First, the system helps the participant to find the shelf where the book is located: providing a strong alert (using any combination of visual, auditory or tactile feedback) when the participants have their hand in the right shelf. In other words, requiring participants to first perform vertical movement. Once the row of books has been identified, a radar-like feedback is used in which the stimulus increases as the user's hand gets closer to the desired book, guiding participants to perform lateral hand movement. See Figure 3.1a for a graphical explanation of this strategy.

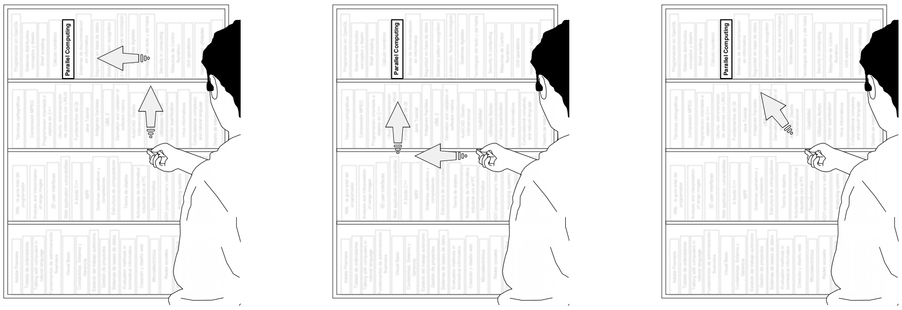
Find Column First - FCF

This strategy also has two steps. First, it guides participants performing a lateral hand movement. In this step, the system provides feedback to help the participant finding the column where the book is located. In the second step, the feedback increases as the hand gets closer to the book, while the participant scans the bookshelf in vertical direction. See Figure 3.1b for a graphical explanation of this strategy.

Euclidean Distance - EUC

This is the most straightforward of the three strategies we implemented. It only has one step, and the idea is that the system increases the stimulus frequency as the user's hand gets closer to the book from any direction. See Figure 3.1c for a graphical explanation of this strategy.

We followed the same protocol as in the main experiment. Participants began their session filling out a survey to record their demographics. Next, they watched an introductory video where they received all the information about the experiment, instructions on how to use the application and instructions about how each searching strategy works.



(a) Find row first (b) Find column first (c) Euclidean distance

Figure 3.1: The best scenario for each one of the three searching strategies we evaluated.

The first task participants had to perform in this part of the experiment was a searching task without any help (the baseline condition). Then, they tried out every search strategy in a random order to ensure there was no effect of ordering on the results. We compared the performance of participants completing the searching task (three books with each one of the searching strategies) and also their subjective opinion.

We quantify user performance as the average time to complete each trial. Before participants started searching for the books using a particular strategy, we explicitly informed them which strategy they were about to use and described it again, if necessary.

We invited 14 participants for this part of the experiment: 11 men and 3 women. They were 27.6 ± 5.6 years, ranging from 21 to 43 years. All of them were right handed, and in this case, all the participants were PhD. students at our university.

We performed a *One-Way Repeated Measures ANOVA*. We selected the *pairwise T-test with Holm correction* method to perform post-hoc tests. All tests were carried out using a 95% confidence level.

The results showed that the *Find Row First* strategy produced faster searching times ($F(3, 39) = 4.01$, $p = 0.019$*, *generalized* $\eta^2 = 0.135$) compared to the other two strategies and also with respect to the baseline condition. The * indicates significant differences with $p < 0.05$; we will keep this notation throughout this thesis where the text or tables describe the results of an statistical analysis.

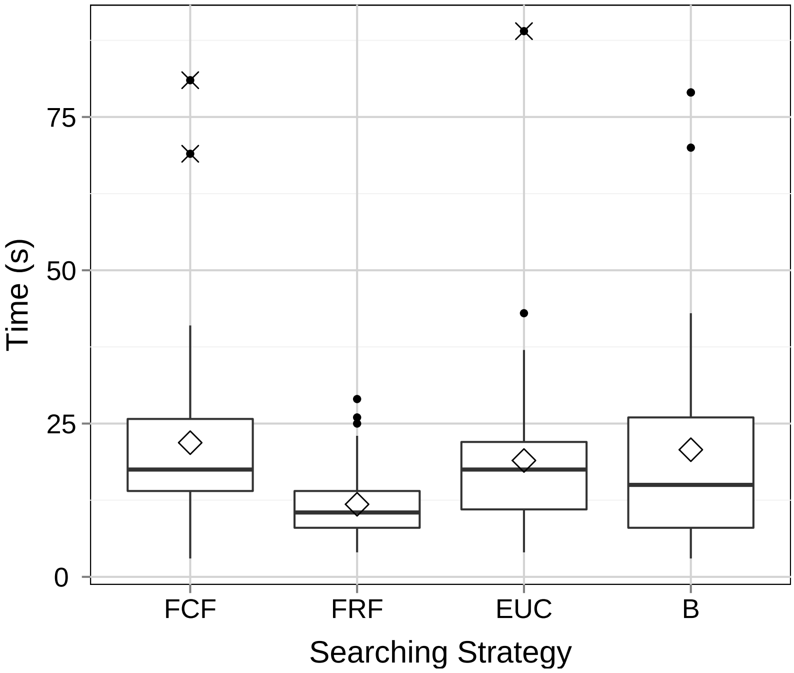


Figure 3.2: The results of the pilot study we performed to select the best searching strategy. In this figure, and in all the figures showing a Box plot throughout this thesis, boxes represent the interquartile range (IQR), horizontal lines inside each box represent the median value, white diamonds depict the average value, and dots with a cross represent outliers that are more than 3 times the IQR from the box.

Figure 3.2 shows the Box plot of the search time. We removed one far outlier of the Baseline group to improve the visualization of the rest of the boxes in the chart. This sample belongs to a participant who completed one trial in 200 seconds. Nevertheless, it is worth to note that we included all the samples in the analysis after applying data transformation techniques to the data in order to fulfill the ANOVA requirements.

Besides better performance results, 11 out of 14 participants mentioned that the *Find Row First* strategy was the best strategy they tested. They mentioned that it was faster and provided information about the location of the book clearly, as “it shows rapidly the shelf where the book is”.

Some comments about the *Euclidean Distance* strategy indicated that it was confusing because of the lack of a more detailed indication of the direction to follow for the initial movement. Therefore, we selected the *Find Row First* strategy to perform the experiment.

3.1.3 Participants

For this experiment, we recruited 24 participants, 18 men and 6 women. They had a mean age of 27.6 ± 3.5 years, ranging from 19 to 34 years. All the participants were right handed and were either undergraduate, master or PhD. students at the Universitat Politècnica de València. All participants volunteered their time, as they received only a couple of chocolate cookies as their reward.

None of the participants for the pilot study to select the searching strategy described before was considered for the main study.

Participant's previous experience using the Nintendo Wii remote was not relevant in our experiment, because we did not use most of the characteristics used in a regular Wii game (like the accelerometers or the infrared sensor).

3.1.4 Apparatus

We implemented the searching assistant using a Nintendo Wii remote, a Microsoft Kinect camera and a desktop PC. The experimental setup fulfilled the requirements for our study: it provides the three types of feedback signals, it is not heavy or cumbersome, has tracking capabilities, and does not require modifying the structure of the bookshelf. An actual implementation of this system could use, for example, a smartphone with QR scanning capabilities to recognize the bar code of the books to provide the guiding aids to find the desired book.

The user controls the interaction with the Nintendo Wii remote's button for three actions: request the next book title, ask the system to repeat the title in case it is necessary, and confirm the target acquisition.

At the same time, the Nintendo Wii remote provides visual and tactile feedback to the user. Visual feedback is provided using the Wii remote's LEDs and the vibrator incorporated in this device provides tactile feedback. We employed the Wiiuse library (Laforest, 2008) to connect the PC with the Nintendo Wii remote wirelessly via Bluetooth.

The application uses the Festival 2 (Clark et al., 2004) text-to-speech library to utter the target book title. The Wiiuse library does not support control of the Nintendo Wii remote speaker, hence, our system provides auditory feedback using the PC speakers.

The Microsoft's Kinect used to track the participant's hand is placed on top of the bookcase, pointing downwards (see Figure 3.3). We used the OpenNI (OpenNI Organization, 2010) library to connect the Kinect camera to a PC running Windows 7 and to perform user's hand tracking.

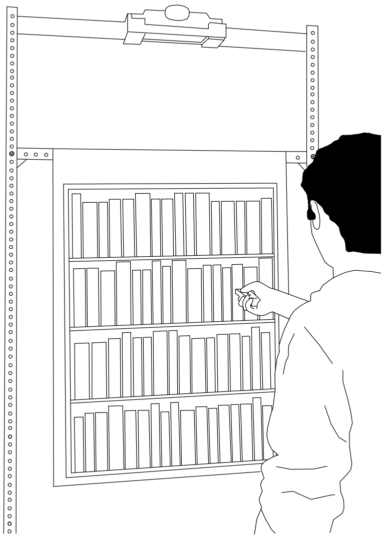


Figure 3.3: The bookcase model, the tracking system and the interaction interface of our experiment.

In order to gain flexibility in our experiment, we designed a model of a bookcase instead of using a real one. Thus, we were able to test different book distributions without changing the computer setup neither requiring a large space to set up the experiment. We printed seven book distributions on A0 paper sheets, one per modality (visual, auditory, tactile and their combinations). Each distribution had 65 books in four shelves: 16 books in the top three shelves and 17 books in the last shelf.

The system generates four discrete signals to provide information on the distance from the current hand's position to the desired book. The four signals are: book found, close to book, medium distance, and far from book. The system provides visual feedback using the LEDs on the Wii remote. The number of lit LEDs increases as the hand gets closer to the target (one LED means far, four LEDs, book found). For tactile and auditory feedback, the frequency of the pulses changes according to the distance between the Wii remote and the target. The frequencies range from 0.77 Hz (far) to 4 Hz (book found). The duration of each pulse is 100 ms. Figure 3.4 shows the four signal areas, which vary according to the distance from target.

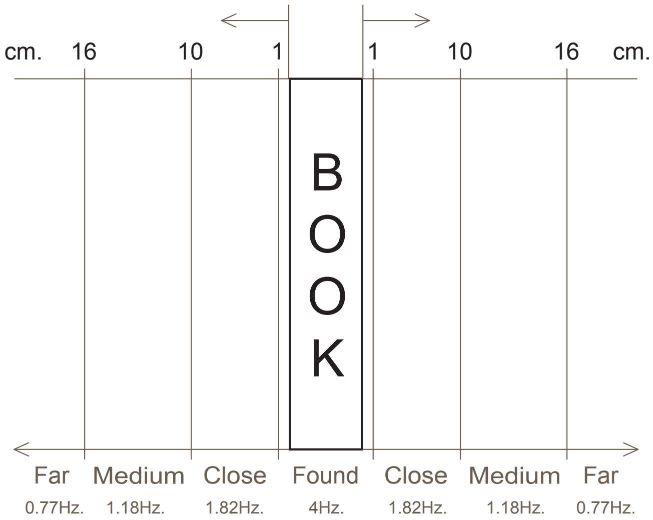


Figure 3.4: Feedback frequency is defined by the four discrete areas depicted in the figure.

3.2 Statistical Analysis

We assessed user performance for each feedback combination measuring the time spent searching for the book (T_s). Additionally we evaluated the user experience through a post–study questionnaire. We collected 576 samples in total from all the participants (72 per feedback modality combination).

According to previous research, we expect that providing any type of feedback to the participants should increase their performance with respect to the base line condition with no feedback. However, as the task is an intensive visual task, we also hypothesize that providing visual feedback should not dramatically improve user performance. Therefore our research questions are:

- Does providing any type of help to the participant reduce T_s compared to the T_s without assistance?

- How does visual proximity cues affect user performance in this evidently visual task?

The single factor to account for in our experiment was the feedback modality received by the participant. This factor has eight levels (no feedback, and the seven possible combinations of visual, auditory and tactile modalities). We used a *within-subjects* experimental design, where the participants tested all seven modalities and the baseline condition. We used a *Latin Square* design to counterbalance the order in which participants tested the feedback modalities.

To study the data captured in the experiment, we performed a *One-Way Repeated Measures ANOVA*. We selected the *pairwise T-test with Holm correction* method to perform post-hoc tests. All the tests were carried out using a 95% confidence level.

3.2.1 Performance Results

The outcome of the One–Way Repeated Measures ANOVA shows a significant effect of feedback modality on T_s ($F(7, 161) = 23.53$, $p < 0.01^{**}$, *generalized* $\eta^2 = 0.37$). The ** indicates significant differences with $p < 0.01$; we keep this notation where the text or tables describe the results of an statistical analysis throughout this thesis.

The post-hoc analysis to compare each group with respect to the others revealed a significant difference between the baseline condition (no feedback) and every other feedback combination tested ($p < 0.01$). It also showed a significant difference between the visual modality alone and all the other combinations. The visual-only feedback resulted in poorer user performance, presenting the longest completion time.

Conversely, there were no significant differences between the other modalities (including those where visual feedback is combined with other modalities).

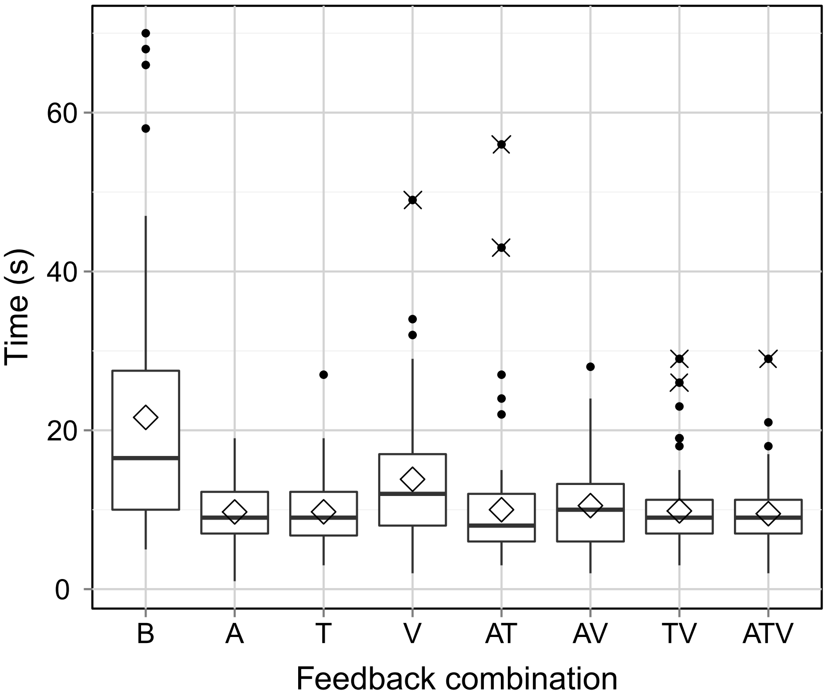


Figure 3.5: Box plot as a function of feedback modality.

Figure 3.5 shows the Box plot for the average task completion time for each modality.

In Figure 3.5, all feedback modalities and the base line condition are represented as follow: Baseline (B), Auditory (A), Tactile (T), Visual (V), Auditory-Tactile (AT), Auditory-Visual (AV), Tactile-Visual (TV) and Auditory-Tactile-Visual (ATV) modalities.

3.2.2 Questionnaires

Besides the performance measures, we also gathered subjective information from the participants using questionnaires. The goal was to study how they perceived each of the types of feedback.

The questionnaires included yes/no, open, ranking and Likert scale questions using a 5-point scale. In the last type of questions, 1 represented the lowest rate (e.g. nothing at all) and 5 the highest rate (e.g. very much).

We asked the following questions:

- Likert questions.
  - How much did you like using the application?
  - How easy was using the application?
  - How much do you think the application helped you to find the books?
- Yes/no and open questions.
  - Would you use this application in your library?
  - Were you reading the book titles while you were searching for the target?
  - What would you suggest to improve the application?
- Ranking questions.
  - Indicate, in ascending order, the importance of the individual feedback modalities while using the application.

Regarding how much participants liked using the application, the majority (66.7%) chose the highest value on the scale (5–very much), 29.1% liked it and 4.2% didn't like it. For 50% of the participants using the system was very easy; 45.8% said it was quite easy and 4.17% reported that it was neither easy nor complex to use.

Participants rated the help they thought they received from the device as follows: 50% of the participants reported that it helped them a lot; 45.8% stated that it had helped them in some degree, and 4.2% said the system neither helped nor hindered task execution.

All participants with the exception of one reported that they would use a system like this in their public libraries. It is worth to mention that 20.8% of the participants did not read the book titles while searching for the target, relying exclusively upon the feedback provided by the system.

Table 3.1 shows how participants ranked the visual, auditory and tactile feedback channels with respect to their perceived importance for the task. We can observe how tactile and auditory feedback were more important to provide the proximity cues for the participants. Visual feedback was rated as the least important modality to receive the proximity cues.

Table 3.1: Feedback modality importance

		Level of importance		
		Highest	Medium	Lowest
	Auditory	25.0%	58.3%	16.7%
Feedback modality	Tactile	58.3%	25.0%	16.7%
	Visual	16.7%	16.7%	66.6%

3.3 Discussion

The results we got with our experiment gave us a positive answer to the first research question. Participants performed better when they received any type of feedback, allowing them to complete the task in less time. The *Pairwise T-tests with Holm correction* revealed statistically significant differences between performing the search process without any help and using any of the feedback combinations assessed.

Our experiment also confirms that providing proximity cues using visual feedback increases T_s, as this sensory channel is already occupied reading the books' titles. The average task completion time for participants receiving visual-only feedback was the highest of all combinations. Visual feedback was rated by participant as the least useful feedback in this experiment.

The results of our experiment are similar to those reported in the literature for other type of applications. However, we believe that the poor results of the visual feedback in our experiment are due to its design. The device deviates users' visual attention from the bookcase, when using visual feedback. Moving the visual feedback from the Wii remote to the shelves (e.g. using one LED per book in the bookshelf and turning it on to highlight its location), or using a smartphone as the tracking and guiding device, employing the screen as a Magic Lense (Bier et al., 1993) that shows proximity and directional cues mixed with the reality (the books' names), would probably improve user performance using visual feedback.

Regarding participants' subjective opinion about the experiment, the questionnaires show that 66.6% of the participants thought it was the least helpful feedback modality. An interesting observation is that, from the four participants that chose the visual feedback as the best, two obtained the worst task completion times using this modality alone; one got the second worst time, and the other participant, her third best average time. Tactile and auditory feedback modalities were regarded as the most helpful (58.3% of the participants chose the vibrotactile cues as the best modality, and 25% chose the auditory cues). Participants mentioned that tactile feedback was the most natural way to receive the feedback, and also the most discrete, as auditory feedback could disturb other people in the room, e.g. when using the application in a library.

Only one participant complained about getting confused when the three modalities were used, as the auditory feedback distracted her from understanding the location feedback. This situation might be related to the findings of Bresciani et al. (2008), as the participant got confused by the presence of different modalities at the same time.

3.4 Chapter Summary

Results described in this chapter reinforces the theory that using any combination of feedback for the human-computer interface of a searching assistant is better than no feedback at all. We have also observed that providing feedback through an already saturated sensory channel (visual feedback in this experiment) represents a disadvantage rather than a benefit for user performance. However, we cannot go further in the analysis, as there are no statistically significant differences among the other feedback modalities combinations.

Results presented in this chapter are the first step in a series of experiments we conducted to understand multimodal human–computer interfaces. This experiment gave us a base line to design future experiments to confirm or reject the findings that have been described here.

In the following chapter, we go one step further and evaluate a more complex AR application in a very popular domain: gaming.

CHAPTER 4

A MULTIMODAL AUGMENTED REALITY GAME

Alerting cues are commonly used in different application domains to inform users about abnormal situations that demand their attention. For example, in a driver assistance system, alerting cues can inform the user that the distance to the car ahead is too small. In a surgical simulator, alerting cues can inform the surgeon about critical vital signs of a patient. In a military simulator, alerting cues can be used to provide the position of possible targets or enemies. Gaming is another domain in which alerting cues are widely used to capture player's attention. Gaming is also one of the most popular application domains in AR. In most cases, one constraint of the design is that the visual attention of the user must remain focused on the main task.

Considering the results of the experiment described in Chapter 3 as a baseline—multimodal feedback improves user performance—the goal of the experiment described in this chapter is to improve the understanding of how to properly use multimodal feedback in AR interfaces.

Key contribution - The findings presented in this chapter can enhance the design of many AR applications, since the viewing angle offset is a very common problem for current desktop AR applications. On the other hand, multimodal alerting cues can be used in different ways, for example, to enhance selection of virtual objects by informing the user when the task has been completed, or to get the attention of the user during abnormal or dangerous situations. The outcome can be applied to scenarios where user's attention needs to be focused on the main task, for example: gaming, military or medical training simulators and portable tourism guiding applications.

4.1 Experiment Description

Most of the currently available 3D games involve certain types of manipulation of virtual elements. Either controlling where a character moves while exploring the virtual world or driving a car through a race circuit require changing the position and orientation of some virtual elements, taking into account their surroundings in the virtual world. We chose to develop the Augmented Reality version of the Wire Loop Game to evaluate how multimodal feedback can be used as alerting cues in an AR application that requires the user to manipulate objects in the three-dimensional space. This represents a step forward with respect to the previous experiment (see Chapter 3), as now we evaluate a more complex task that requires precise movements. The user needs to change the position and orientation of a virtual object by manipulating a real object in the real world. The Wire Loop Game (see Figure 4.1) is a popular dexterity game for children that requires users to traverse a twisted wire path with a wire loop without touching one another.

We compare how providing alerting cues through all the combinations of visual, auditory and tactile feedback help users to accomplish the goal of the game. We asked 210 volunteers to play our game and recorded various performance measures to determine how well they perform in relation to the feedback modality combination they received. Additionally, we evaluated how each modality is subjectively perceived by the users.

4.1.1 Methodology

The session with each participant started with a brief explanation describing the goal of the game: finish every level as quickly as possible and with the lowest

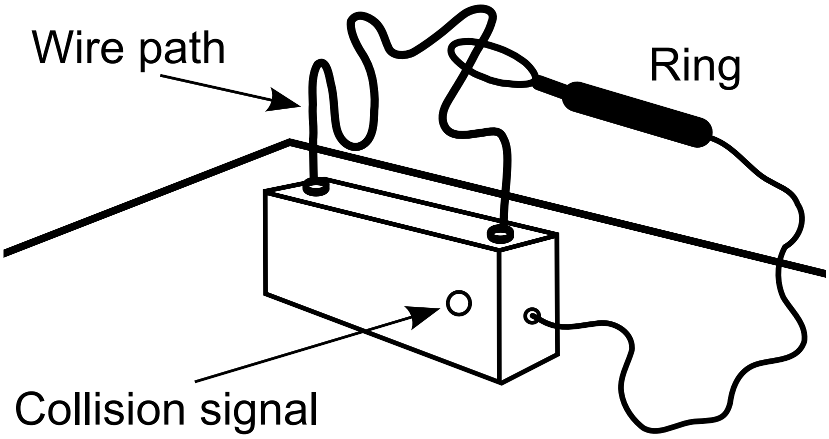


Figure 4.1: A model of the real Wire Loop Game: an electric circuit is closed when the ring touches the wire path, switching on the collision signal to alert the user about the collision.

number of errors. Then, we asked participants to fill out a questionnaire regarding their personal information. An introductory 3.5 minutes video provided instructions on how to play the game.

The game has three complexity levels (see Figure 4.2). The complexity of each level depends on the number of curves in the virtual wire path: one, three or five curves for the easy, intermediate, and difficult level, respectively. We used the first two levels of the game to allow the participants to get used to control the virtual loop using the real wand, and also to learn the spatial relationship between the real and the virtual objects. We employed the third level of the game, where the participant had to perform the most complex movements, for evaluating the performance results.

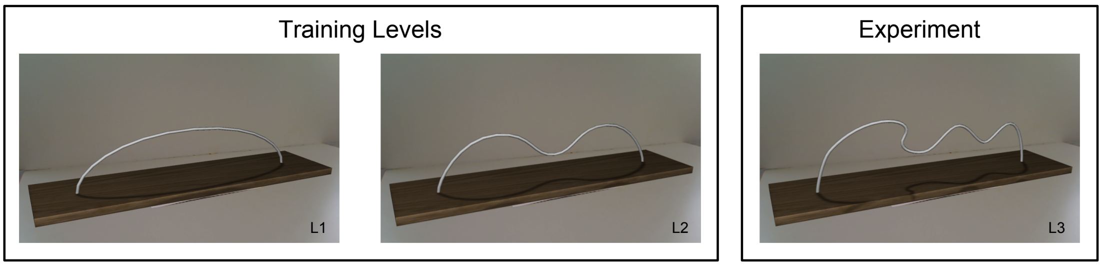


Figure 4.2: The game has three difficulty levels. We only take into account the results of the third level in our experiment.

We used a *between subjects* design, where each participant was randomly assigned to one of the following groups, determining what combination of feedback they received during the experiment:

Group 1 (A) : Auditory feedback only.

Group 2 (T) : Tactile feedback only.

Group 3 (V) : Visual feedback only.

Group 4 (AT) : Auditory and tactile feedback.

Group 5 (AV) : Auditory and visual feedback.

Group 6 (TV) : Tactile and visual feedback.

Group 7 (ATV) : Auditory, tactile and visual feedback.

Participants played each level (starting with level one) seated in front of the laptop with the webcam to their left/right side, depending on their dominant hand. They observed virtual objects (that appear between them and the laptop) from the camera's point of view. The camera was placed 15 cm above the main marker on the table, and 30 cm away, oriented at 45° with respect to the main axis of the path (see Figure 4.3).

We decided to use this configuration because the relative position of the loop with respect to the path was better perceived. At the same time, this configuration increased the complexity of the task, forcing participants to perform a movement in the 3D space to complete the paths, instead of just moving their hand up and down and laterally in a 2D plane once they find the correct depth to position their hand. By keeping the camera position fixed, we ensured that any mistake (collision between objects or taking the ring out of the path) was caused by the hand movements (hand-eye coordination) and not because of camera movements (e.g. when mapping head movements to the camera position).

Observing the hands in the working space from a different point of view is also commonly used in several professional domains, such as laparoscopic surgery (see Figure 4.4). In this domain, surgeons use images from a camera shown on a screen that should be placed in the line of vision of the surgeon (Levy and Mobasheri, 2014). Thus, surgeons have to receive specific training to learn how

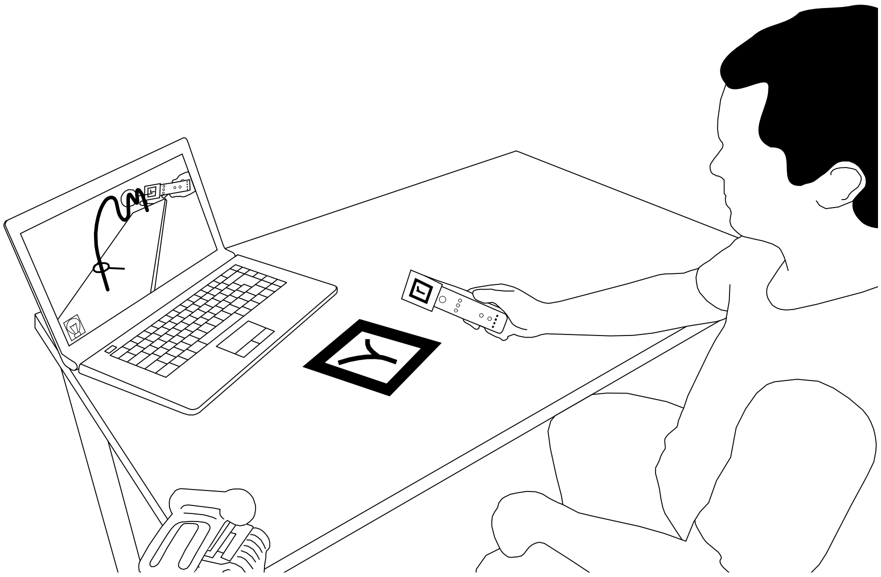


Figure 4.3: The webcam is placed to the side of participants, above the main marker. Participants were seated in front of the laptop controlling the virtual ring using the Wii remote.

to transform the 2D image from the camera inside the patient to 3D movements, while observing the scene on the monitor, requiring good hand-eye coordination (Grantcharov and Funch-Jensen, 2009).

After each level, we allowed participants to take a one minute break. Once all levels had been completed, we asked the participants to fill out a questionnaire and asked a few open questions about their experience playing the game in an interview format.

4.1.2 Participants

We recruited 210 computer science students for our experiment. They were between 17 and 33 years old (mean of 21.79 ± 3.62). In this group, 149 participants were men (70.95%) and 61 were women (29.05%). Only 6 participants were left-handed (2.86%) and 23 of the participants (10.95%) had previously used an AR application. None of the participants had played our game before. Most of the participants (197 or 93.81%) had experience with playing a game on the Wii

[2] Image from `http://www.defense.gov/photos/newsphoto.aspx?newsphotoid=6052`

Figure 4.4: Laparoscopic surgery is one example of a professional domain where people have to perform a manual task observing a different point than the one where their hands are.[2]

console using a Wii remote. We grouped participants in two categories with respect to the amount of hours they spent playing video games: 134 participants (63.81%) were frequent gamers (more than 5 hours a week) and 76 (36.19%) were non-frequent gamers (between 1 to 5 hours a week). We randomly divided all users into seven groups of 30 users, one group per feedback combination as explained before.

Table 4.1 shows the distribution of the different participants across the seven test groups.

Table 4.1: Participants distribution per test condition.

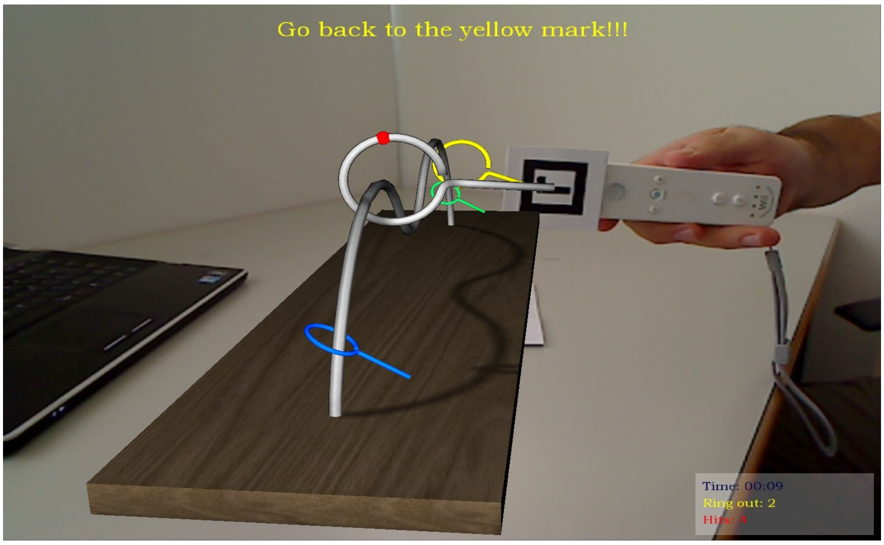
	Gender		Dominant hand		AR Exp.		Wii Remote Exp.		Game Exp.	
Group	Men	Women	Right	Left	Yes	No	Yes	No	Non-Frequent	Frequent
A	20	10	29	1	10	20	30	0	10	20
T	18	12	30	0	3	27	30	0	12	18
V	22	8	30	0	4	26	26	4	8	22
AT	19	11	29	1	2	28	26	4	12	18
AV	24	6	29	1	1	29	26	4	10	20
TV	24	6	30	0	1	29	29	1	12	18
ATV	21	9	27	3	3	27	30	0	12	18

4.1.3 Apparatus

The real version of the Wire Loop Game demands the steady visual attention of the player to keep track of the spatial relationship between the elements of the game. It also requires good hand-eye coordination and precision in hand movements. In case of a collision between the objects in the real game, the player receives a sound warning, and haptic feedback when both objects collide.

We used the typical configuration of a Desktop AR application. The computer monitor shows the augmented scene: the video feed with the (real) desktop and the hand of the user, and the AR wire loop game. The video is captured with a webcam connected to the computer, and the registration between the real and virtual scene is achieved using two printed markers.

Figure 4.5 shows the *standard game* and the second training level in our experiment. The standard game includes the following features: the virtual path, the loop, and the shadows. It also includes two additional elements that guide the user in the game: a red sphere in the loop shows the point where the loop last touched the wire—during half a second—and a yellow "phantom" loop shows where the user last crossed the wire, until the user returns to that point. The starting and ending points are highlighted by a green and a blue loops respectively. All these elements are present in any of the conditions.



Go back to the yellow mark!!!

Time: 00:09

Ring out: 2

Hits: 4

Figure 4.5: The standard game and its visual elements.

If the participant tries to cheat by accident or on purpose, the game forces the participant to go back to the place where the ring left the path, indicated by the yellow phantom ring. In this way, the participant cannot start the game and take the loop from the beginning to the end without going along the path. Also, the game cannot end while the loop is outside the path.

The wire path and its support are drawn on top of a marker placed on a table in front of the webcam. That marker establishes the position of the path, and it is not moved during the game. The virtual loop is handled by the user by means of a Nintendo Wii remote with a smaller marker attached.

The visual feedback of the system is provided by the computer display. The auditory feedback is generated by the speakers of the laptop computer. The vibrotactile feedback is provided by the Nintendo Wii remote.

We ran our experiments on a Dell XPS 1647 laptop, with an Intel Core i5 CPU running at 2.40 GHz, 4 GB RAM and an ATI Mobility Radeon HD 4670 graphics card with 1 GB of dedicated memory, and a 15.6” Full HD screen. The webcam was a Logitech Pro 9000 HD (1600 × 1200 pixels and up-to 30 fps) connected to a USB 2.0 port. The operating system was Windows 7. The Wii remote was connected to the PC via Bluetooth.

The Augmented Reality Wire Loop Game was implemented using the osgART library (Looser et al., 2006). This library provides fast integration of real-world and virtual 3D objects, as well as high quality graphics. It is based on ARToolKit (Kato and Billinghurst, 1999) and uses printed markers to compute the position and orientation of the virtual objects from the camera point of view. The Wii remote was controlled with the Wiiuse library (Laforest, 2008).

There are three possible states during the game: *(i)* the path is inside of the loop, without touching it, *(ii)* the loop is touching the path, a *collision event*, and *(iii)* the loop has completely crossed the path, a *crossing event*. Feedback is provided to the participant depending on these states.

We designed the game to be as close as possible to its real counterpart. However, we decided to use different sounds and vibration patterns for *collision* and *crossing events* as explained in the following paragraphs.

4.1.3.1 Visual Alerts

Visual alerts are provided using three elements: a light bulb in the lower left/right corner that changes colour, text messages that appear above the wire, and changes in the colour of the loop. The position of the light bulb depends on the hand used by the participant.

The light bulb is lit red on every *collision event*, as long as the wire loop is touching the path. A text message displays the number of times the user has touched the wire with the ring; the message it is visible during 2 s. The ring becomes transparent (changing the alpha channel value from 1.0 to 0.3) immediately after a *crossing event* and will remain that way as long as the wire loop is out of the path. Figure 4.6 shows the elements that make up the visual alerts of the game.

The lowest frame rate in our game was 28 fps. This occurs when using all three feedback channels while playing the third level of difficulty. The highest frame rate while playing our game was 36 fps.

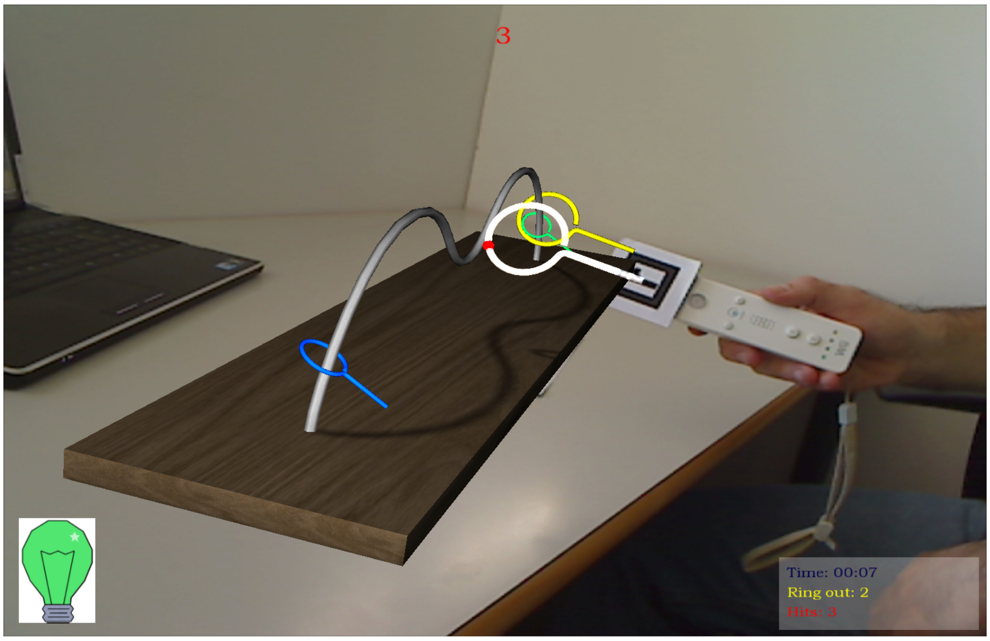


Figure 4.6: Visual feedback elements: the light bulb in the lower-left corner, the text element displaying the number of collisions on top of the wire, and the change in the colour of the ring when it is outside the path.

4.1.3.2 Auditory Alerts

Auditory alerts provide feedback on a *collision event*, playing a loud sound, which lasts as long as the user touches the path with the loop. This behaviour imitates the real game. The audio track used for this alerting cue was a 1 s long WAV audio file (16 bit PCM format), with only one channel, with levels in the range $[-1, 1]$ and a frequency of 11.025 Hz. We adjusted the volume of the speakers to reproduce the sound at 80 dB. We measured the decibels, using the SPL Meter mobile application for iOS, at approximately the location of the participant's head.

The system plays a warning sound, softer than the previous one, if the game is in the third state (a *crossing event*). This warning feedback is repeated continuously while the ring is outside of the path. The audio track used for this alerting cue was a 0.8 s long WAV audio file (32 bit PCM format), stereo, with levels in the range $[-0.08, 0.08]$ and a frequency of 44.100 Hz. This sound was played at 55 dB.

There is no auditory feedback when the participant is moving the loop correctly. Our system needs 0.04 s to play the auditory feedback when a collision or crossing event occurs.

4.1.3.3 Tactile Alerts

Tactile alerts are provided by means of the vibrator of a Nintendo Wii Remote. As in the case of auditory feedback, when the user is on the right path, there is no tactile feedback. When a *collision event* occurs, the participant receives a 0.5 s long vibrotactile pulse. If the loop crosses the path completely, a continuous vibration is given through the Wii remote.

We performed a pilot study to decide the duration of the tactile collision warning. 10 colleagues played the game with 3 different durations of the tactile collision alert (0.2, 0.5 and 1 s). All participants tested the three different durations while playing the first level of the game. We asked participants to cause a collision event on purpose at the beginning of the path so they received the tactile feedback. We balanced the order in which participants tested the three durations, 4 participants tested the 0.2 s duration first, 3 participants the 0.5 s duration first, and 3 participants the 1 s duration first. Then we asked to pick one of the three durations: 8 of the 10 participants chose the 0.5 s feedback and 2 the 0.2 s feedback. Participants who chose the former one mentioned that it was easier to perceive the collision warning without being too annoying.

Hence, we decided to use the 0.5 s tactile feedback as the alerting cue for the Augmented Wire Loop Game. The Wii remote motor state has only binary control (on and off), thus once a collision event is detected, our system requires 0.03 s to activate the vibrator of the Wii remote.

4.2 Statistical Analysis

We studied the data collected for five performance measures:

- Feedback modality performance indicators.

  Average collision time (T_c) : average time it takes the loop to stop touching the path after a collision.

  Average time outside the wire path (T_o) : average time it takes the participant to bring the loop back to the path, after crossing it.

- Game scores.

  Completion time (T) : time required to complete a game level.

  Number of collisions (N_c) : number of times the loop touches the path in a game level.

  Number of crossings (N_x) : number of times the loop crosses the path in a game level.

The first two measures can be considered as direct performance indicators for our experiment, because they can be affected by the feedback received by the participant. The last three measures represent scores directly related to the game and they are not affected by the feedback modality the participant received. In our experiment, the alerting cues were not designed to guide users' hand throughout the virtual path helping them making less mistakes.

We designed this experiment to answer the following research questions:

- Do vibrotactile cues in a desktop AR precision task affect user performance negatively? Previous research has proven that vibrotactile feedback can be an effective way for alerting the user, but not that beneficial in a target finding task.

- Does auditory feedback outperform vibrotactile feedback as an alerting cue? Auditory feedback is less intrusive for this precision task.

- Does visual feedback, being the predominant sense for most humans, outperform conditions that does not include it?

- Does multimodal feedback produce better user performance compared to individual modalities? Previous research in other domains has shown that multimodal interfaces result in better performance compared to providing feedback with one modality.

We employed the *ANOVA* test to compare the effect of the feedback modality on the performance of the participants. We did not consider factors such as gender, gaming experience, previous knowledge of AR, previous experience using the Wii remote and the hand participants used to play the game in our analysis, because of the small number of participants in some of the groups.

A *Shapiro-Wilk* test of normality and *Bartlett Test of Homogeneity of Variances* showed that our data did not fulfil *ANOVA* prerequisites. Therefore, we used data transformation techniques to be able to use parametric tests in the analysis. In those cases where we found a statistically significant difference, we used the *Tukey HSD* test as *post-hoc* method at the 95% confidence level.

To analyse the results of the questionnaires, we used the *Kruskal-Wallis Rank Sum* test to compare the effect of feedback modality on the responses of the participants for the *Likert* scale questions. We also provide *effect sizes* for every test to have a measure that is independent from the sample sizes and to have a magnitude of the significant differences found, which is not measured by *p-values.* We consider the following thresholds for the partial η^2 effect size[3]: *small* ≥ 0.01, *medium* ≥ 0.06, *large* ≥ 0.14.

[3] From http://yatani.jp/teaching/doku.php?id=hcistats:anova

4.2.1 Performance Results

The ANOVA test for average collision time did not show any statistically significant differences for feedback modality factor ($F(6,203) = 0.8199$, $p = 0.55$, *partial* $\eta^2 = 0.02$). Table 4.2 shows the descriptive statistics for every feedback modality combination.

Table 4.2: Descriptive statistics for T_c.

Group	Mean	SD	Min. Value	Max. Value
A	0.12	0.04	0.06	0.25
T	0.14	0.10	0.04	0.41
V	0.13	0.06	0.05	0.25
AT	0.15	0.07	0.05	0.32
AV	0.14	0.07	0.07	0.33
TV	0.13	0.07	0.05	0.38
ATV	0.13	0.07	0.04	0.33

The ANOVA test for the average time outside the wire path (T_o) showed statistically significant differences for the feedback modality factor ($F(6,203) = 3.05$, $p = 0.007^{**}$, *partial* $\eta^2 = 0.08$). The post-hoc analysis revealed statistically significant differences between the AV group (0.39±0.37 s in average) and: T only (0.62±0.38 s), AT (0.65±0.42 s) and TV (0.0.71±0.56 s) groups. Table 4.3 shows the descriptive statistics for average time outside the path by feedback modality. Figure 4.7 depicts the box plot and the average T_o for every feedback modality group.

Table 4.3: Descriptive statistics for T_o.

Group	Mean	SD	Min. Value	Max. Value
A	0.56	0.35	0.03	1.22
T	0.78	0.73	0.09	3.74
V	0.62	0.38	0.08	1.54
AT	0.65	0.42	0.11	1.70
AV	0.39	0.37	0.04	1.83
TV	0.71	0.56	0.16	2.55
ATV	0.62	0.38	0.08	1.61

The ANOVA test for the completion time (T) for the feedback modality factor revealed statistically significant differences ($F(6,203) = 2.20$, $p = 0.04^{*}$, *partial* $\eta^2 = 0.06$). However, the post-hoc comparisons showed no statistically significant difference between groups. Figure 4.8 depicts the box plot and the average completion time for every feedback modality group. Table 4.4 shows the

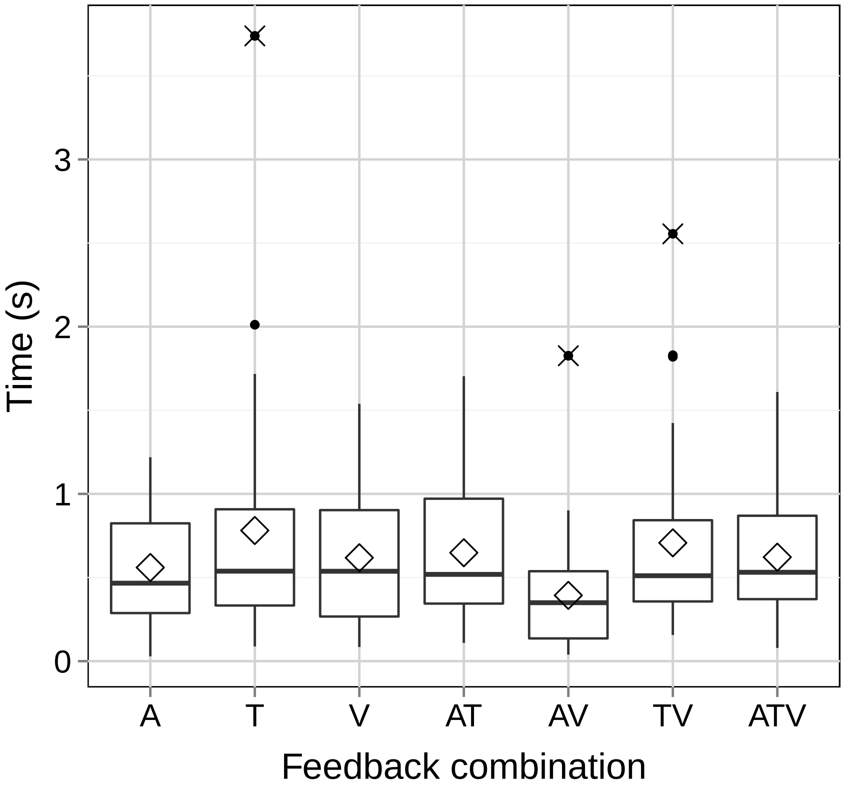


Figure 4.7: Average time outside the wire path (T_o) box plot as a function of feedback modality group for Level 3.

descriptive statistics for completion time for every feedback modality combination.

Table 4.4: Descriptive statistics for T.

Group	Mean	SD	Min. Value	Max. Value
A	58.54	30.07	16.52	138.49
T	65.78	30.98	20.85	127.69
V	48.87	31.16	16.99	164.96
AT	55.55	25.85	17.39	109.25
AV	46.12	22.91	20.03	123.15
TV	57.65	27.81	16.44	124.17
ATV	61.34	28.73	24.72	154.06

The ANOVA test for number of collisions did not show any statistically significant differences for feedback modality groups ($F(6, 223) = 1.65$, $p = 0.13$, *partial* $\eta^2 = 0.04$). Table 4.5 shows the descriptive statistics for number of collisions for every feedback modality combination.

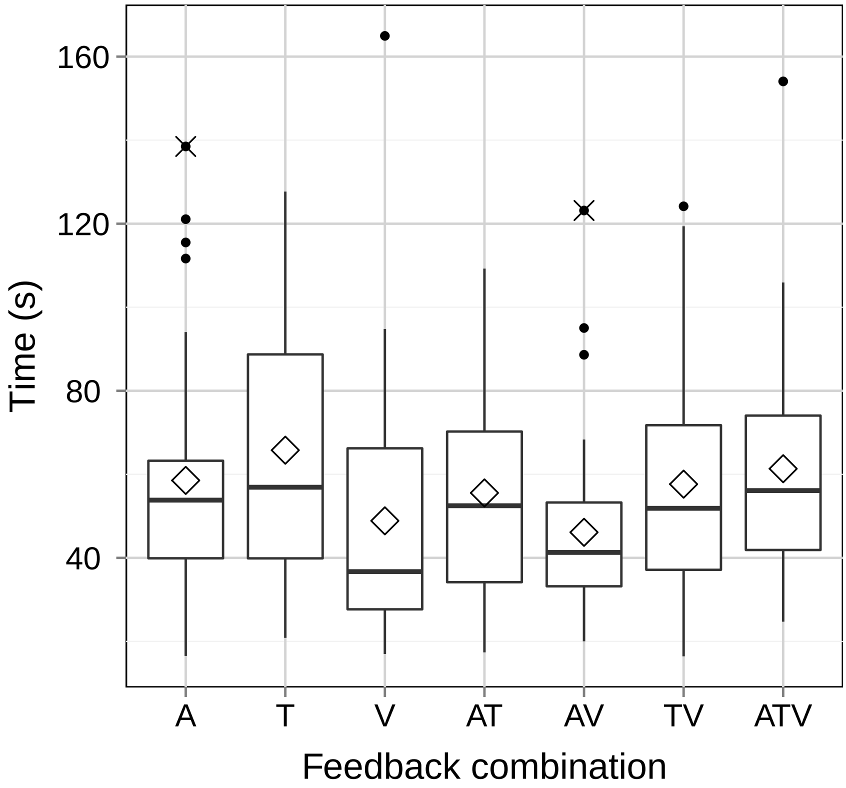


Figure 4.8: Completion time (T) box plot as a function of feedback modality group for Level 3.

Table 4.5: Descriptive statistics for N_c.

Group	Mean	SD	Min. Value	Max. Value
A	35.06	18.02	11	74
T	38.00	17.89	10	76
V	30.23	17.54	14	85
AT	32.10	15.39	12	73
AV	25.57	10.11	11	47
TV	34.07	16.21	8	65
ATV	33.87	18.83	13	90

The ANOVA test for number of crossings also did not show any statistically significant differences for feedback modality factor ($F(6, 203) = 2.03$, $p = 0.06$, *partial* $\eta^2 = 0.05$). Table 4.6 shows the descriptive statistics for number of crossings for every feedback modality combination.

Table 4.6: Descriptive statistics for N_x.

Group	Mean	SD	Min. Value	Max. Value
A	15.43	12.52	2	43
T	17.63	11.51	3	51
V	12.63	11.88	1	53
AT	12.00	8.12	1	33
AV	9.47	5.69	1	22
TV	14.43	9.00	2	39
ATV	14.90	9.87	1	33

4.2.2 Questionnaires and Interviews

After the experiment we gathered subjective information from the participants through questionnaires and interviews. The questionnaires included yes/no questions, open questions and *Likert* questions (from one to five, being one the lowest and five the highest score for each question):

- Yes/no and open questions.
  - Q1 - Have you ever played the real version of this game?
  - Q2 & Q3 - If you have played the real version of this game, which one do you prefer? Why?
  - Q8 - Describe your experience playing our Augmented Reality game.
  - Q9 - What would you suggest to improve the game?
- Likert questions.
  - Q4 - How much did you enjoy playing the Augmented Reality game?
  - Q5 - How difficult do you think it was playing the first level of the game?
  - Q6 - How difficult do you think it was playing the second level of the game compared to Level 1?
  - Q7 - How difficult do you think it was playing the third level of the game compared to Level 2?

Only 66.67% of the participants (140 of 210) in our experiment had played the real version of the Wire Loop Game. Column Q1 in Table 4.8 presents the

Table 4.7: Statistical analysis for the answers to the Likert scale questions. The effect size column for feedback modality presents the maximum effect size of all the comparisons among the groups.

Question	Median	d.f	Statistic	p-value	Effect size
Q2	1	6	$\chi^2 = 6.073$	0.425	$r <= 0.316$
Q4	2	6	$\chi^2 = 6.611$	0.358	$r <= 0.217$
Q5	2	6	$\chi^2 = 2.430$	0.876	$r <= 0.182$
Q6	3	6	$\chi^2 = 6.444$	0.375	$r <= 0.273$
Q7	3	6	$\chi^2 = 9.308$	0.157	$r <= 0.192$

Table 4.8: Descriptive statistics for the questionnaire responses.

	Q1		Q2		Q4	Q5	Q6	Q7
Group	Yes	No	Preferred AR	Median	Median	Median	Median	Median
A	22	8	14	AR	2	2	2	2
T	21	9	14	AR	2	3	3	2
V	20	10	13	AR	2	2	3	2
AT	19	11	11	AR	3	2	2	2
AV	17	13	8	Real	2	2	3	2
TV	20	10	9	Real	2	3	2	2
ATV	21	9	16	AR	2	2	3	2
Totals	140	70	85					

detailed count of the participants who knew the real version of Wire Loop Game by feedback modality group.

We asked them to choose between the real version of the game and the AR version they played in our experiment: 60.71% (85 participants) preferred our game to the real version because of its flexibility to include more levels without requiring any extra material. Column Q2 in Table 4.8 shows the detailed count of the participants that preferred our system (Preferred AR column) and the median answer by feedback modality group (Median column). We did not ask participants who had not played the real version of the game about their preference, because they did not have a reference frame for the comparison. As shown in Table 4.7 (row labelled Q2), feedback modality factor did not have a statistically significant effect on the percentage of users that preferred our version of the game.

Regarding the enjoyment level of participants playing our game, feedback modality factor did not have any statistically significant effect, as shown in row Q4 in Table 4.7. The median value chosen by participants of every feedback modality group is presented in Table 4.8, column Q4.

As for the difficulty of playing the game (rows Q5, Q6 and Q7 in Table 4.7), feedback modality factor also did not have any statistically significant effect.

Columns Q5, Q6 and Q7 in Table 4.8 present the median value of the responses of participants of every feedback modality group.

When we asked for suggestions on how to improve the game, the most common answer was to increase the number and difficulty of available levels (69 participants, 32.85%). Participants also asked to include additional challenges in the game, for example, adding an animated character to distract the player, or an arcade mode to force the player to finish the level in a few seconds. They also asked for the option of two players playing the game at the same time in a challenge mode.

Participants liked shadows as positional and depth cues, because they were helpful to perceive the relative position of the ring and the wire. 5 participants (2.38%) mentioned that the colour of the wire loop and the wire path were too similar and made it difficult to distinguish the difference between them. Another participant asked for a stereoscopic version of the game, because *"it would improve visual perception of the relative position of the wire and the ring."*

Through the interviews and our observations during the study, we found that tactile feedback as produced by the vibration of the Wii remote made participants nervous. The constant vibration while the ring was outside of the wire path annoyed participants. We also found that the sound used to indicate a collision was too noisy for participants. One interesting finding was that participants who played without tactile feedback, asked for it, and participants who played without auditory feedback asked for a sound to be played when a collision or crossing occurred.

Only 9 participants (4.29%) reported problems with depth perception while they were playing the first training level. Still, they all were able to complete the three game levels. They mentioned that, at the beginning, it was difficult to move the hand because of the different point of view, but once they got used to it, the game was less difficult. Participants who had difficulties finishing the training levels mentioned that the task required too much physical effort, because it implied keeping the arm lifted as long as they were completing the level.

Participants also mentioned during the informal interview that the game was similar to the psychometric tests required in some countries to obtain a driving license. They recommended evaluating the game as a tool for this kind of testing.

We also got suggestions about using the game as a tool for physical rehabilitation.

4.3 Discussion

Regarding the analysis of the effect of individual feedback modalities, the use of vibrations or loud sounds as feedback caused a negative effect on participants' performance. Those groups that played with any combination including tactile feedback resulted in poorer performance. We believe this negative effect is due to the fact that participants were already under pressure because of the task: finish the level in the shortest time and with the least number of errors. Thus, as we observed during the experiment, both types of alerting cues induced participants to a more stressed state, causing more frustration. Herbst (2005) highlighted the fact that the position in which vibrotactile feedback is given to the user is important to achieve a good user experience in virtual environments. In a similar way, our experiment highlights the importance of choosing the appropriate frequency for vibrotactile stimuli when used as feedback within desktop AR applications that require precise interaction.

In general, our results of feedback modality analysis are similar to those obtained by Oron-Gilad et al. (2007), where tactile feedback results are poorer than results obtained with visual feedback only in a target finding task in a military context. However, Van Erp and Van Veen (2004) reported that tactile feedback improved performance of drivers when used to alert about a future turn. The difference of both situations, is the level of precision required to accomplish the task in each situation. Vibrotactile feedback have a negative effect in precision tasks, such as acquiring a target or, in our experiment, interacting with virtual objects in a desktop AR application.

Although tactile feedback resulted in poorer performance, participants mentioned that tactile feedback made the game more similar to the real one, and those who did not experience the tactile feedback, asked for it. Thus, in other situations, for instance when realism is important and precise interaction is not, vibrotactile feedback can definitely be useful.

We expected visual alerts after collision events to deviate the attention of participants to the light bulb in the corner of the screen, decreasing their performance. We did not find any statistical evidence to support this, and we observed that the light bulb in the corner of the screen and the text label above the wire path that showed the number of errors simply went unnoticed. Participants were too focused on the virtual objects (the path and the wire loop).

Softer auditory crossing alerts, on the other hand, produced better results than loud auditory and tactile feedback, especially when combined with visual feedback. Auditory and visual feedback complemented each other in an efficient way to provide the crossing alerting cues. Visual feedback was given directly in the region of focus of participants, while the auditory alert was soft enough to not increase the stress level, but yet, accomplished its purpose.

As reported in previous research in other domains (Burke et al., 2006; Prewett et al., 2006, 2012), multimodal interfaces typically produce better performance compared to providing feedback with individual sensory modalities. We can observe a similar trend in our results, however, the only statistically significant difference between multimodal and unimodal feedback was for time outside the wire path, between participants playing with AV feedback and participants playing with T feedback only. We believe the combination of constant vibrations and the demanding task in our experiment affected the performance of participants that received the other combinations of multimodal feedback. With our results, we do not have enough evidence to confirm that vibrotactile feedback affected the performance of participants while they were completing the paths of the game. We can neither confirm that multimodal feedback produced better performance, as the objective measures were not statistically significant different from every individual modality.

We used the typical approach of desktop AR applications, where users observe the augmented scene from a different point of view than their eyes. Thus, our results are specific for this (common) type of desktop AR applications. Using the static camera setup, even with a different point of view, we ensure that participants could perform precise movements without worrying about a change in the point of view. During our experiment, only 9 participants (4.28%) reported having problems with the used point of view. However, all of them could finish the third level of the game, after the training period with the two first levels of the game. One of these participants mentioned *"it was hard to understand how* *to play at the beginning, but once you understand how to move your hand, it is easier"*. Analysing the effect of the point of view in AR is a very interesting topic which is still open for future research.

In our experiment, we decided to focus on the use of vibrotactile feedback, as previous work demonstrated its positive effect as alerting cue in a military context (Lindeman et al., 2005) or in driving assistants (Van Erp and Van Veen, 2004). We employed the Wii remote as the output device, but more complex feedback, such as force feedback can provide a more realistic experience, and thus produce different results. We refer the reader to the work of Unger et al. (2002) for more information about the effect of force feedback in virtual environments.

The design of the experiment aimed to study the effect of multimodal feedback—all the combinations of visual, auditory and tactile modalities—on the user performance. Hence, we only balanced the groups according to the feedback modality. On the other hand, as we did not evaluate any particular cognitive ability that depends on the professional background, the participants we recruited for our experiment are still a valid sample of potential users of AR applications. Nevertheless, future experiments should consider a more heterogeneous sample.

4.4 Chapter Summary

We presented the evaluation of all the combinations of visual, auditory and tactile feedback as alerting cues in a typical desktop Augmented Reality application where users observe the augmented scene from a different point of view than their eyes. We performed the experiment using an AR version of the Wire Loop Game. In total, 210 volunteers participated in our experiment.

Our experiment shows that, for a task where participants need good hand-eye coordination to complete a demanding task, combining visual and auditory feedback produces the best results (reducing reaction times) when soft sounds and visual alerts in the region of focus are combined. The visual attention of the participants was already focused on the virtual elements to avoid collisions and crossings, the crossing alert (the change in the transparency of the ring) was given in that same region and soft auditory alerts efficiently informed the error without being annoying. We believe that our results could be applied to diverse application areas: gaming, navigation or driver assistance systems, industrial maintenance and repair assistants, to cite a few examples. In general, in any AR application where the sight of the user is focused on task related information and require precise interaction with virtual objects.

Designers should pay special attention when choosing sounds and vibration patterns, because they might annoy the user: their frequent use increases the stress in users who are already under pressure because of the demanding nature of the task. The use of auditory alerts should also be evaluated in the context where the application is going to be used, as this type of feedback may not be suitable for quite places such as a library, or too noisy environments where users could have problems to hear it.

CHAPTER 5

A STEREOSCOPIC MULTIMODAL AR GAME

The use of stereoscopic 3D graphics has become more common in different domains. Three examples are: computer games with complex and visually appealing 3D environments, movies that show scenes rendered with stereoscopic images to create a better immersion illusion, and computer-aided design software that helps professionals in different domains. Furthermore, affordable off-the-shelf display hardware to render that content is also popularizing. This hardware can be either televisions, computer screens, projectors or head-mounted-displays. In this last category, we find devices such as Sony HMZ-T2 Personal 3D Viewer, Vuzix Wrap 1200DXAR or Google Glass.

It seems a logical step for mobile and desktop Augmented Reality systems to take advantage of the available hardware to use stereoscopic rendering for giving more realism to the augmented scenes, and improving the depth perception of virtual objects in the real scene.

The previous chapter describes the experiment we performed to find the best combination of visual, auditory, and tactile feedback modalities as alerting cues in an AR game. In that experiment, we used a normal computer screen to show the augmented scene. One interesting question that came out after the analysis of the results is how different display technologies would affect them. We wonder if using a stereoscopic display would improve depth perception of virtual objects, and thus, would improve the overall user performance and enjoyment level of the game.

In this chapter, we analyse the effect of the most common display technologies used in AR: normal screens and Head-Mounted-Displays. We evaluate this effect on user performance when completing a task that requires good hand-eye coordination. Furthermore, we compare the effect of the use of non-stereoscopic and stereoscopic graphics as rendering technique for both display technologies.

To accomplish this goal, we employed a modified version of the AR Wire Loop Game we used for the experiment described in Chapter 4. The game requires the player to perform precise 3D movements to complete each path in the shortest time and with the least number of errors. Hence, perceiving the correct depth relationship between both, the wire path and the wire loop is vital for the game.

We study the effect of combining display technologies and rendering techniques both, from the objective point of view, measuring different performance measures, and from the subjective point of view, assessing user experience when playing the game.

Key contribution - The results of this experiment can help the AR community to understand the role of both display technologies—2D screens and head-mounted-displays—and rendering technology—stereoscopic and non-stereoscopic images—when they are used in applications that require precise interaction between the user and the virtual elements of the scene.

5.1 Experiment Description

We designed the experiment to evaluate the impact of the two most common display technologies for AR (regular 2D screens and head-mounted displays) on the performance of users completing a task that requires good hand-eye coordination and precise 3D movements. We also evaluate whether using stereoscopic rendering improves user performance for both display conditions.

5.1.1 Methodology

The session with every participant started with an explanation about the goal of the experiment and the purpose of the game. Then, we asked participants to fill out a questionnaire with their demographics. To ensure every participant received the same information about the game, we showed them an introductory 3.5 minute video to provide the instructions on how to play it.

The game we used for the experiment has three complexity levels, as we explain in Chapter 4. Participants played Levels 1 and 2 to learn how to control the virtual wire loop using the real wand. We only considered data from the third level for the statistical analysis reported in this chapter.

The goal of the game was to finish each level in the shortest possible time and with the least number of mistakes. Mistakes, as also explained in Chapter 4, occur when participants cause a collision between virtual objects or when participants take the virtual ring out of the path, a *collision* and a *crossing event*, respectively.

For this experiment, all participants played receiving the combination of visual, auditory and tactile feedback. Each subject was randomly assigned to one of the four groups defined by the combinations of the two levels of display technology and rendering technique (i.e. a *between subjects* design):

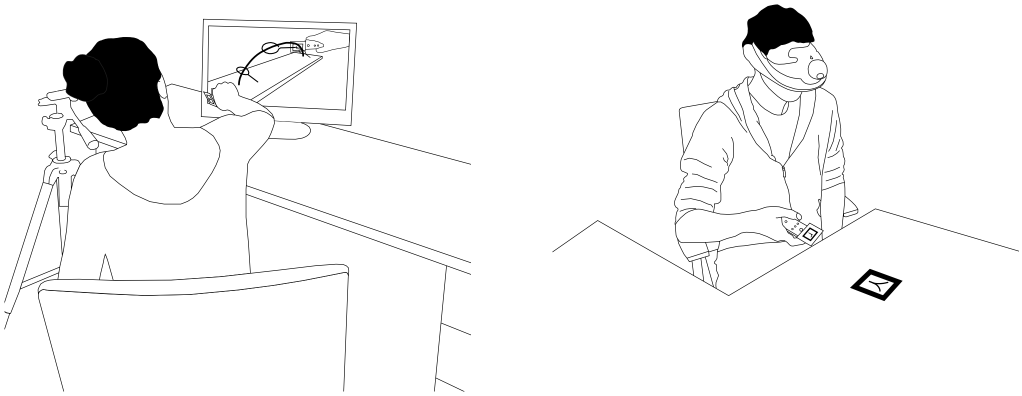
Screen 2D (2DScr) Using a screen without stereoscopic rendering.

Screen 3D (3DScr) Using a screen with stereoscopic rendering.

HMD 2D (2DHMD) Using a Head-Mounted-Display without stereoscopic rendering.

HMD 3D (3DHMD) Using a Head-Mounted-Display with stereoscopic rendering.

Participants using the normal screen played using the same setup explained in Chapter 4 (seated in front of the computer screen observing the virtual objects from the point of view of the static camera), see Figure 5.1a. On the other hand, participants who played using the HMD, observed the augmented scene from the point of view of the camera that was located close to the position of the eyes, attached to the HMD, see Figure 5.1b.



(a) Normal screen.

(b) Head-Mounted-Display.

Figure 5.1: Participants played using either a normal screen or a HMD.

Participants that played with stereoscopic rendering and the head-mounted-display passed a test (before the actual experiment) to evaluate if they could correctly distinguish the relative position of simple geometric forms in 3D: a sphere and a cube. The objective of this pre-experiment evaluation was to discard those participants that could not adjust their eyes to the content showed on the HMD. However, all the participants from the first group of people that came to the experiment were able to complete the task, and thus, participated in the experiment.

During the experiment, we allowed participants to take a one minute break after completing each level. Once all the levels were completed, we asked the participants to fill out a questionnaire and made a few open questions in an interview format.

5.1.2 Participants

We performed this study with 80 university students, 54 men (67.50%) and 26 women (32.50%). They were between 17 and 40 years old (mean of 25 ± 5 years). Only 5 participants were left handed (6.25%). None of the participants had played our game before, but 26 participants (32.50%) had previous experience with an AR application. A little bit more than half of the participants, 46 or 57.50% were frequent gamers, playing more than 5 hours a week, the other 34 participants played between 1 and 5 hours a week (non-frequent gamers).

We randomly divided all participants into the four groups (20 participants per group), as explained before. All of them volunteered their time and they were invited through different mailing lists, fliers or ads posted at the university buildings.

5.1.3 Apparatus

We employed the game described in Section 4.1.3 of Chapter 4, but we did the necessary changes to add stereoscopic rendering capabilities.

The osgART (Looser et al., 2006) library we used to build our game provides all the functionality required to render stereoscopic images using Open Scene Graph. Stereoscopic viewing was achieved using the NVidia GeForce 3D Vision Kit (active stereoscopic glasses and an NVidia Quadro 600 graphics card) and the LG Flatron W23630 monitor for the screen scenario. We used the Sony HMZ-T2 Personal 3D Viewer for the HMD scenario. In both cases, we used the same screen and HMD for both conditions of stereoscopy.

The LG monitor is a 23 inches screen with an aspect ratio of 16:9 and a maximum resolution of 1980 × 1080. Its response time is 5 ms, and it has a maximum brightness of 400 cd/m^2. The Sony HMD has two OLED panels with a resolution of 1280 × 720 also with a 16:9 aspect ratio, and a 45° FOV.

Participants who played using the screen observed the scene from the perspective of a Logitech Quick cam Pro 9000 webcam, that was placed 15 cm above the main marker and 30 cm away, oriented at 45° with respect to the main axis of the path (see Figure 5.1a).

5.2 Statistical Analysis

We studied the data collected for the five performance measures listed below. We refer the reader to Section 4.2 in Chapter 4 (page 57 of this document), for a detailed description of these measures.

- Game scores.

  Completion time (T)

  Number of collisions (N_c)

  Number of crossings (N_x)

- Reaction times.

  Average collision time (T_c)

  Average time outside the wire path (T_o)

For this experiment, we consider that the three game scores are the measures that better characterize the effect of combining display technologies and rendering techniques. These measures can change according to the quality of depth perception and to how the participant perceives the relationship between virtual objects in the 3D space.

On the other hand, the reaction times measured by the Average collision time and Average time outside the wire path can be indirectly affected by the quality of depth perception. Once any type of feedback alerts the participant about the *collision* or *crossing event*, the depth perception can help to solve the problem faster.

We designed our experiment guided by the following research questions:

- Does participants' performance improve when playing the game using stereoscopic rendering? This condition should offer a better depth perception of virtual objects.

- Does participants' performance improve when playing with a HMD? This display, with the webcam attached to it, offers a closer point of view of the scene to the one participants would have with their own eyes, and thus, can make the game easier to play.

We employed the *Multifactor ANOVA* test to compare the effect of the following factors:

- display technology,

- rendering technique,
- gender,
- gaming experience and
- AR previous experience

We did not consider the hand participants used to play the game in our analysis because of the small number of left-handed participants. *Shapiro-Wilk* test of normality and *Bartlett Test of Homogeneity of Variances* showed that our data did not fulfil *ANOVA* prerequisites. Therefore, we used data transformation techniques to be able to use parametric tests in the analysis. In those cases where we found a statistically significant difference, we used the *Tukey HSD* test as *post-hoc* method at the 95% confidence level.

To analyse the results of the questionnaires, we used the *Kruskal-Wallis Rank Sum* test to compare the effect of feedback modality on the response of the participants to *Likert* scale questions and *Mann-Withney U* test to compare the effect of gender, gaming experience and AR previous experience on the response of participants to *Likert* scale questions.

We also provide *effect sizes* for every test to have a measure that is independent from the sample sizes and to have a magnitude of the significant differences found, which is not measured by *p-values*.

5.2.1 Performance Results

Table 5.1 presents the summary of Multifactor ANOVA test for the Completion time (T). We observe statistically significant differences only for rendering technique factor (medium effect size). We did not observe any statistically significant interaction between any of the factors under analysis.

Participants playing with stereoscopic rendering needed on average 41.66 ± 16.09 s to complete Level 3 of the game, while participants playing without stereoscopic rendering needed on average 50.59 ± 17.50 s to complete the same level. Figure 5.2 depicts the box plot and the average completion time both rendering conditions.

Table 5.1: Multifactor ANOVA for completion time (T).

Factor	F	p	Effect Size (*partial* η^2)
Rendering technique	4.02	0.04*	0.063
Display technology	1.52	0.22	0.025
Gender	0.16	0.68	0.002
Gaming Experience	0.70	0.40	0.011
AR previous experience	0.91	0.34	0.015

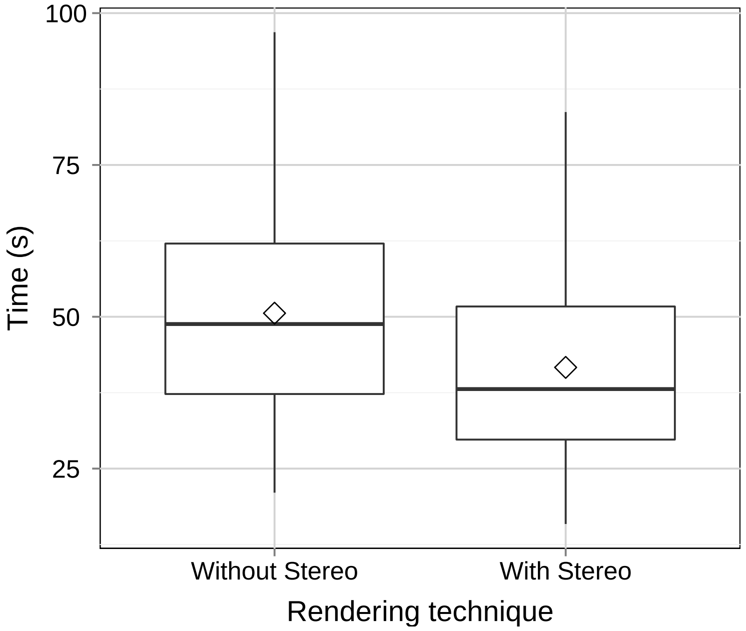


Figure 5.2: Completion time (T) box plot as a function of Rendering technique for Level 3.

The analysis of the Number of collisions (N_c) revealed statistically significant differences for two factors: Stereoscopic rendering (large effect size) and AR previous experience (medium effect size). The analysis did not show any statistically significant interaction between any of the factors. Table 5.2 presents the summary of Multifactor ANOVA test for the N_c.

Once again, participants playing with stereoscopic rendering had better performance, made less mistakes (18.10 ± 9.11 collisions) than participants playing without stereoscopic rendering (28.60 ± 13.36 collisions). Figure 5.3 depicts the box plot and the average number of collisions for both rendering techniques.

On the other hand, participants with previous experience using an AR ap-

Table 5.2: Multifactor ANOVA for number of collisions (N_c).

Factor	F	p	Effect Size (*partial* η^2)
Rendering technology	23.40	1×10^{-5}**	0.284
Display technology	0.59	0.44	0.009
Gender	0.04	0.83	0.001
Gaming Experience	0.01	0.89	0.002
AR previous experience	4.16	0.04*	0.065

plication made less *collision events* (22.31 ± 10.55) than participants without previous experience (23.85 ± 13.44).

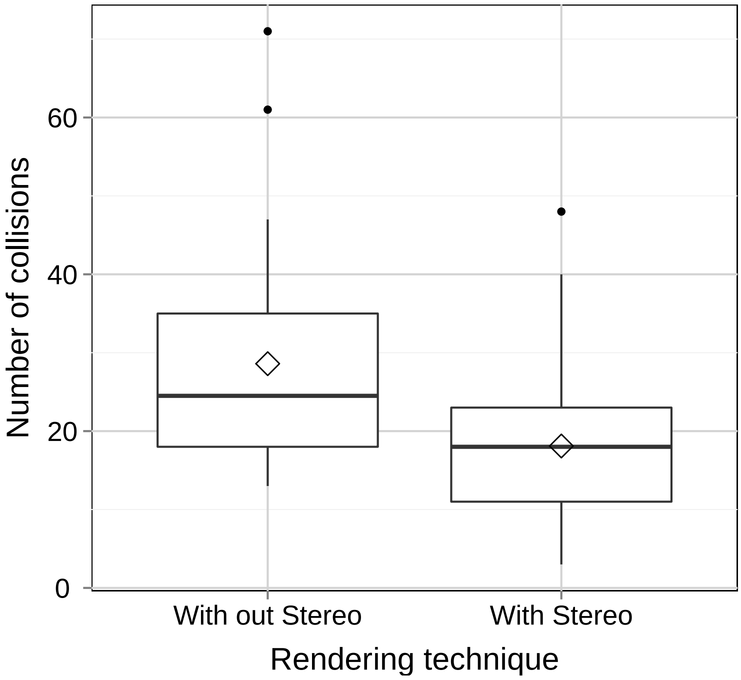


Figure 5.3: Number of Collisions (N_c) box plot as a function of Rendering technique for Level 3.

The Multifactor ANOVA for the Number of Crossings did neither show any statistically significant differences for any of the factors, nor any statistically significant interaction.

Table 5.3 presents the summary of the Multifactor ANOVA test for the Collision time (T_c). We observe statistically significant differences for Rendering technique factor, with a large effect size. However, we also observe a statistically significant interaction between Rendering technique and Display technology factors. Thus we discard the main effect and further analyse the interaction.

Table 5.3: Multifactor ANOVA for collision time (T_c).

Factor	F	p	Effect Size (*partial* η^2)
Rendering technique	37.43	1×10^{-6}**	0.38
Display technology	2.99	0.08	0.04
Gender	1.04	0.31	0.01
Gaming experience	0.16	0.96	0.00
AR previous experience	0.01	0.87	0.00
Rendering tec. : Visualization tec.	19.28	5×10^{-5}**	0.24

The analysis of the interaction revealed statistically significant differences between participants playing with stereoscopy (0.12 ± 0.06 s) and without it (0.02 ± 0.03 s) when they used the normal screen as a display ($F(1, 38) = 76.593$, $p = 1.211e{-}10$**, *generalized* $\eta^2 = 0.6683$). We also observed statistically significant differences between participants playing with stereoscopy (0.07 ± 0.03 s) and without it (0.14 ± 0.05 s) when they used a HMD ($F(1, 38) = 31.025$, $p = 2.2141e - 06$**, *partial* $\eta^2 = 0.4494$). We can observe a large effect size in both cases. Figure 5.4 depicts the interaction plot between both factors.

Table 5.4 presents the summary of the Multifactor ANOVA test for the Time outside the wire path (T_o). We observed statistically significant differences for the Rendering technique, Display technology and AR previous experience factors, with a large effect size in all cases. We did not observe any statistically significant interaction between any of the factors.

Table 5.4: Multifactor ANOVA for crossing time (T_x).

Factor	F	p	Effect Size (*partial* η^2)
Rendering technique	43.79	1×10^{-5}**	0.42
Display technology	12.84	6×10^{-4}**	0.17
Gender	0.20	0.65	0.01
Gaming experience	0.24	0.62	0.01
AR previous experience	10.11	2.3×10^{-3}*	0.14

Participants playing with stereoscopic rendering needed less time to take the ring back into the path after a *crossing event* (0.37 ± 0.38 s) compared to participants who played without stereoscopic rendering (0.60 ± 0.40 s). Figure 5.5

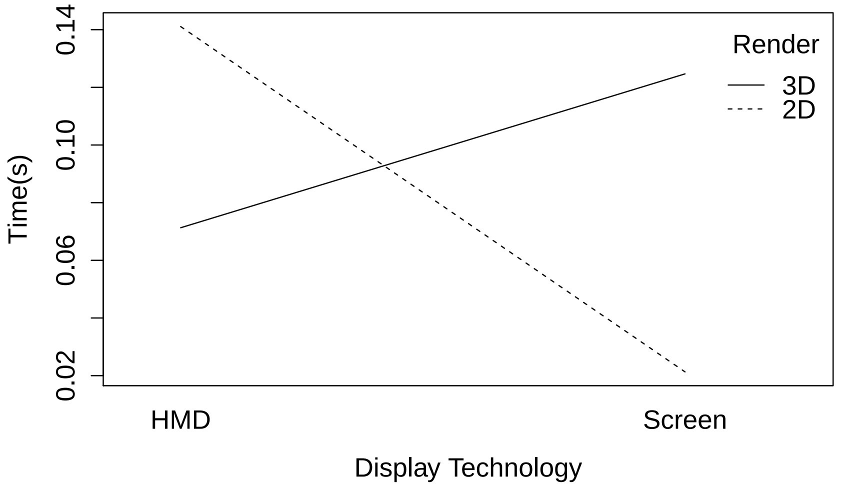


Figure 5.4: Collision Time (T_c) interaction plot for Stereoscopic rendering and Display Technology.

depicts the box plot with the distribution of the results for both rendering techniques.

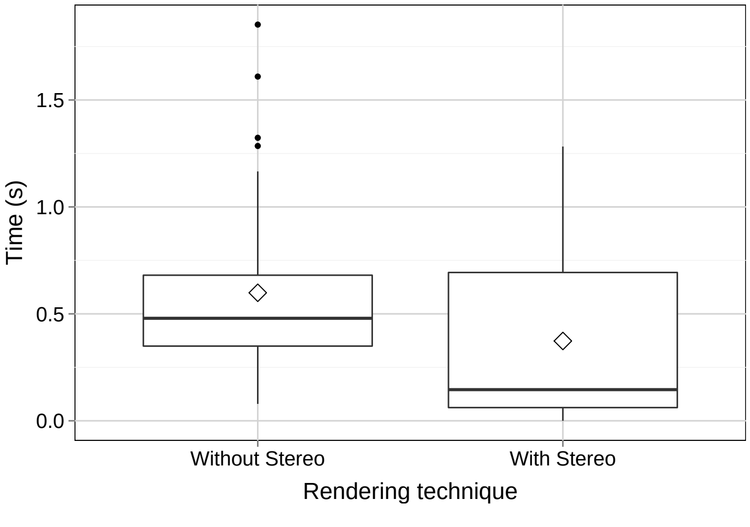


Figure 5.5: Time outside the wire path (T_o) box plot as a function of Rendering technique for Level 3.

On the other hand, participants who played using the screen as a display needed less time (0.35 ± 0.40 s) than participants who played using the HMD (0.62 ± 0.37 s) to correct the trajectory after a *crossing event*. Figure 5.6 depicts the box plot with the distribution of the results for both display technologies.

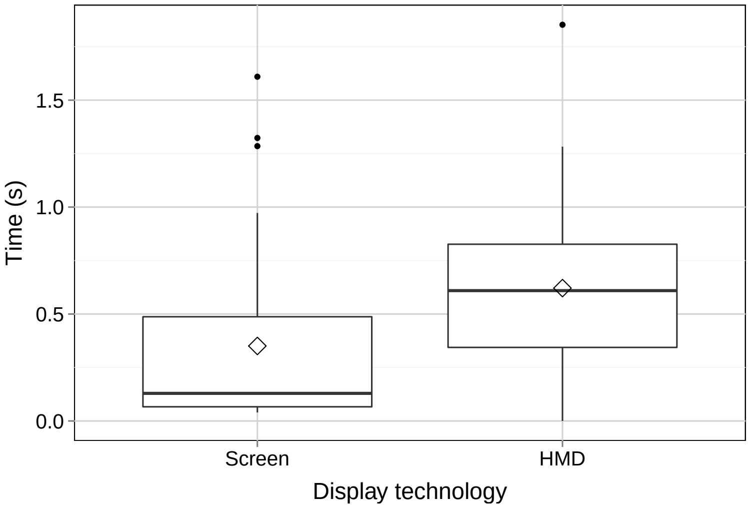


Figure 5.6: Time outside the wire path (T_o) box plot as a function of Display technology for Level 3.

Finally, participants who had previous experience using an AR application needed more time than participants with no previous experience, 0.68 ± 0.28 s versus 0.60 ± 0.31 s respectively.

5.2.2 Questionnaires and Interviews

After the experiment, we also gathered subjective information from participants using questionnaires and interviews. The questionnaires include yes/no, open and *Likert* scale questions.

The list of questions participants answered during the experiment is shown below.

- Yes/no and open questions.
  - Q1 - Have you ever played the real version of this game?
  - Q2 & Q3 - If you have played the real version of this game, which one do you prefer? Why?

  - Q8 - Did stereo rendering helped you to improve your performance in the game?
  - Q9 - Describe your experience playing our Augmented Reality game.
  - Q10 - What would you suggest to improve the game?

- Likert questions.
  - Q4 - How much did you enjoy playing the Augmented Reality game?
  - Q5 - How difficult do you think it was playing the first level of the game?
  - Q6 - How difficult do you think it was playing the second level of the game, compared to the first level?
  - Q7 - How difficult do you think it was playing the third level of the game, compared to the second level?

Question number 8 (Q8) was answered only by those participants who played the game with stereoscopic rendering, either with the screen or the HMD.

From the 80 participants in our experiment, 65 (81.25%) had played the real version of the Wire Loop Game. We asked them to choose between the real version of the game and the AR version they played in our experiment: 69.23% (45 out of the 65 participants) preferred our game to the real version. Statistical analysis revealed statistically significant differences only for the Stereoscopic rendering factor (with a small effect size) (see Table 5.5 rows labelled Q2). The preference for our game was higher between those participants that played with stereoscopic rendering (independently from the display technology): 35 out of 40 participants that played with stereoscopic graphics preferred the AR version of the game, 30 of the 40 who played without stereoscopic rendering had the same preference.

Regarding the enjoyment level of participants playing our game, we found statistically significant differences for Stereoscopic rendering (large effect size) and Gaming experience (medium effect size): rows labelled Q4 in Table 5.5. Further analysis of the Stereoscopic rendering factor showed that participants who played with stereoscopic rendering enjoyed the game more (4 in a 5 points Likert scale) than participants who played without it (2 in a 5 points Likert scale). Regarding

Table 5.5: Statistical analysis for the answers to the Likert scale questions.

<table>
<tr><th colspan="5">Stereo Rendering</th></tr>
<tr><th>Question</th><th>Median</th><th>Statistic</th><th>p-value</th><th>Effect size</th></tr>
<tr><td>Q2</td><td>1</td><td>$U = 402.5$</td><td>0.04*</td><td>$r \leq 0.225$</td></tr>
<tr><td>Q4</td><td>4</td><td>$U = 266.0$</td><td>1.13×10^{-7}*</td><td>$r \leq 0.593$</td></tr>
<tr><td>Q5</td><td>3</td><td>$U = 928.0$</td><td>0.20</td><td>$r \leq 0.142$</td></tr>
<tr><td>Q6</td><td>3</td><td>$U = 504.5$</td><td>0.002*</td><td>$r \leq 0.332$</td></tr>
<tr><td>Q7</td><td>4</td><td>$U = 994.5$</td><td>0.04*</td><td>$r \leq 0.224$</td></tr>
<tr><th colspan="5">Display Technology</th></tr>
<tr><th>Question</th><th>Median</th><th>Statistic</th><th>p-value</th><th>Effect size</th></tr>
<tr><td>Q2</td><td>1</td><td>$U = 555.5$</td><td>0.65</td><td>$r \leq 0.050$</td></tr>
<tr><td>Q4</td><td>4</td><td>$U = 748.0$</td><td>0.60</td><td>$r \leq 0.057$</td></tr>
<tr><td>Q5</td><td>3</td><td>$U = 863.5$</td><td>0.52</td><td>$r \leq 0.070$</td></tr>
<tr><td>Q6</td><td>3</td><td>$U = 905.5$</td><td>0.29</td><td>$r \leq 0.118$</td></tr>
<tr><td>Q7</td><td>4</td><td>$U = 643.5$</td><td>0.10</td><td>$r \leq 0.180$</td></tr>
<tr><td>Q8</td><td>1</td><td>$U = 224.0$</td><td>0.39</td><td>$r \leq 0.096$</td></tr>
<tr><th colspan="5">Gender</th></tr>
<tr><th>Question</th><th>Median</th><th>Statistic</th><th>p-value</th><th>Effect size</th></tr>
<tr><td>Q2</td><td>1</td><td>$U = 399.5$</td><td>0.50</td><td>$r \leq 0.075$</td></tr>
<tr><td>Q4</td><td>4</td><td>$U = 694.0$</td><td>0.94</td><td>$r \leq 0.009$</td></tr>
<tr><td>Q5</td><td>3</td><td>$U = 545.0$</td><td>0.09</td><td>$r \leq 0.186$</td></tr>
<tr><td>Q6</td><td>3</td><td>$U = 592.5$</td><td>0.24</td><td>$r \leq 0.131$</td></tr>
<tr><td>Q7</td><td>4</td><td>$U = 692.0$</td><td>0.92</td><td>$r \leq 0.012$</td></tr>
<tr><td>Q8</td><td>1</td><td>$U = 199.5$</td><td>0.89</td><td>$r \leq 0.016$</td></tr>
<tr><th colspan="5">Gaming experience effect</th></tr>
<tr><th>Question</th><th>Median</th><th>Statistic</th><th>p-value</th><th>Effect size</th></tr>
<tr><td>Q2</td><td>1</td><td>$U = 438.0$</td><td>0.21</td><td>$r \leq 0.139$</td></tr>
<tr><td>Q4</td><td>4</td><td>$U = 499.0$</td><td>0.004**</td><td>$r \leq 0.318$</td></tr>
<tr><td>Q5</td><td>3</td><td>$U = 884.5$</td><td>0.30</td><td>$r \leq 0.115$</td></tr>
<tr><td>Q6</td><td>3</td><td>$U = 553.0$</td><td>0.02*</td><td>$r \leq 0.260$</td></tr>
<tr><td>Q7</td><td>4</td><td>$U = 892.0$</td><td>0.25</td><td>$r \leq 0.128$</td></tr>
<tr><td>Q8</td><td>1</td><td>$U = 107.0$</td><td>0.35</td><td>$r \leq 0.105$</td></tr>
<tr><th colspan="5">AR previous knowledge effect</th></tr>
<tr><th>Question</th><th>Median</th><th>Statistic</th><th>p-value</th><th>Effect size</th></tr>
<tr><td>Q2</td><td>1</td><td>$U = 513.0$</td><td>0.49</td><td>$r \leq 0.077$</td></tr>
<tr><td>Q4</td><td>4</td><td>$U = 750.0$</td><td>0.61</td><td>$r \leq 0.056$</td></tr>
<tr><td>Q5</td><td>3</td><td>$U = 535.0$</td><td>0.72</td><td>$r \leq 0.039$</td></tr>
<tr><td>Q6</td><td>3</td><td>$U = 508.5$</td><td>0.03*</td><td>$r \leq 0.232$</td></tr>
<tr><td>Q7</td><td>4</td><td>$U = 605.0$</td><td>0.28</td><td>$r \leq 0.119$</td></tr>
<tr><td>Q8</td><td>1</td><td>$U = 165.0$</td><td>0.92</td><td>$r \leq 0.013$</td></tr>
</table>

the analysis of the Gaming Experience factor, non-frequent gamers enjoyed the game less than frequent-gamers, 2 and 4 in the 5 points Likert scale respectively.

As for the difficulty of playing the game (rows Q5, Q6 and Q7 on Table 5.5), there were statistically significant differences for Stereoscopic rendering (Levels 2 and 3), Gaming Experience (Level 2), and previous knowledge of AR factors (Level 2). Participants playing with stereoscopic rendering found the second and third levels of the game easier than participants playing without it. The same difference exists between frequent gamers and non-frequent gamers. The former group found the second level easier than participants of the latter group. Participants with previous experience using an AR application found the second level of the game easier than participants without previous experience.

Regarding the depth perception during the game; 10 participants explicitly mentioned that it got better as they played each one of the levels: 3 of these participants played using the screen without stereoscopic rendering, 3 with the screen and stereoscopic rendering and 4 with the HMD and stereoscopic rendering.

Only 5 participants reported problems to correctly perceive the relative depth between virtual objects: 3 played using the HMD and stereoscopic rendering, 1 played using the HMD without stereoscopic rendering, and 1 played using the normal screen with stereoscopic rendering.

From the group of 40 participants that played using the HMD, only 4 (10%) participants reported that it was uncomfortable to use because of its weight and made them sick after playing the third level of the game. Another issue with the HMD, as reported by 3 (7.5%) participants, was that trying to keep the head steady to avoid unintended movements of the virtual objects made the game more difficult.

From the entire group of participants, 2 asked to include feedback to guide their hand in case of making too many mistakes. Participants also mentioned that constant vibrations were uncomfortable. Only 1 participant explicitly mentioned that observing virtual objects appear all the time over his hand was confusing.

When we asked participants for suggestions on how to improve the game, just as in the case of the experiment described in Chapter 4, participants of this experiment asked for more levels with higher complexity for the game.

5.3 Discussion

Using stereoscopic rendering in application domains such as video games and movies can improve depth perception and increase immersion feeling. On the other hand, most AR applications use a computer screen or TVs to show the augmented scene, with the user observing the scene from the point of view of a camera in some position different than their eyes. However, the appearance of affordable HMD highlight the need to evaluate the benefits of such display technology for future AR applications design.

We designed this experiment to evaluate the effect of the two most common display technologies used in AR on user performance. Participants were asked to complete an spatial task that requires precise 3D movements. Furthermore, we evaluate if the use of stereoscopic rendering and HMDs represent an advantage and eventually a worthwhile investment for new AR applications.

Participants playing with stereoscopic rendering had better performance than participants playing without it in almost all the measures we analysed. Stereoscopic rendering helped participants complete Level 3 of the game in less time and with less *collision events*. Participants playing with stereoscopy also required less time to correct the trajectory after a *crossing event*. Stereoscopic rendering also increased the number of participants that preferred our game instead of the real version of the game. It also increased the reported enjoyment level of playing the game with a 4 in a 5-points Likert scale (where 5 is the highest enjoyment level). Thus, we can give a positive answer to our first research question because, in general terms, the use of stereoscopic rendering improved the performance and user experience of participants, both, objectively and subjectively.

However, designers of AR stereoscopic applications should be careful. Even when performance and user experience was better with stereoscopic rendering, participants reported that the augmented scene looked weird to them because of the virtual objects being over their hand all the time. Other participants had problems accommodating their vision to the 3D objects, and even when they were able to play the game correctly, they reported eye strain after playing. All these perception issues are common to other AR applications, and have also been reported before, for example by Kruijff et al. (2010).

On the other hand, participants who played without stereoscopic rendering needed less time to correct their trajectory after a *collision event* when they played with the screen, but needed less time when they played with stereoscopic rendering using an HMD. Participants playing with the screen needed less time to correct their trajectory after a *crossing event.* But the analysis did not reveal more statistically significant differences for the display technology factor, neither for the rest of performance measures nor for the user experience evaluation through the questionnaires. The data collected through this experiment is not enough to prove if user performance improves when playing the game using the HMD (our second research question).

However, participants who played with the HMD, mentioned that it was heavy and tiring after a while. Participants also mentioned that it was somehow difficult to keep the head steady and at the same time, moving the hand to control the virtual ring in order to avoid extra *collision events.* While these are two important factors to consider in future designs, the first will, for sure, be tackled as the HMD technology improves. Today HMDs such as Oculus Rift (Oculus VR., 2014), are lighter and more affordable than its predecessors, such as the Sony HMZ-T2 we used in this experiment. On the other hand, the second aspect is more important to take into account for future designs. The use of HMD for applications such as the one we have evaluated might not be the best choice. If the task in which the AR application is assisting or involving the user requires good hand-eye coordination and precision, the application should consider stability strategies to avoid normal head movement affecting the user's performance and experience. Forcing users to keep their head steady can cause uncomfortable postures and the rejection of the application.

In this experiment, gaming experience did not affect user performance for any of the factors we measured. Gaming experience analysis showed statistically significant differences only in the case of Enjoyment Level, where frequent gamers enjoyed the game more than non-frequent gamers. One of the reasons for this difference between frequent gamers and non-frequent gamers is related to the perceived level of difficulty. We found statistically significant differences between both groups on the perceived difficulty degree of Levels 2 and 3. Both levels were easier for frequent gamers.

Previous experience using AR applications helped participants to need less time to correct their trajectory after a *crossing event*. However, this factor did not affect user performance according to the analysis of the rest of measures. Previous experience using an AR application also affects the perceived difficulty degree of Level 3. Even when the difference in the perceived difficulty level is small, it is surprising that participants without previous experience with an AR application found it less difficult. Unfortunately, we do not have enough information to give a good explanation to this fact.

Finally, we did not observe any statistically significant differences for any of the performance measures neither the user experience questions for the Gender factor. Performance and experience playing the game for participants of both groups of participants was similar in our experiment.

We believe that our results are, of course, application dependent. We would like to point out the fact that using a HMD with the camera attached to it, might not be the best option for applications that require precision in order to interact with the virtual objects, because normal head movements can interfere with user experience and affect their performance, when it causes changes in the point of view of the augmented scene. Using an different point of view through a normal screen, like we did in this experiment, can be a good alternative (and a cheaper one), as we did not find any statistically significant differences between performance of participants playing with both display technologies. Besides the fact that observing the scene from a different point of view it is widely used in professional domains, such as laparoscopic surgery.

5.4 Chapter Summary

Summarizing, stereoscopic rendering improves user performance compared to participants playing without it. However, the fact that participants mentioned eye strain and a strange appearance of the AR scene because of virtual objects appearing on top of real objects all the time deserves attention in future designs.

On the other hand, display technology did neither affect user performance nor user experience in our experiment. Hence, designing desktop AR applications using a normal screen and a webcam can be still considered a good option.

Continuing with the evaluation of multimodal user interfaces, another important Augmented Reality application domain is mobile computing. In the next chapter, we compare different the effect of visual, auditory and tactile feedback modalities, and their combinations, in a pedestrian navigation assistant.

CHAPTER 6

A MULTIMODAL NAVIGATION ASSISTANT

Smartphones combine, among other features, a high resolution display with multitouch capabilities, different tracking sensors, permanent Internet connection and good processing power. The widespread usage of these devices has created new opportunities for Human-Computer Interaction researchers for developing novel types of applications and novel interaction techniques. One of the main challenges in this area is that smartphones are used, most of the time, in highly demanding environments for users' attention, especially when they are on the move (Jameson, 2002; Oulasvirta et al., 2005). For that reason, developing mobile applications that implement efficient user interaction is very important. In particular, pedestrian navigation assistants have to (1) provide robust guiding cues and (2) avoid distracting users by letting them to keep an eye on the road or socially interact with friends.

Different studies in this particular application domain have investigated how to use auditory, tactile and visual cues to improve how pedestrians receive navigation assistance (Liljedahl et al., 2012; Pielot and Boll, 2010). Multimodal feedback offers many benefits for navigation assistance: eyes-free operation, language independence, faster decision-making, and reduced cognitive load (Jacob et al., 2011). Multimodal feedback is also appreciated by the users (Pielot et al., 2012b):

complementary tactile feedback was used in one third of the logged routes where users had the option to receive only visual feedback on a digital map similar to Google Maps. However, as also mentioned in many of those works, there is still room for analysis of the properties of multimodal feedback and thus, improvement of the mobile interfaces. We are especially interested in the scenario where visual feedback is presented using the smartphone's screen as a video see-through display where the user can observe both the road and the assistant's directions.

We developed a pedestrian navigation assistant for smartphones named *LeadMe*. It gives directions to find buildings at the university campus. Visual direction cues were presented as virtual arrows or paths superimposed on the rear camera feed, showing an augmented view of the path as depicted by Figure 6.1. Auditory and tactile direction cues were presented as spoken directions and vibration patterns, respectively. Based on the results of Vainio (2009), we designed our system to provide continuous rhythmic feedback to guide the user in the navigation task.

Key contribution - The main contribution of this experiment is to study the scenario where visual directional cues are presented using the smartphone's screen as a video see-through display where the user can observe both the surroundings and the visual feedback. We build our multimodal guidance system upon the findings of previous research, and evaluate how using the augmented view of the path to provide visual directional cues affects the experience and performance of the user in our mobile application.

6.1 Experiment Description

The domain of mobile applications are another important application area for Augmented Reality. In this context, besides gaming, navigation is one of the most common tasks that can be improved by augmenting the users' perception, especially considering that outdoor environments, in which most of the mobile applications are used, are highly demanding for users' attention. We designed this experiment considering an outdoor task—completing a predefined route on the university premises—to study the effects of multimodal feedback when Aug-

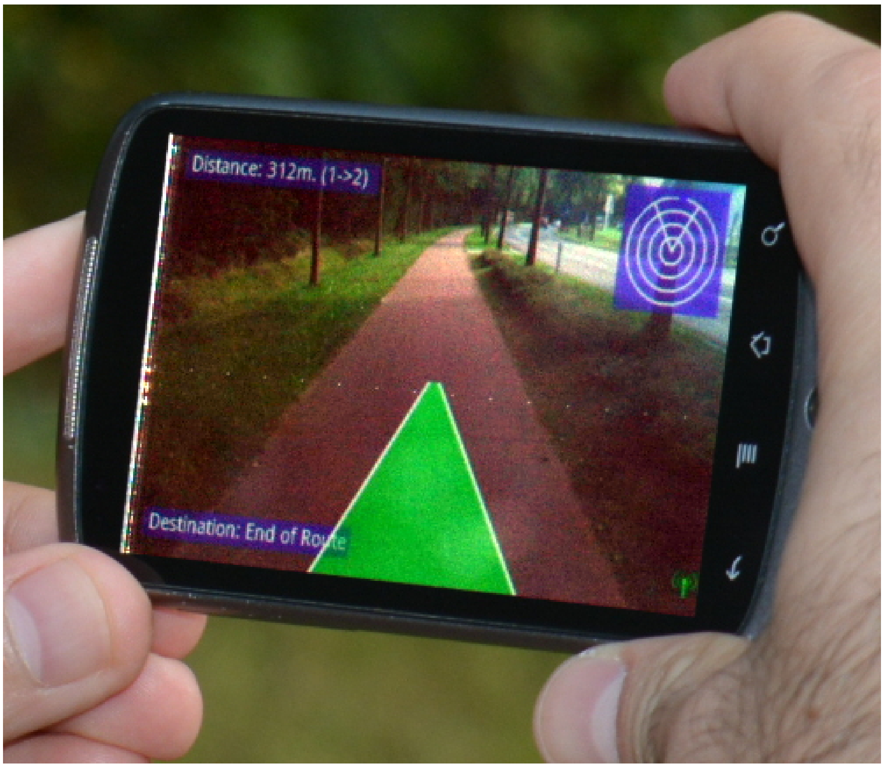


Figure 6.1: Visual directional cues augment the video feed of the rear camera on the smartphone's display, letting the user perceive both the environment and the guidance at the same time.

mented Reality is used beyond the spatial limitations of a desktop.

The aim of our experiment was to understand how multimodal directional information is perceived by users of a mobile navigation assistant for pedestrians. Directional cues are provided through visual, auditory and tactile channels.

6.1.1 Methodology

Every participant in the experiment started the trial reading the written instructions about how to interpret the information given by the application. They also filled out a questionnaire with their demographics.

Then, a member of our team confirmed that the instructions were understood, and gave some final explanation when it was required. Participants were also instructed about the purpose of the experiment: follow the directions provided by the system as accurately as possible. They were kindly asked not to stop during the trial for any reason that was not related to the experiment, such as talking to a friend.

Participants were also explicitly instructed to keep the phone pointing along their way to avoid noise in the measurements. We did not observe any miss-pointing problems and we did not have to discard any sample.

When they were ready to start, participants had to select the destination from a menu option. In this case, there was only one available destination to choose from. Then, participants had to start the route, followed by one member of our team from an approximate distance of 5 m. No interventions were required during the whole experiment.

When participants reached the destination point at the main entrance of one of the university buildings, they had to touch the screen of the smartphone two times—two taps—in order to confirm the reception of the end-of-route message. After completing the path, participants answered a written questionnaire regarding their experience.

Our application is able to show visual elements either in portrait or landscape modes. However, we performed a pilot experiment with members of our research institute to evaluate both layouts. The 10 participants of the pilot experiment preferred holding the smartphone horizontally, in landscape mode. One of the comments was that "the text and the rest of the content was better viewed this way". For this reason, we modified the application to only show the content in landscape mode, and not responding to the smartphone's orientation changes.

We performed the experiment with first year students during their first month at the university. We did not mention the destination building to avoid biasing the results. We also performed the experiments during class hours, to avoid revealing the route in advance.

Participants were randomly assigned to one of the following groups, determining what combination of feedback they received during the experiment (*between subject* design):

Group 1 (A) Auditory feedback only.

Group 2 (T) Tactile feedback only.

Group 3 (V) Visual feedback only.

Group 4 (AT) Auditory and tactile feedback.

Group 5 (AV) Auditory and visual feedback.

Group 6 (TV) Tactile and visual feedback.

Group 7 (ATV) Auditory, tactile and visual feedback.

6.1.2 Participants

There were 77 participants in our experiment, 54 men and 23 women. All of them first year University students. They were between 17 and 24 years old (19.03±1.71 in average). Only 11 participants (14.29%) had used an AR application before; 31 participants (40.26%) had used a GPS before, and 23 participants (29.87%) owned a smartphone.

6.1.3 Apparatus

After evaluating the available options for developing AR mobile applications, we decided to develop our pedestrian navigation assistant for a smartphone running Android OS. The application has the same purpose as a commercial GPS system: providing direction cues to reach a desired destination, but with subtle differences that are explained in this section.

We use the digital compass, the GPS and the built-in camera of a smartphone with a microprocessor running at 1 GHz. It has a 3.7″ display (480 × 800 pixels resolution), a 5 MP rear camera that is able to record video at 24 fps at a maximum resolution of 640 × 480 pixels.

We constructed a virtual map of one portion of the university and designed a path to follow during our experiment. It is 155 m long, and it has 6 checkpoints where the system provides directional cues: 3 left turns, 2 right turns, and the destination point. We carefully chose the path to avoid zones which may cause tracking issues because of buildings or trees blocking the GPS signal. Figure 6.2 shows the map of the route: it shows 7 possible buildings between the starting and ending point of our experiment, and all the alternative routes to follow from one point to another.

Battery consumption was not an issue for our experiment, hence the screen was turned on during the entire experiment, we collected GPS data every 200 ms and requested the orientation data from the digital compass as fast as possible, using the constant SENSOR_DELAY_FASTEST[1] to set-up the compass sensor manager.

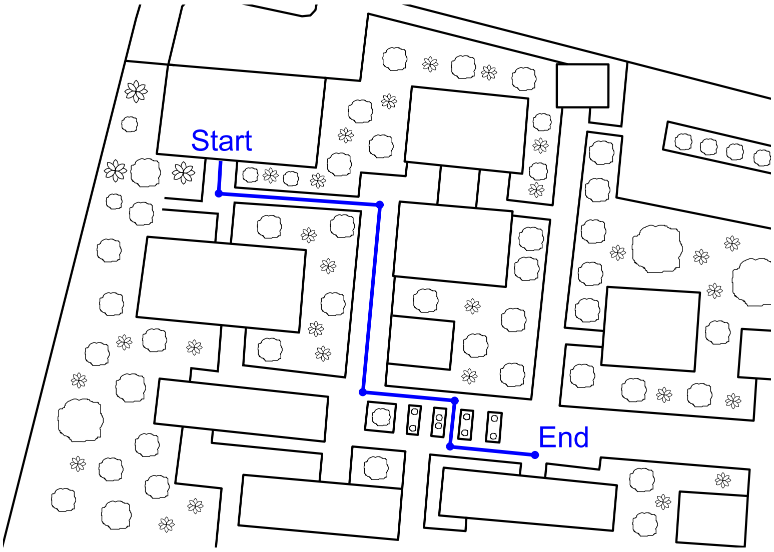


Figure 6.2: The route for the experiment.

The standard interface of the application includes four basic elements: the video stream from the mobile phone's camera, the application menu (on demand), the destination and a distance to destination labels. These elements are always visible, regardless of the feedback combination under evaluation.

During the execution, the application can be in one of four states: *(i)* the participant points the phone in the right direction, *(ii)* the participant has to turn, *(iii)* the participant has missed a *checkpoint* and has to go back, and *(iv)* the participant has reached the final destination. We define a checkpoint as a place on the path where the participant has to make a decision: turn left or right, keep walking straight ahead, or confirm the arrival at destination.

Figure 6.3 shows the transitions that might occur between the states, and thus how the feedback will be given to the participant. For example, consider the following scenario: when the participant approaches a point where she has

[1] http://developer.android.com/reference/android/hardware/SensorManager.html

to turn left, the application passes from *state 1* to *state 2*; if she misses the turn, and keeps walking, there is a transition *to state 3*. When the participant walks back pointing the mobile phone in the direction of the checkpoint, there is a transition to *state 1*, and when the missed checkpoint is reached, *state 2* is activated again to point the participant in the right direction. Finally, when the participant reaches the destination, the system presents a message and stops providing feedback.

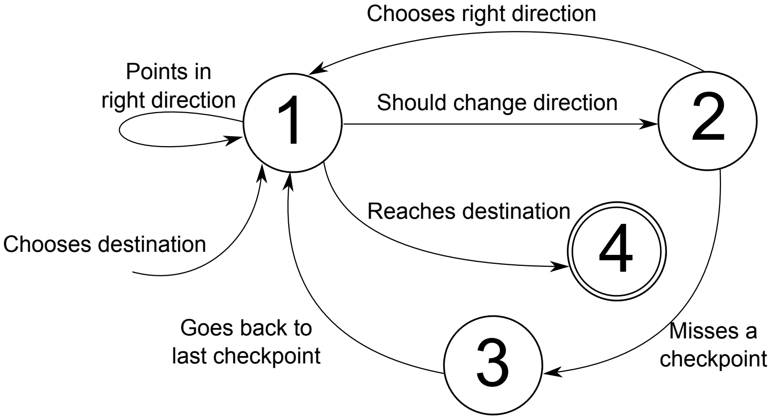


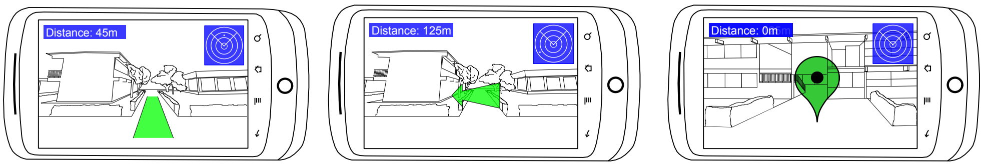
Figure 6.3: Transitions between the four states of the application.

The following sections describe the different types of feedback provided during the experiment.

6.1.3.1 Visual Feedback

When visual feedback is active, the participant is presented with these elements: a virtual path drawn on the camera's view when the participant points the mobile phone in the right direction towards the next checkpoint, an arrow pointing to the correct direction when the participant has to turn or has deviated from the right path, and a radar image that codes direction and distance to target.

Figure 6.4a depicts the path and the radar elements when the user is pointing the phone in the right direction (*State 1*). Figure 6.4b shows how the user sees the turning signal in the camera's view (*State 2*). Visual feedback in case the participant missed a checkpoint and needs to go back is represented with an arrow pointing backwards.



(a) Go ahead (b) Turn (c) Route completed

Figure 6.4: The Visual Feedback in *LeadMe*

When the subject completes the path and reaches the destination, a visual signal is shown on the screen to indicate that the task has finished (see Figure 6.4c).

We employed OpenGL ES to draw the virtual elements of our application on top of the camera's video feed.

6.1.3.2 Auditory Feedback

In the first state, a synthetic voice informs the participant to continue in that direction: "Walk straight ahead". When the participant has to turn left or right, the synthetic voice indicates so: "Turn left" or "Turn right". "You missed a checkpoint and you need to go back" is the spoken message when the participant has missed a checkpoint. Finally, when the participant reaches the destination, the voice utters the message: "You have reached your destination".

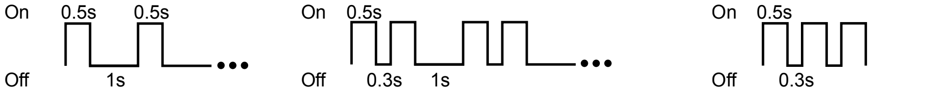
Each one of the voice directions is repeated once, when a state change occurs. However, participants could ask for the repetition of the last message selecting an option from the menu. We used the Text-To-Speech features of the Android library to generate the voice instructions.

6.1.3.3 Tactile Feedback

Tactile feedback is provided by means of the vibrator of the mobile phone using the magic wand metaphor (Fröhlich et al., 2011), that requires the participant pointing the mobile phone in the right direction to receive the directional information. This metaphor has been used before in way finding tasks with good results (Pielot et al., 2012a; Raisamo et al., 2012).

We follow a similar approach to the one proposed by Raisamo et al. (2012), changing the vibration pattern at a checkpoint to attract participant's attention. Our system also required participants to turn the phone scanning the space around them to find the new direction. Our system uses the following vibration patterns:

- when the participant is on the right path, a repetition of 0.5 s vibration pulses separated by 1 s periods with no vibration (see Figure 6.5a),
- an approaching turn is represented as a sequence of 0.5 s vibration, 0.3 s with no vibration, 0.5 s vibration and 1 s with no vibration (see Figure 6.5b),
- when the destination is reached, three 0.5 s vibration pulses, separated by 0.3 s with no vibration as a confirmation, and then the application stops producing feedback (see Figure 6.5c).



(a) Go ahead (b) Turn (c) Route completed

Figure 6.5: The Tactile Feedback in *LeadMe*

6.2 Statistical Analysis

We evaluated the effect of multimodal feedback for pedestrian navigation assistance guided by the following research questions:

- How useful visual feedback is—the augmented view—to provide directional cues? This type of feedback typically requires to deviate visual attention from the path to the screen and can also be affected by visibility problems, caused for example by sun glares.
- How useful is auditory feedback? In a real-world situation, this type of feedback can be missed because of noise. However it also represents an intuitive way to give direction cues, and can be used when vision is not available.

- How useful and intuitive is tactile feedback to provide direction cues?
- Does the combination of feedback modalities improve user performance?

We employed the *Multifactor ANOVA* test to compare the effect of the following measures on participants performance.

- feedback modality,
- gender and
- GPS navigation previous experience.

We also quantified participants' performance according to:

- Time to complete the route (T): Time required to complete the route, from the moment the participant chooses the destination, to the confirmation of the end-of-route signal.
- Time to confirm the reception of the end-of-route signal (T_e): Time a participant takes to acknowledge the end of route signal.
- Average number of missed checkpoints (N_{mc}): Number of times a participant missed a checkpoint where she was supposed to turn. A checkpoint is considered missed when the participant reaches the checkpoint but does not turn as indicated and continues walking in the wrong direction. We set the threshold distance to 5 m from the geographical coordinate of the checkpoint.

Shapiro-Wilk test of normality and *Bartlett Test of Homogeneity of Variances* showed that our data did not fulfil *ANOVA* prerequisites. Therefore, we used data transformation techniques to be able to use parametric tests in the analysis. In those cases where we found a statistically significant difference, we used the *Tukey HSD* test as *post-hoc* method at the 95% confidence level.

To analyse the results of the questionnaires, we used the *Kruskal-Wallis Rank Sum* test to compare the effect of feedback modality on the response of the participants to *Likert* scale questions and *Wilcoxon rank sum test with continuity* *correction* to compare the effect of gender, previous experience using a GPS on the response of the participants to *Likert* scale questions. We also provide *effect sizes* for every test to have a measure that is independent from the sample sizes and to have a magnitude of the significant differences found, which is not measured by *p-values.*

The questionnaires included yes/no, open, multiple choice and *Likert* scale questions (from 1 to 5, the lowest and the highest score for each question respectively).

The list of questions asked after the experiment is presented below. The questionnaire included a general section for every participant in the experiment, but it also included specific questions regarding visual, auditory and tactile feedback for those participants who completed the path receiving them.

- Yes/no and open questions.
  - Q3 - Did you find auditory feedback annoying?
  - Q4 - Do you think the auditory feedback was insufficient to complete the route?
  - Q5 - Do you think the frequency of the auditory feedback was enough?
  - Q7 - Do you have any comment regarding auditory feedback?
  - Q8 - Did you find tactile feedback annoying?
  - Q9 - Do you think the tactile feedback was insufficient to complete the route?
  - Q10 - Do you think the frequency of the tactile feedback was enough?
  - Q12 - Do you have any comment regarding tactile feedback?
  - Q13 - Did you find visual feedback annoying?
  - Q14 - Do you think the visual feedback was insufficient to complete the route?
  - Q15 - Do you think the frequency of the visual feedback was enough?
  - Q17 - Did you have problems to see the elements on the screen?
  - Q18 - Do you have any comment regarding visual feedback?

  - Q19 - Do you have any comments regarding *LeadMe*?
  - Q20 - Do you think that using more than one feedback type improves the way the application gives information?
  - Q22 - Did the sun glares bother you?

- Likert questions.
  - Q1 - How much did you enjoy using *LeadMe*?
  - Q2 - How difficult do you think it was using *LeadMe*?
  - Q6 - How helpful do you think auditory feedback was?
  - Q11 - How helpful do you think tactile feedback was?
  - Q16 - How helpful do you think visual feedback was?
  - Q21 - How difficult were to understand the directions to complete the route?

- Multiple choice.
  - Q23 - In which of the following situations would you use *LeadMe*?

6.2.1 Performance Results

The Multifactor ANOVA revealed that only the Feedback modality factor has a statistically significant effect on Time to complete the route (T) with a large effect size. Table 6.1 shows the summary of the analysis. There was not any statistically significant interaction among any of the factors under analysis with respect to T.

Table 6.1: Multifactor ANOVA for completion time (T).

Factor	$d.f.$	F	p	Effect Size ($partial\ \eta^2$)
Feedback modality	6	13.52	1.0×10^{-3}**	0.609
Gender	1	3.37	0.07	0.061
GPS Previous experience	1	4.3×10^{-3}	0.94	0.000

The post-hoc analysis for Feedback modality factor and T revealed statistically significant differences between Tactile feedback and the rest of the groups.

Table 6.2 depicts statistically significant differences with an * mark for the effect of the Feedback modality. Figure 6.6 shows the box plot for T for every feedback group.

Table 6.2: Tukey HSD Post-Hoc Analysis for Feedback Modality Factor on T.

	A	T	V	AT	AV	TV	Average (s.)
A	-						159.53 ± 18.12
T	*	-					232.26 ± 35.14
V	-	*	-				163.14 ± 13.09
AT	-	*	-	-			173.37 ± 19.31
AV	-	*	-	-	-		153.20 ± 16.38
TV	-	*	-	-	-	-	169.69 ± 19.77
ATV	-	*	-	-	-	-	148.16 ± 12.27

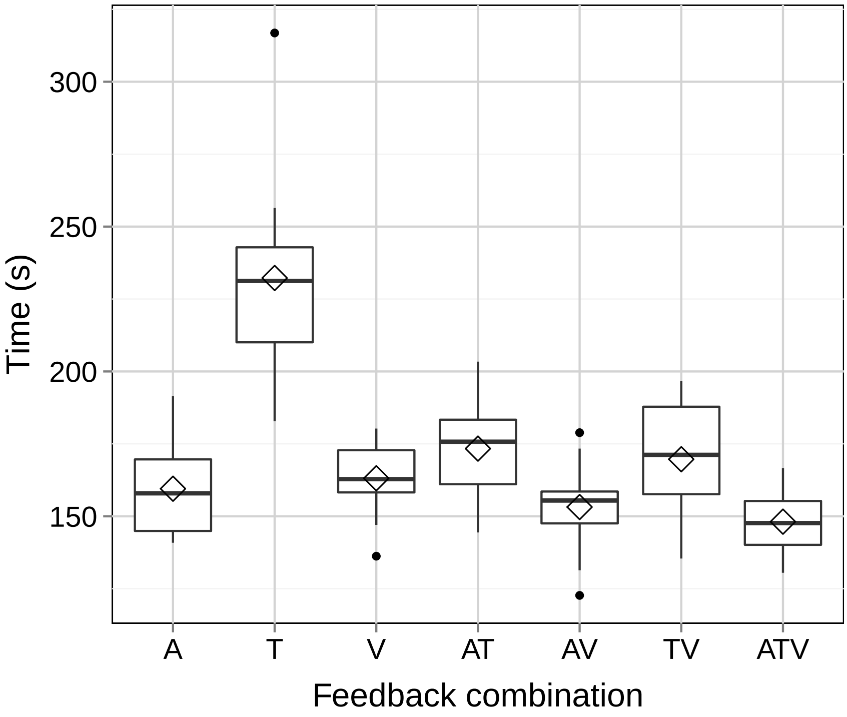


Figure 6.6: Time to complete the route box plot (T).

The Multifactor ANOVA did not show any statistically significant differences regarding the Time to confirm the reception of the end-of-route message. Finally, no participant missed a checkpoint, and thus, there was no data to compare the Average number of missed checkpoints.

6.2.2 Questionnaires and Interviews

When asked how much participants liked the application, 59 out of the 77 (76.62%) chose the level 5 in the Likert scale; 15 (19.48%) chose level 4, and 3 (3.90%) participants chose level 3. Neither the *Kruskal-Wallis Rank Sum* of Feedback modality nor the *Wilcoxon rank sum test with continuity correction* for Gender and Previous Experience using a GPS showed any statistically significant effect on the degree to which participants liked or disliked our application.

Regarding how easy to use the application was, 52 participants (67.53%) rated it with a 5 in the Likert scale; 9 participants (11.69%) with a 4, 11 participants (14.29%) with a 3, 4 participants (5.19%) with a 2 and only one participant (1.30%) with a 1. In this case, only the Gender factor had a statistically significant effect on the measure (Wilcoxon rank sum test with continuity correction $W = 452.5$, $p = 0.02$). Women rated the easiness level with a median of 4 ($IQR = 2$) in the Likert scale, while men rated the easiness of using the application with a median of 5 ($IQR = 0$).

The sun glares caused problems when reading the messages on the screen for 41 participants (53.25%). We should take into account that even participants of groups without visual feedback had the distance to target and the destination labels on the screen. None of the factors we analysed had a statistically significant effect on the number of participants that reported problems with sun glares. On the other hand, all the participants believed that using more than one feedback modality would improve the effectiveness of the system.

None of the participants that completed the path receiving auditory feedback (A, AT, AV or ATV groups) considered it annoying. They also reported having all the information they needed to complete the route. The frequency of auditory feedback was appropriate for 37 of the 44 participants (84.09%). However, none of the factors under analysis present statistically significant differences. Auditory feedback usefulness was rated with the highest score by 30 of the 44 participants (68.18%); 7 participants (15.91%) chose level 4 in the Likert scale; 5 participants (11.36%) chose level 3 and only 2 participants (4.55%) rated it with a 2. The volume of auditory feedback was a problem for 6 participants, who mentioned that the voice was too quiet (even when the smartphone speaker was set at its maximum level). Another comment made by a participant was that listening to the auditory feedback was boring, because it was too repetitive.

Participants that completed the route receiving tactile feedback (T, AT, TV and ATV groups) mentioned that it was not annoying. However, tactile feedback was not enough to convey guiding information for 39 out of the 44 participants (88.63%). Only 5 participants (from the ATV group) mentioned that it was enough. Regarding the frequency of the tactile feedback, 39 out of the 44 participants (88.63%) considered it appropriate. The other 5 participants belong to one of the multimodal groups (AT, TV and ATV). The perceived usefulness was lower for tactile feedback: 20 participants out of 44 (45.46%) rated it with a 1 in the Likert scale; 11 participants (25%) with 2, 7 participants (15.91%) rated it with 3, 5 participants (11.36%) with 4 and only one participant (2.27%) with the highest score, 5. This last participant belonged to the ATV group.

Comments about tactile feedback were more varied: 3 participants mentioned that the vibration frequency was too slow. On the other hand, 2 participants gave positive comments about tactile feedback, mentioning that it keeps you alert on your route. They remarked tactile alerts combined with auditory and visual feedback made the application very useful.

Finally, 43 out of 44 participants that completed the path with visual feedback mentioned it was not annoying. The only participant that thought otherwise belonged to the TV group. Only 4 participants (9.09%) mentioned they felt they needed more information for completing the route. On the other hand, 37 participants (84.09%) reported that the frequency of the feedback was appropriate. Regarding the usefulness of Visual feedback to provide the guiding information, 28 participants (63.64%) rated it with the highest score in the Likert scale; 8 participants (18.18%) rated it with a 4; 6 participants (13.64%) with a 3; one participant rated it with a 2, and one participant rated visual feedback with the lowest value of the Likert scale. Sun glares caused 37 participants (84.09%) to report trouble and discomfort reading the on-screen messages. Combining visual with auditory and tactile feedback was mentioned as a good alternative for this situations.

The ease to follow the instructions to complete the route was rated with a 5 in the Likert scale by 57 participants (74.03%). It was rated with a 4 by 10 participants (12.99%), with a 3 by 4 participants (5.19%); with 2 by 5 participants (6.49%), and with the lowest rate by one participant only (1.30%). Participants that considered it was somehow difficult belonged to A, V, TV and ATV groups. The participant that considered it very difficult belonged to AV group.

6.3 Discussion

Visual and auditory feedback were the most intuitive way to receive navigation assistance for participants in our experiment. Spoken directions and the visual elements mixed with the video feed of the smartphones' camera are both promising ways to provide navigational cues to the user. Especially if we consider the fact that wearable devices such as Google Glass provides the necessary hardware to reduce the visibility and hearing problems reported in our experiment.

Our results align with what has been reported by previous studies (Jacob et al., 2011; Magnusson et al., 2010; Szymczak et al., 2012): multimodal feedback improves user performance. Participants in our experiment mentioned that receiving a different vibration pattern at a checkpoint increased their awareness of the decision point, efficiently attracting their attention to the smartphone screen or the spoken directional cues. Participants mentioned that tactile and auditory feedback helped them to keep focus on the route.

On the other hand, the fact that we considered in our design the precision of commercial GPS systems, including a radius of 5 m around the checkpoints to provide feedback, caused that none of the participants missed a checkpoint during the experiment. This highlights the importance of timing in pedestrian navigation assistance, because even when participants receiving only tactile feedback required more time to complete the route, all the participants were able to complete the itinerary.

Participants mentioned that sun glares made it difficult to observe and read the on-screen messages, and that the volume of auditory feedback was not high enough (even when the volume of the smartphone was set at its maximum level). This problem highlights the importance of a careful design considering the context in which feedback is going to be used. For example, auditory feedback would not be appropriate in contexts such as a library or a museum.

6.4 Chapter Summary

Visual and auditory feedback produced better results, both from the subjective appreciation of the participants and regarding the performance measures. In our approach, the augmented view provided similar information than a commercial GPS, but with the advantage of showing the real environment instead of just a map, while auditory feedback provided spoken instructions about the direction to follow. Tactile feedback is the least effective way to provide directional cues, which is not surprising, as previous research found similar results when tactile feedback was compared to traditional GPS systems feedback (Pielot and Boll, 2010). However, according to the user experience, tactile feedback is helpful to alert the user about an approaching decision point in the route, and also as an alternative when visual or auditory feedback are not available.

The following chapter analyses gesture-based interaction as an alternative input modality for 3D User Interfaces, specifically, for ODV interaction.

CHAPTER 7

OMNI-DIRECTIONAL VIDEO INTERACTION

Omni-Directional Video is an emerging media format that offers viewers a 360° panoramic video. The immersive experience is typically shown in a CAVE-like setup (see Figure 7.1), or a personal display (e.g. a head-mounted display) in combination with a tracking system to calculate the viewer's viewpoint. Recent efforts such as Microsoft's Illumiroom (Jones et al., 2013) provide interesting possibilities for ODV, as they show how a living room environment could be turned into a small CAVE-like theatre. Benko and Wilson (2010a) show different scenarios in which ODV can be used. They describe a portable dome setup in which users can interact with applications such as a 360° video conferencing system, a multi-user game or an astronomical data visualization system.

We envision ODV content becoming more and more common in the future and accessible within the context of our living rooms. As a result, traditional television watching experiences will change, since multiple viewers no longer have the same region of focus (i.e. the television screen in front of them), but are able to watch video content in any direction. This change also implies that traditional interaction methods, such as a remote control or the current gesture-based TV interfaces, need to be re-evaluated.

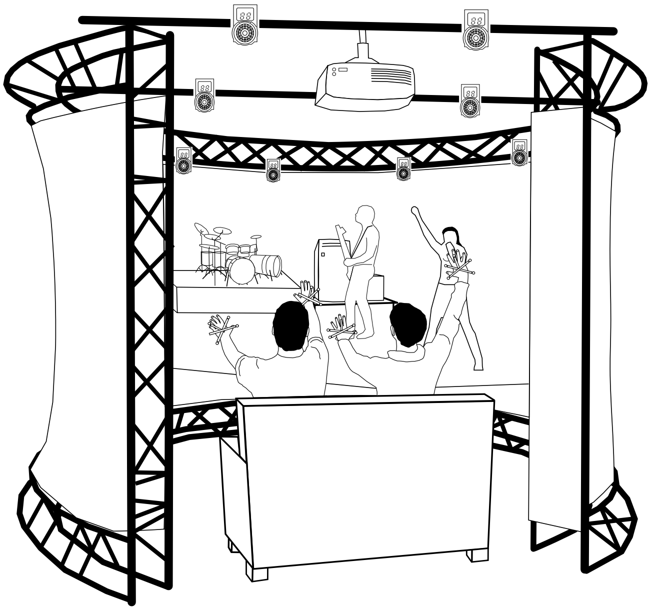


Figure 7.1: An ODV CAVE setup

Key contribution - The study described in this chapter presents a first step to address the specific challenges that multi-user ODV interaction poses. We compare the gestures performed by a group of participants in two different scenarios: interacting with the system on their own, and sharing the workspace with another participant. We study the mid-air gestures they performed, their properties and the strategies that participants adopted to complete the required tasks. Although we focus on designing appropriate gesture-based interactions for ODV, our findings can also be useful in other domains that require spatial or time-related operations (e.g. interacting with a home cinema or controlling the viewpoint during navigation of predefined sequences inside a virtual environment).

7.1 Study Description

The aim of the study is to evaluate the set of mid-air gestures that people consider the most appropriate for interacting with ODV, not only when users are on their own, but also in a collocated scenario, in the presence of other viewers who might want to interact with the ODV.

For this purpose, we gather both qualitative and quantitative data through observations, motion capture, questionnaires and interviews. We look into interactions for control operations typically performed with video content, either on television or digital video players, and also on some control operations that are typically used for spatial exploration. As a result, the control operations considered in this study are commands that manipulate time (i.e. play, pause, skip scene, fast forward and rewind) or space (i.e. panning and zooming).

7.1.1 Methodology

Literature has demonstrated that user-generated gesture sets tend to have a higher acceptance level among users (Morris et al., 2010). We adapted the gesture elicitation methodology of Nielsen et al. (2004) to gather the gestures that participants consider most appropriate for the aforementioned control operations. The study consisted of two sessions: first, a participant was asked to perform the gestures alone, and in the second session, two participants had to perform the gestures in a collocated setting.

Participants started the first session filling out a questionnaire with their personal information such as age, gender and experience with gesture interfaces. Then, a member of our team—that was also present as an observer, taking notes during the sessions—explained the list of control operations by showing an actual ODV to the participants. During the sessions, however, only still images of an ODV were shown to avoid unnecessary distractions.

The ODV did not respond to the gestures of the participants. Similar to Wobbrock et al. (2009), we decided against a Wizard of Oz approach to avoid that participants constrain or adapt their gestures according to the feedback they receive (e.g. to compensate for a delay or mismatch). A Wizard of Oz approach would also be impractical in the collocated scenario, because providing feedback for each participant simultaneously would inevitably result in inconsistencies.

Participants were asked to perform one easy to repeat and easy to understand gesture for each operation. They were informed that they had complete freedom of action to devise a gesture or posture using hand(s) and/or finger(s), and that the same gesture could be repeated for more than one action, if considered appropriate. Before the collocated session, the observer explained that they would be interacting with the ODV independently, but at the same time.

The observer asked the participants to devise an appropriate gesture for each operation, one by one. The observer did not impose any time constraints. Participants simply had to signal when they were ready to perform a gesture, and next, the observer gave the “go-ahead” to execute the gesture. During the collocated session, both participants had to perform their gesture at the same time, so the observer waited to give the “go-ahead” until both participants were ready.

Participants were seated on a couch, inside a CAVE-like ODV setup, as seen in Figure 7.1. This kind of setup helps participants to explore the interaction possibilities, since it clearly reveals the spatial properties of the ODV content. It also prevents participants from being influenced by the form factor of the output device (e.g. when using a rectangular screen, participants are more likely to unnecessarily frame gestures within a rectangle in front of them). On the second session, participants were seated reasonably close to each other (as it would happen in a living room, sitting next to each other on a couch) to evaluate the impact of the collocated setting. Participants were not forced to sit uncomfortably close to each other, however, and had sufficient space to sit without invading each other’s personal space.

When both participants finished performing the gestures for all control operations during the collocated session, they were asked to swap positions on the couch and repeat the trial. In other words, participants performed gestures for each control operation three times in total: once alone, and twice when sitting next to the other participant.

To control order effects, we divided the participants into two groups: each group received the control operations in a different order during the sessions. However, we did not simply randomize the order of the control operations, but decided to maintain a logical structure (e.g. by grouping related operations such as fast forward and rewind), to make it easier for the participants to devise gestures.

After completing both sessions, participants filled out a questionnaire regarding their experience and discussed their opinions with the observer.

7.1.2 Participants

Sixteen participants took part in our study: twelve male and four female, with ages ranging from 23 to 52 years old (average age 31.5). All of them were colleagues at our research centre. Two participants are left handed, two ambidextrous, and the others are right handed. Most participants are experienced touch screen users (12 participants use them daily), but merely 2 participants play video games on consoles like the Nintendo Wii or Microsoft Xbox with Kinect more than once a month. Only 2 participants make regular use of gestures to interact with their PC, either by performing mouse gestures to control the web browser, or by using a multi-touch mouse. None of the participants had experience interacting with gesture-based TVs. Finally, 12 participants knew beforehand what ODV was, but only 2 of them had previously interacted with an ODV system.

For the collocated session of the study, we formed 8 pairs according to the following criteria: 4 pairs with participants who were used to interact with each other and 4 pairs with participants who rarely interacted with each other. We based our grouping criteria on the *"friendship ties"* described by Haythornthwaite and Wellman (1998):

Close friendship people who work in the same office, usually have lunch together, and would go to the movie theatre together.

Working together people who know each other, but rarely interact with each other in the work environment.

7.1.3 Apparatus

To gather data about the gestures that participants performed, we used motion capture, allowing us to measure the spacial dimensions of the gestures. For this purpose, we used 8 OptiTrack V100:R2 cameras and the Natural Point Tracking Tools software (see the cameras mounted on the structure depicted in Figure 7.1). The OptiTrack cameras have a 640 × 480 pixels image resolution and a maximum capture frame rate of 100 fps. They are capable of tracking markers with submillimetre accuracy. We also used a normal video camera to record the sessions, to make classifying the different gesture easier during analysis.

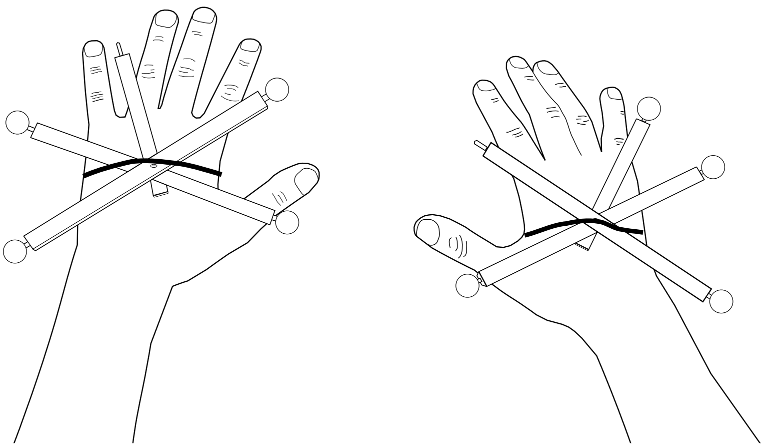


Figure 7.2: Rigid body markers composed of small IR reflective balls, used for motion capture.

To track participants, a rigid body marker composed of small IR reflective balls had to be attached to each hand. Before the actual study, we ran a pilot study for two purposes: *(i)* to verify whether the instructions and study design were clear and *(ii)* to uncover limitations and issues with our apparatus. It allowed us to optimise the rigid body markers in order to avoid occlusion problems when participants turned their hands. We therefore built the markers with wooden sticks that exceeded the size of the participant's hand (Figure 7.2). In this manner, only the hands' centres are tracked and not the small finger movements, but this suffices for our purposes, since we have complementary video recordings.

7.2 Results

To compare the gestures performed by participants for every control operation and study their properties, we analysed all the video recordings, following the strategy proposed by Peltonen et al. (2008): first extracting the video segments with useful data and then extracting all the required information.

For this purpose, we defined parameters to annotate the videos of each gesture according to the suggestions of Nielsen et al. (2004). We described all the gestures in natural language using these parameters, in such a way that others would be able to understand and reproduce them. The parameters we used are:

- hand usage (*one or two hands*),
- trajectory of the movement (*linear or circular*),
- type of gesture (*movement or steady posture*) and
- granularity (*fine-grained finger movements or coarser hand movements*).

Table 7.1 shows the results of the analysis. The value in each cell represents the number of participants who used the property for the specific operation during the study: 16 participants performed a gesture for each control operation 3 times, resulting in 48 samples per control operation in total. The * mark represent a statistically significant difference between levels (non-parametric binomial test, $\alpha = 0.05$)

The *Hand usage* column of Table 7.1 shows that participants had no clear preference for using one or both hands for most gestures. They did prefer to use one hand for performing a pause gesture and both hands for zooming (these differences are statistically significant, based on a non-parametric binomial test between the two possibilities). Furthermore, participants preferred to use linear movements rather than circular movements. As indicated in the *Trajectory* column of Table 7.1, the difference is statistically significant for all control operations. This confirms the findings previously reported for pan-and-zoom interaction with wall-sized displays (Nancel et al., 2011). People prefer linear movements when they are asked to devise easy to perform, easy to remember and easy to repeat gestures.

We classified all the gestures as either static (e.g. a steady hand posture that uses both index fingers to represent the typical pause symbol), or dynamic (e.g. performing a “push” gesture by moving a hand away from the body and back, with the palm outwards). Both examples are depicted in Figure 7.3. The *Gesture type* column of Table 7.1 shows a clear preference for using dynamic movements rather than static hand postures to represent most control operations. Pause and stop are control operations for which the participants’ preference is less clear, but for the other operations, the differences are statistically significant. We believe

Table 7.1: Results of the gesture properties analysis.

	Hand usage		Trajectory		Gesture type		Granularity	
Control operation	One	Two	Linear	Circular	Static	Dynamic	Fine	Coarse
Play	31	17	48*	0	14	34*	14	34*
Pause	36*	12	48*	0	27	21	9	39*
Stop	19	29	48*	0	26	22	0	48*
Skip scene	24	24	33*	15	3	45*	2	46*
Fast forward	31	17	41*	7	3	45*	7	41*
Rewind	29	19	39*	9	6	42*	9	39*
Pan	18	30	46*	2	1	47*	4	44*
Zoom	3	45*	48*	0	0	48*	10	38*

that the number of participants using steady postures is higher for pause and stop, because they both implicitly denote turning the video into a standstill state. A number of participants used the same gesture with different speed/timing to represent different control operations. Participant 15 (P15), for instance, explicitly mentioned that he did the same gesture for play and fast forward, moving his right hand to the right, but varying the time he kept pointing in that direction (longer for fast forward).

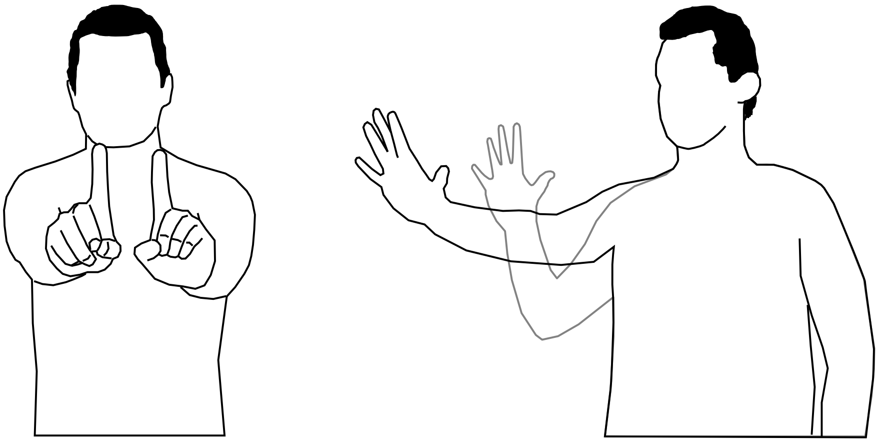


Figure 7.3: Two gestures representing the pause operation. Grey lines represent the initial position and black lines the final state.

We also found statistically significant differences comparing the usage of fingers (fine granularity) and whole hands (coarse granularity) to perform the gestures (*Granularity* column of Table 7.1). Participants preferred to use coarser hand movements instead of fine-grained finger movements, even though they were informed that they could use finger movements to represent the control operations.

As expected, participants extrapolated their knowledge from real-life devices and software applications (in this case, mostly video or DVD players), an observation that was also made by Henze et al. (2010) in the context of gestures for music playback. This was especially true for play, pause, stop and zoom. Participants for instance tried to transform a symbolic representation into a gesture or posture, such as a triangle for play or a square for stop (e.g. P14 explicitly asked *"... do I have to do the square for stop?"*). Another form in which participants extrapolated their real-life knowledge is when they considered that play, pause and sometimes stop should be represented by the same gesture, as these control operations are often mapped to the same button on devices or in software applications (e.g. a lot of media players use the same button for play/pause and do not have a stop button). For zooming, twelve of the sixteen participants employed the typical spread-and-pinch gesture, even participants were not frequent multi-touch users.

7.2.1 Collocated Interaction

An interesting part of our study consisted on analysing the changes the participants made to represent each control operation when they were interacting with the ODV system together with another participant. To this end, we employed the answers to the post-study survey participants filled out about their experience doing the gestures, the motion capture data (mo-cap) from the tracking system and the video recordings.

The outcome of the survey analysis revealed that "Avoid invading the other participant's private space" and "Avoid colliding with the other participant's gestures" were chosen by seven out of the sixteen participants as factors that influenced their decision to perform a gesture. "Avoid blocking the other participant's view of the video", however, only received two votes, one of which belongs to P14, who felt his field of view was blocked and reported collisions with his fellow participant while performing the gestures. We believe the fact that "Avoid blocking the other participant's view of the video" received few votes is due to the absence of a particular task to perform with the ODV. Participants did not need to be engaged with the content and thus did not consider blocking the other participant's view an important factor.

The analysis of our study notes and video recordings shows that participants of the four pairs of the *"close friendship"* category had no problems performing gestures side by side. One pair of participants even made jokes about synchronized dancing, because they performed nearly identical movements for some control operations. Participants who were part of a *"working together"* pair, on the other hand, were more uncomfortable and some of them expressed that feeling during an informal interview after the study. P5 reported, for instance, that *"It was not comfortable doing the gestures with the other participant."* and P16 reported that *"I felt limited by the presence of the other participant. She invaded my private space."*

By analysing the video recordings and mo-cap data, we identified three interesting situations that resulted from the collocated interaction: participants adapted the size of their gesture, changed hands to perform the same gesture, and chose a completely different gesture for the same control operation. We discuss each of these gesture adaptations in the next sections.

7.2.1.1 Size Adjustment

A number of scripts were implemented to automatically analyse the motion capture data gathered by the OptiTrack system. We first measured the space participants can cover when they completely stretch their arms to the side, to the front and to the top. The areas created on each plane (*XY* - *frontal*, *XZ* - *top* and *YZ* - *lateral*) represent the maximum distances that a participant is able to reach. These areas were used to create baseline bounding boxes. Next, we decomposed the captured hand movements and determined the 2D bounding boxes that represent the areas covered by each gesture on the three planes. Finally, we calculated the ratios between the sizes of the bounding boxes for each gesture and the participant's baseline bounding boxes.

We used these ratios to detect changes in size of a gesture for each control operation across the sessions, to investigate if participants used this as a strategy

Table 7.2: Size adjustments analysis. Values in each cell represent the number of participants adapting the size of the gesture by more than 10% between the specified sessions. Column names stand for S: single participant session, A: first collocated trial, and B: second collocated trial.

	Lateral adjustment (X)			Vertical adjustment (Y)			Depth adjustment (Z)		
Control operation	S vs A	S vs B	A vs B	S vs A	S vs B	A vs B	S vs A	S vs B	A vs B
Play	3	3	2	2	0	1	4	7	2
Pause	1	1	0	4	3	0	2	3	0
Stop	2	3	1	1	1	0	4	3	0
Skip scene	6	3	1	2	3	0	7	2	1
Fast forward	5	5	0	2	3	0	5	5	2
Rewind	4	5	1	4	3	1	5	6	3
Pan	8	4	1	1	2	0	7	5	3
Zoom	8	7	0	1	1	0	6	5	0

to adapt gestures in the collocated setting. Table 7.2 presents the number of participants who reduced the size of their gestures by more than 10%, for each of the three axes. The size adjustment is especially noticeable for control operations that typically involved lateral movements (e.g. fast forward, rewind, pan, zoom), due to the presence of the other participant.

Analysis of the friendship ties revealed an expected trend: participants of *"working together"* pairs adjusted the size of their gestures more often. They adjusted 42.2% of all the gestures performed during the sessions (for all the control operations and in the three movement directions), while participants of the *"close friendship"* pairs adjusted only the 17.2% of their gestures. We did not find a statistically significant correlation between friendship ties and size adjustments, which can probably be attributed to the limited number of pairs per friendship tie.

As an example, we briefly discuss the gestures for the zoom operation of a *"working together"* pair. When P8 and P13 performed the zoom gestures for the first time, their hands collided. The second time, after switching positions, both participants adjusted their gesture by displacing the movements to the free space (Figure 7.4). While looking into another *"working together"* pair, P4 and P16, we clearly noticed both participants reducing the movement of their hands between the single and collocated session to represent the skip scene operation.

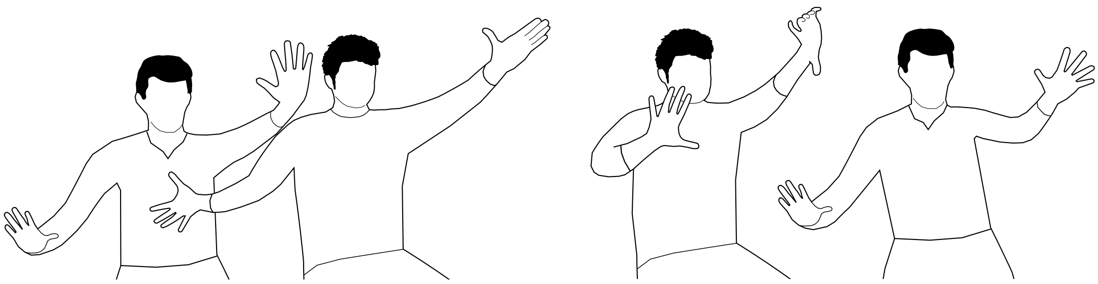


Figure 7.4: Participants adjusted the size of their gestures in collocated trial.

7.2.1.2 Gesture Mirroring

Analysis of the video recordings shows that participants also adapted their gestures by using a different hand to perform the same gesture. Table 7.3 depicts how many participants used gesture mirroring across the different sessions. The values in each cell represent the number of participants who mirrored gestures between the specified sessions.

In total, 5 participants adapted their gestures in this manner. Only 1 of those participants was ambidextrous, and 4 were part of a *"working together"* pair. In total, the *"working together"* pairs used gesture mirroring for 17.2% of the gestures performed during the sessions and the *"close friendship"* pairs for 4.17% of their gestures, but again, no statistically significant correlation was found between friendship ties and the adaptations.

Table 7.3: Gesture mirroring analysis.

Control operation	S vs A	S vs B	A vs B
Play	1	2	3
Pause	1	2	1
Stop	1	1	0
Skip scene	1	1	2
Fast forward	0	0	1
Rewind	1	1	0
Pan	2	1	3
Zoom	0	0	0

To illustrate the gesture mirroring strategy, we briefly discuss three examples. P8 used his left hand for the fast forward gesture when performing the gesture during the single session, but he used his right hand when sitting to the right of P13 (Figure 7.5). Similarly, P14 used his right hand when he was sitting to the right of P7 when doing the skip scene gesture, and then his left hand when he was sitting to the left of P7. Finally, P1 did the play gesture using her left hand when P10 was sitting at her right, and her right hand after exchanging positions on the couch.

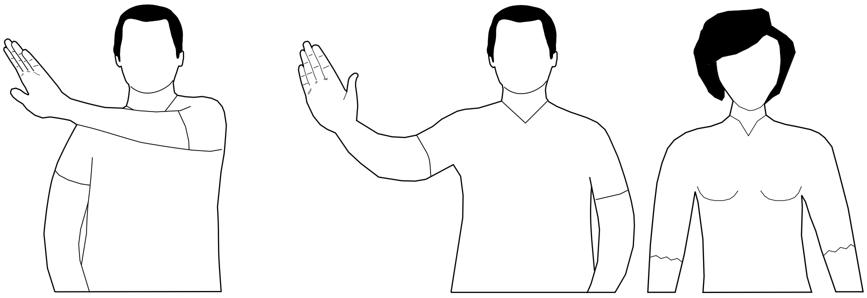


Figure 7.5: Participants used different hand to perform the same gesture when another participant was present.

Although only a limited number of participants adopted this mirroring strategy, it is still interesting to note that users will expect a gesture to be recognized by the system in both cases, whether they are using their left or right hand. The Microsoft Kinect development guidelines already suggest this strategy to create flexible gestural interfaces (Microsoft Inc., 2013).

7.2.1.3 Choosing New Gestures

Table 7.4 depicts the number of participants who changed gestures across the sessions. Eleven participants changed at least one of their gestures. In total, 11.5% of the gestures performed by participants of a *"working together"* pair were changed across the sessions, and 9.38% in case of *"close friendship"* pairs. No statistically significant correlation was found between friendship ties and choosing new gestures.

The main reason for this behaviour is the extra time participants spent thinking about the gestures. P15, for example, mentioned that the second time he had to perform the gestures, he *"tried to put some logic"* in them, and P2 mentioned that he tried *"to do more energy efficient"* gestures.

Most pairs (all the *"close friendship"* pairs and two of the *"working together"* pairs) discussed the reasons and implications of their gestures, with comments

Table 7.4: Choosing new gestures analysis.

Control operation	S vs A	S vs B	A vs B
Play	2	3	1
Pause	1	1	0
Stop	2	2	1
Skip scene	1	3	2
Fast forward	3	3	1
Rewind	4	3	1
Pan	3	2	1
Zoom	0	0	0

like *"your gesture is not energy efficient"* or *"your gesture is error prone"*. This interaction between participants sometimes resulted in them changing the gesture. For instance, P4 used both hands for panning when she represented the operation the first time. She completely stretched both arms to the front, making a clockwise circle with her right arm to pan right and a counter clockwise circle with her left arm to pan left. The second time, she used only one hand, copying the gesture of her fellow participant: moving her right hand, pointing with the index finger to the left and then to the right.

7.2.2 Agreement Level

We classified all the gestures for every control operation into groups of similar gestures, based on the parameters *hand usage (one or two hands)*, *trajectory of the movement* (*linear or circular*), and *type of the gesture* (*movement or steady posture*). In addition to these parameters, we considered the overall movement pattern of the gesture (e.g. the directions of the movements). We did not include *granularity* (*fine-grained finger movements or coarser hand movements*), so we classified spread-and-pinch gestures performed with two fingers or with both hands as similar gestures.

Next, we calculated the percentage of participants who used a particular type of gesture and the agreement level for each operation. For this purpose, we used the formula of Wobbrock et al. (2005):

$$A_i = \sum_{j=1}^{n} \left(\frac{G_{ij}}{G_i} \right)^2 \tag{7.1}$$

A_i is the agreement level of the *ith* operation, G_i the total number of gestures performed for the *ith* operation and G_{ij} the number of elements in the *jth* group of gestures for the *ith* operation. Park and Han (2013) used this formula in a similar manner.

We illustrate the formula's usage by applying it to the gestures used for panning (Equation 7.2). We found three groups of similar gestures, with sizes 30, 2 and 16. As a result, the agreement level was 0.5036 for the panning operation.

$$A_{Panning} = \left(\frac{30}{48}\right)^2 + \left(\frac{2}{48}\right)^2 + \left(\frac{16}{48}\right)^2 = 0.5036 \tag{7.2}$$

We also calculated the percentage of participants who chose a particular type of gesture, taking into account the 48 gestures performed for each operation. Table 7.5 depicts both these percentages and the agreement levels for the top-rated gesture for each operation. In this table the Rate column represents the percentage of participants who performed this type of gesture during the study. The agreement level column gives an indication about the variety of gestures that were performed for an operation.

Table 7.5: Top-rated gestures for the eight control operations.

Control operation	Rate (%)	Agreement level
Play	35.41	0.18
Pause	41.67	0.22
Stop	22.92	0.21
Skip scene	33.33	0.16
Fast forward	35.42	0.26
Rewind	41.67	0.30
Pan	62.50	0.50
Zoom	75.00	0.63

7.3 Discussion

As a result of the analysis and comparison of the gestures performed by the participants during the three sessions in our study, we propose the following user-defined gesture set:

Play Push gesture, moving the hand, with the palm outwards, toward the front and back in a fluent movement.

Pause Halt gesture, holding the arm completely stretched with the palm outwards for a few seconds.

Stop Halt gesture, holding both arms completely stretched with the palm outwards for a few seconds.

Skip scene Moving the hand from right to left or from left to right one time and returning to the starting position.

Fast forward Left to right movement, holding the hand pointing to the right for a few seconds.

Rewind Right to left movement, holding the hand pointing to the left for a few seconds.

Pan Using one hand to "grab" the video and then move it from left to right or from right to left.

Zoom Using the spread-and-pinch gesture, moving two hands apart (spread) and bringing them back together (pinch).

We believe these gestures will lead to a high acceptance level among users. Figure 7.6 gives a graphical representation of this gesture set. The different states of gestures that require movements are represented with different line colours: grey colours represent early states of a gesture and the black line the final state.

We considered the most repeated gestures across all the sessions to assemble our gesture set. In 62.5% of the cases, the difference between the most repeated gesture and the second most repeated gesture was very large. For the fast forward and rewind operations, however, the differences were small. There were six groups of similar gestures for the fast forward operation (representing 35.42%, 29.16%, 18.75%, 10.42%, 4.17% and 2.08% of the participants) and eight groups for the rewind operation (representing 41.67%, 33.33%, 8.33%, 6.26%, 4.17%, 2.08%, 2.08% and 2.08% of the participants). In case of the stop operation, performing the halt gesture with one hand was the most repeated one (performed by 33.33% of the participants), but as we have already chosen this gesture to represent the

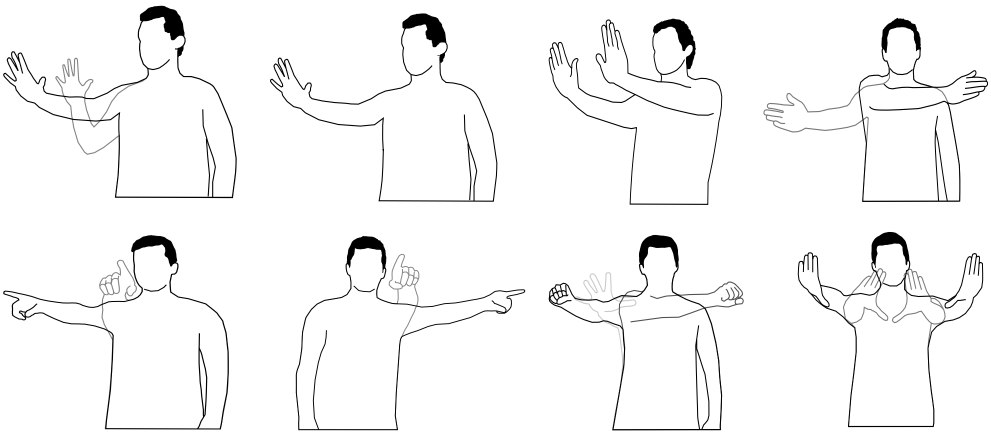


Figure 7.6: From left to right and top to bottom: play, pause, stop, skip scene, fast forward, rewind, pan and zoom. Grey lines represent early states of the gesture, black lines represent the final state.

pause operation, we decided to use the second most repeated gesture: the halt gesture using both hands (performed by 22.92% of the participants).

We can observe in Figure 7.7 that the large variety of gestures to represent certain control operations sometimes causes low agreement levels (for instance the play and skip scene operations). The zoom operation, on the other hand, was the one with the most consensus. This can be explained by the fact that participants regularly relied on existing mental models to devise gestures to represent control operations. The following three examples illustrate this behaviour:

Zoom and pan Not surprisingly, the most repeated gesture for zoom was the widely used spread-and-pinch gesture, as indicated by the agreement level. For panning, "grabbing" the video and moving the hand was the most repeated gesture. Similar gestures have been proposed by Fikkert et al. (2010) for zooming and by Stellmach et al. (2012) for panning, in the context of large display control.

Play and pause Some participants mentioned that video players use the same button to represent play and pause, and thus they also used the same gesture for both control operations. This behaviour was also reported by Henze et al. (2010) in the context of gestures for music playback. Overall, there was a great diversity in gestures to represent play and pause, leading to lower agreement levels.

Fast forward and rewind Most media players and timelines associate "the future" with the right side (e.g. arrows pointing to the right in video players for fast forward), and "the past" with the left side. Participants in our study represented both control operations following this established mental model, which is consistent with the observations of Henze et al. (2010).

Although we considered the most repeated gestures to assemble our gesture set, not all gestures are necessarily the optimal solution. Participants sometimes changed their gesture to copy their fellow participant (imitative behaviour was also reported by Walter et al. (2013), in the context of a public display game), so they considered their first gesture to be suboptimal. Eleven participants also chose a new gesture for at least one of the control operations, due to reasons such as wanting a more *"energy efficient"* gesture. This implies that using a gesture set over a prolonged period of time might lead to a different prioritisation of gestures' properties. Our gesture set thus needs further validation and refinement before its actual implementation, and the next step is to benchmark the chosen gestures, as suggested in the methodology of Nielsen et al. (2004). Another component to consider before implementing the set, is how to discriminate between gestures and other movements, for instance by indicating the start of an interaction with

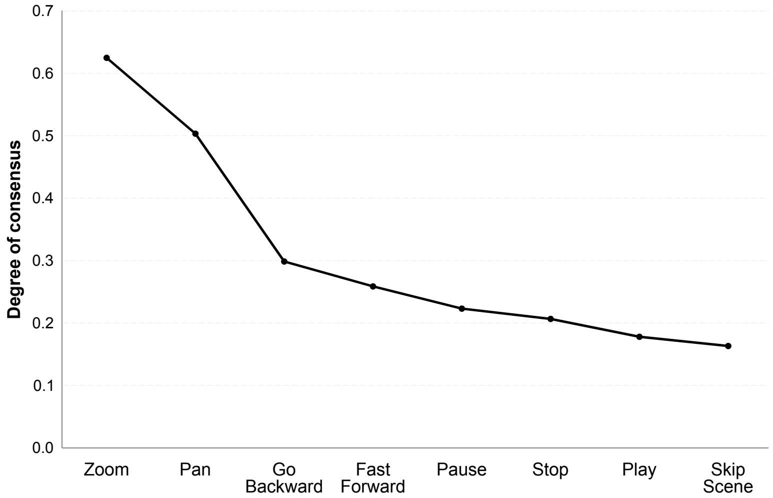


Figure 7.7: Gesture set agreement level in descending order.

a specific body pose (Walter et al., 2013).

We noticed that minimal friendship ties between participants had a negative effect on their experience. Participants of the *"working together"* category often felt uncomfortable performing gestures close to each other. This must be taken into consideration when designing a gestural interface: when users are likely to be unfamiliar with each other, less invasive gestures might need to be considered, while such gestures might be a source of fun for close friends.

We identified a number of gesture adaptations caused by the collocated setting: participants used a different hand to perform the same gesture, changed the size of the gesture, or performed the gesture more to the left or to the right to avoid colliding with their fellow participant. Participants expect their gestures to be recognized in all cases, regardless of the hand they use or the scale of their gesture. Therefore, the system has to be designed to recognize all (or at least the most common) forms of a gesture (e.g. recognize a spread-and-pinch gesture performed with two hands, but also one performed with two fingers). The need to support gesture variations was already observed in other contexts, such as multi-touch surfaces, Hinrichs and Carpendale (2011), and interactive public displays, Walter et al. (2013).

We only looked into which gestures participants found to be the most appropriate for every operation. We neither took into account the parametrization of control operations (e.g. how users express how many degrees to pan with their gesture), nor the focus point of those control operations (e.g. how users express on which area they want to zoom in). These factors also need to be investigated, because they can influence how gestures are scaled. When a single user performs a small panning gesture, for instance, it probably means that she wants to move only a little. In a collocated setting, on the other hand, the same small panning gesture might be the result of the presence of others. A gesture recognizer should take this into account by scaling the panning operation according to the situation. Care has to be taken, however, that this kind of adaptation does not confuse users. The system needs to provide sufficient feedback about the scale of control operations.

During our study, participants did not need to be engaged with the actual ODV content, nor did they have different points of focus, which is likely to happen in a CAVE-like setup. As a result, they rarely considered aspects such as blocking the other participant's view. Such aspects will become an important factor when participants do engage with the content, which might lead to additional adaptation strategies when performing gestures. However, participants not having different points of focus during the study does allow us to generalise our results beyond CAVE-like setups.

7.4 Chapter Summary

The user-defined gesture set contains the most repeated gestures in our study. We observed a clear preference for using linear movements to represent easy to perform and easy to remember gestures. Participants also preferred to use dynamic movements rather than static hand postures to represent most control operations, and coarser hand movements instead of fine-grained finger movements. We also found that participants tried to extrapolate their knowledge from interaction with real-life devices or software applications.

Analysis of the collocated interactions revealed interesting behaviours that participants exhibited while devising and performing gestures. They adapted their gestures in several ways because of the presence of another participant. The most prominent adaptations were changing the size of the gesture and shifting the hand movements to the opposite side of where the other participant was sitting. Other gesture adaptations were using a different hand for the same gesture or devising a new gesture. These adaptation strategies highlight the importance of a good system design. The ODV system must be able to interpret the user's actions (e.g. adapting the scale of a gesture because of the proximity of another person versus adapting the scale to make smaller adjustments), give sufficient feedback about the scale, and provide sufficient flexibility to cope with different variations of gestures (e.g. a spread-and-pinch with both hands or two fingers).

Although our findings are based on a gesture elicitation study regarding control operations for ODV in a CAVE-like setup, we believe the user-defined gesture set and user expectations can also be useful in other setups and domains that require spatial or time-related operations. Furthermore, this study is only a first step in the exploration of ODV gestures, with many interesting avenues for future research, such as studying collaborative tasks with users who each have different points of focus, or an in-depth analysis of the reasons for choosing a specific gesture, which might reveal certain cultural, educational or generational influences.

CHAPTER 8

DISCUSSION

The Human-Computer Interaction discipline studies all the aspects involved in the two-way communication process between a user and a computer system. One of the paths that has been explored by previous research to improve this communication is multimodal interfaces. On one hand, multimodal feedback uses a combination of different sensory channels to provide feedback to the user, where the most common modalities are visual, auditory or tactile. On the other hand, multimodal input techniques allows the user to control the system using a combination of different input techniques, such as traditional keyboard and mouse, speech recognition and gesture tracking. Multimodal interfaces offer multiple advantages, such as providing extra resilience to issues like noise and distractors. Multimodal interfaces can also have a negative effect if stimuli compete for the same cognitive structures. Personal preference or sensitivity to one type of stimulus can also be an important factor regarding the effectiviness of multimodal feedback, as explained by Coutaz et al. (1995) in the CARE properties for multimodal interaction. It is worth noticing that these properties were defined for multimodal input, however, they can also be applied to multimodal feedback, as is our case.

Considering the CARE properties (Coutaz et al., 1995), the multimodal stimuli we used in our experiments were designed to be *equivalent* (as any of the modalities can be used individually to fulfill the purpose of the feedback). Therefore, when more than one modality is used as a feedback, the information provided to the user is *redundant*, because all the modalities are given within the same temporal window.

3D User Interfaces (3DUIs) are being used in many applications to solve complex problems, and it is very important to study the use of multimodal interfaces in 3DUIs to improve their effectiveness and efficiency. The work described in this thesis tries to address specific questions about multimodal 3DUIs in two areas: Augmented Reality and Omni-Directional Video. Nevertheless, because of the properties of the interactions we studied, our findings can be extrapolated to other 3DUIs with similar purposes and types of interaction. For example, multimodal feedback could benefit the experience of a technician on how to move (change position and orientation) different pieces of a complex engine, by using a simulator trainer.

In this chapter we discuss the implications of the different experiments we did to study multimodal 3DUIs.

8.1 General Discussion

The existing research on multimodal integration has demonstrated how robust human perception is when mixing information from our five senses (Ernst and Bülthoff, 2004; Ernst, 2006; Wozny et al., 2008). For example, tactile and auditory feedback have been successfully used as directional and distance cues in different application domains, such as military personal assistants for target acquisition tasks (Ahmaniemi and Lantz, 2009; Oron-Gilad et al., 2007; Lindeman et al., 2005). Multimodal AR interfaces, however, have not received the same attention, and, as mentioned by Billinghurst (2008), it is important to understand how to create new interface metaphors that can properly integrate the relationship between real and virtual content.

We performed three experiments in three of the major AR application domains, comparing the combinations of visual, auditory and tactile feedback modalities when they are used as directional, proximity and alerting cues. With these experiments, we confirm that Wickens' Multiple Resource Theory (Wickens, 2002) also applies to multimodal AR applications.

Combining feedback modalities is a good alternative to alleviate the problems derived from saturation of information in a sensory channel. In the experiment with the searching assistant, participants chose tactile feedback as the most helpful feedback because it was an intuitive way to provide proximity cues. The feedback was provided directly on the participants' hands. The Nintendo Wii remote could be used as a pointer while looking for the desired book, blending the feedback in the region of focus, but using an alternative sensory channel. Furthermore, considering the scenario of a library in which the experiment was based on, tactile feedback was also preferred because of the possibility to use it without disturbing people around the user. The experiment with the multimodal navigation assistant showed that vibration patterns were also good alerting cues to attract participants' attention when an action was required. Spoken directions on the other hand were an efficient alternative to visual cues when the sunlight affected the visibility of on-screen elements. Auditory feedback also produced good results when participants received soft auditory warnings after a crossing event in the Augmented Wire Loop Game.

The downside of auditory feedback is that it can go unnoticed in noisy environments, as was the case in our experiment with the multimodal navigation assistant. Participants mentioned the volume of the feedback was too low, even when the volume of the smartphone was set to its maximum. The other problem with auditory feedback mentioned by participants was that it can get too repetitive or annoying when the sound is too loud, as we observed during the experiment with the Augmented Wire Loop Game. On the other hand, tactile feedback was the least effective way to provide directional cues in the multimodal navigation assistant, as the user had to scan the surroundings to find out the direction to follow. Furthermore, it made participants nervous when playing the Augmented Wire Loop Game.

When the directional, proximity, or alerting cues were presented in the region of focus, the effect on the participants performance was positive. The crossing alerts in the Augmented Wire Loop Game—the change in the wire loop colour—or the virtual path shown on the camera's view during the experiment with the multimodal pedestrian navigation system produced good results both subjectively and objectively.

In our experiments, the performance of the participants decreased when they had to deviate their attention from the region of focus required by the task, e.g. when they had to look down to receive the visual proximity cues from the Nintendo Wii remote. We also observed that visual collision alerts that are not shown in the region of focus—for example, the bulb in the corner of the screen—went unnoticed by participants playing the Augmented Wire Loop Game. Participants were too concentrated on the virtual objects to notice them.

Another important aspect for improving multimodal Augmented Reality applications is depth perception. We were especially interested in evaluating the benefits of using an HMD compared to a normal screen, which is the most common display technology for Augmented Reality. The use of stereoscopic graphics improved the depth perception in the Augmented Wire Loop Game, reducing the number of *collision* and *crossing events*. Stereoscopic graphics also increased the degree in which participants liked the game. However, since our prototype did not solve the visibility between real and virtual objects, participants also mentioned that the scene looked strange, because the virtual objects appeared on top of their hands all the time. This is an important issue, and designers should take it into account. New technology such as Google's Project Tango (Google Inc., 2014) or the sensor *Structure* (Occipital Inc., 2014) are working in solving this issue.

Using the HMD produced better results when displaying stereoscopic graphics, but the performance was better without stereoscopic graphics for those participants playing with the normal screen. Participants who used the HMD mentioned different issues during the experiment: controlling virtual objects was difficult because they had to keep their head as steady as possible, and at the same time, control the movement of their hand to move the virtual ring. The HMD we used in the experiment was also too heavy and somehow uncomfortable for long term use, and participants reported eye strain and dizziness. The new generation of HMD addresses these problems. Oculus Rift (Oculus VR., 2014) is one example of this improvement, as it is a lighter HMD than the Sony HMZ-T2 we used in our experiment. Visualisation hardware is also more affordable than a few year ago, putting this technology within the reach of more people outside the research laboratories.

Regarding interaction with Omni-Directional Video, the study we performed to understand mid-air gestures for controlling operations gave us valuable insights to improve their design. Comparing all the gestures the participants performed to represent the eight control actions for ODV lets us propose a user-defined gesture set. However, the most interesting outcome of this study are the properties of these gestures we could observe. Participants preferred to use one-handed gestures, with linear movements that requires moving the entire arm to activate the desired action. They also tried to extrapolate their previous mental mappings, such as representing operations with an implicit cut-off meaning (e.g. as stop and pause) with static hand postures like the halt gesture, or using the common spread-and-pinch gesture for zoom. This is worth noticing, because, even when it is not easy to talk about universal gestures that can be applied to a specific action, we found similar gestures reported in other application domain, such as controlling an audio player (Henze et al., 2010).

Studying the gesture elicitation results, we could observe different adaptation strategies that participants applied when performing the gestures: adjusting the size of their gestures, using a different hand, or choosing a completely new gesture between each session. It is interesting to see how users expect the system to be fully and easily configurable according to their preferences. We also observed that the level of comfort performing the gestures varied according to the level of friendship with the other participant during the collocated session. Participants performed the task comfortably when they were alone or together with a *close friend*. On the other hand, when participants performed the gestures with a co-worker they do not have a close relationship with, they felt more uncomfortable.

It is important to notice that even when the time span between the sessions in our study was small, participants changed their gestures. The adjustments were due to the presence of another participant, as a result of a short discussion about the implications of their gestures, or, as explained by the participants, after spending more time thinking about better gestures for each control operation.

In general terms, during the study, participants preferred to perform gestures employing complete arm-hand movements. However, it is to be expected that in the long term, these gestures would change, e.g. to use smaller gestures requiring less physical effort.

8.2 Validity Evaluation

Wohlin et al. (2000) describe a list of threats that can affect the results of an experiment with computer systems. In this section, we discuss them together with the approach we followed to avoid or reduce their effect on the work described throughout this thesis.

8.2.1 Conclusion validity

Conclusion validity refers to the threats that could mislead the researcher from drawing the correct conclusion about the relationship between the independent variables (the treatment) and the result of the experiment. It considers the *reliability of the measures*, *reliability of the application of treatments to subjects*, and the *random heterogeneity of the subjects*.

The analysis of multimodal feedback that we performed is based on the objective measures that were automatically recorded by our prototypes (e.g. the task completion time or the number of errors). However, the questionnaires to evaluate the experience of the participants were designed explicitly for the specific scenario under evaluation. This is a drawback in our experiments.

On the other hand, we randomized the combinations of the feedback modalities on every experiment when we used a between subjects design, and the order of the tasks in the experiment with a within subjects design. We tried to ensure a standardized application of the treatments (the feedback modality and the task order) defining experimental procedures that were followed to the letter with every participant.

All the participants in our experiments were university students and in many cases they have a computer sciences background. We tried to ensure a balanced distribution among the different experimental conditions. None of the participants in our studies had previous experience with any of the systems we developed, ensuring a relatively homogeneous sample of participants. Even if they had previous experience using the Nintendo Wii remote, the granularity of the movements required in our experiments was different enough to those required by the video games on the Nintendo Wii console. Therefore, participants had a similar level of experience in the tasks we evaluated.

8.2.2 Internal validity

Internal validity describes the issues that might cause the results to show causal relationships even when there is none. It focuses on two aspects: *instrumentation* and *maturation*. We validated the mechanisms to collect the objective measurements in the applications we developed for our experiments. We also tried to reduce the learning effect in our experiments. For example, by using different levels of the Augmented Wire Loop Game for training and for the evaluation, or by changing the set of books in the bookshelf mock-up for every combination of feedback modalities. When we used paper questionnaires to ask participants about their experience, the transcription of their responses was double-checked. We opted to use electronic questionnaires for the last couple of experiments, minimizing any possible transcription errors. To reduce the effect of fatigue on participants during the experiments, we allowed them to take breaks when they considered necessary.

8.2.3 Construct Validity

We chose three of the major application domains for Augmented Reality to evaluate the benefits of multimodal feedback. All of the applications we used for the experiments are typical AR application. We used the hardware configuration that users might have at home to interact with modern AR applications. Regarding the study about gesture elicitation for Omni-Directional Video, we used the ODV CAVE to give participants a better idea of the spatial properties of this new media type. This helped us to ensure the validity of our study.

8.2.4 External Validity

This category considers two types of threats: *interaction of selection and treatment*, and *interaction of setting and treatment*. It analyses the ability to generalize the experiment's results beyond the studied case. It is important to notice that even when we tried to invite participants with different backgrounds to our experiments, our sample cannot be considered completely heterogeneous. All the participants were university students and in many of the cases with a background in computer science. This might have affected the results of our experiments. Future experiments should consider a more heterogeneous sample. To minimize the effect of not having representative material in the study, we carefully designed three applications, one for each of three of the most representative Augmented Reality application domains. However, more experiments with a broader range of applications are needed to be able to draw more general conclusions.

8.3 Lessons Learned for 3DUIs

To summarize all our findings about multimodal 3D user interfaces for Augmented Reality and Omni-Directional Video, we present below the key aspects we learned in our studies.

Visual Feedback :

- Visual feedback should be provided without forcing the user to deviate their sight from the region of focus of the task. Performance of the user decreases when they need to shift their attention between different locations.
- It is an intuitive way for providing directional cues in outdoors navigation tasks, when blending the feedback with the environment around the user. However, it is important to provide alternatives to visual feedback when designing mobile applications. Outdoors environments can be quite challenging for visual feedback, either because lighting variations can make it harder to read the screen, or because all the different tasks that demands users' visual attention.

- It is a good feedback alternative in 3DUIs that require fine-grained interaction, such as gaming.

Auditory Feedback :

- It is demonstrated that it is best to minimize the use of loud strident sounds for auditory feedback when it is going to be played repeatedly. Even when this type of sounds are good to attract users' attention, they are annoying and stressing.
- Auditory feedback can go unnoticed in noisy environments or disturb people nearby in quiet environments, such as a library. Alternative feedback should be considered in those cases.
- It is a good alternative for improving the user experience in gaming applications, but care should be taken to avoid repetitive feedback that can be annoying; it is also a good alternative to provide directional cues in outdoor environments, as natural language can express the directions efficiently.

Tactile Feedback :

- It is best to minimize the use of vibrotactile feedback in 3DUIs that require precise movements to interact with virtual objects. Constant vibrations can increase the level of stress and negatively affect user performance.
- Vibrotactile feedback can be a good alternative to attract or guide user's attention. It is less intrusive than auditory feedback in quiet environments, and it can be provided without distracting user's visual attention.
- It is a good alternative for indoor quiet places and outdoor navigation assistants. However, it is not recommended for gaming applications that require precise interaction.

Gesture-based Interaction :

- People adjust their gestures according to the environment, e.g. other users in close proximity or the amount of effort they can do. Gesture recognizers should be designed with enough flexibility to recognize variations of the same gestures. For example, performing the zoom gesture using only fingers or with a movement of both arms, or using left and right hand indistinctly to perform one-handed gestures, such as issuing a "play" command.

Other considerations :

- Feedback should not overwhelm the user. Avoid saturating users with too much information, considering an aesthetic and minimalistic design. Avoid using the same sensory channel to provide all the feedback.
- Using an HMD with a camera attached to it for AR applications that require precise interaction with virtual objects can be tiresome for the user. Natural head movements in these situations make interaction harder, forcing the user to keep their head steady to avoid virtual objects moving while interacting with them.
- HMD can cause motion sickness. Using an HMD did not represent a clear advantage compared to a normal screen in our experiment; thus, the use of this display technology should be carefully considered.

CHAPTER 9

CONCLUSIONS

Human-Computer Interaction is a multidisciplinary research field founded, among others, on Computer Science and Psychology. It studies human-computer interfaces both from the point of view of technology and from the user experience.

Continuous advances in computer hardware, especially in mobile devices, has given computer scientists the opportunity to create more affordable 3D User Interfaces. This type of interfaces is based on 3D spatial input that typically takes place in mid-air. Augmented Reality and Omni-Directional Video applications are two examples of 3DUIs.

The usability evaluation of 3DUIs pose a number of different challenges, described for example by Dünser and Billinghurst (2011) or Billinghurst (2008). Among these challenges are the use of multiple sensory modalities, the absence of a defined and validated methodology for user testing, or the wide variety of hardware used to track the viewpoint and interactions. Thus, understanding how to combine visual, auditory and tactile feedback modalities to effectively provide important information to the user on one hand, and providing the best combination of input modalities for each type of application on the other hand, are still open questions.

This thesis presents a number of experiments and studies that provide valuable insights for the development of multimodal 3D user interfaces, specifically, for Augmented Reality and Omni-Directional Video applications. These findings can also be extrapolated to other applications with similar purposes. For example, to 3D Virtual Environments and simulators where the user has to change the position and orientation of virtual objects.

We adapted Human-Computer Interaction methodologies for user testing, for example, Nielsen's (Nielsen et al., 2004) gesture elicitation approach. We describe the methodology of each experiment and study, hoping that it can serve as a reference for future evaluation of this type of applications.

Our results show that there is no universal solution that provides the perfect combination of feedback modalities for every application. Each situation requires careful design. It is important to consider the experience of the target audience, the environment where is going to be used (indoors or outdoors, in noisy or quiet places), the type of interactions with virtual objects (e.g. precise interactions or not), etc. However, in general terms, we can confirm that multimodal feedback provides a more flexible approach in giving the user a better experience and helping them perform better.

Another interesting finding of our research has been the importance of the social ties between users that share an experience in a 3D environment. The results of the study show that people tend to adapt their gestures according to the familiarity with the users they share the space with. Participants also changed their gestures after spending more time thinking about the implications of their gestures. For example, about the amount of required effort. People change their gestures even when a short period of time passed between each session. Thus, it is important to design flexible gesture recognizers that adapt to the user. Consider, for example, the possibility to perform a zoom gesture using either both hands or just the index and thumb fingers.

The design of gesture recognizers should consider the scenario in which ODV gesture-based interfaces are going to be used. Thus, it should be allowed to adjust the scale of the gestures in such way that less intrusive interaction can be achieved, without invading the other user's personal space. The design should also take into account the location of the system, as there will be changes in user's behaviour in the comfort of a living room compared to the entrance hall of a museum.

9.1 Future Work

In order to generalize our findings, it is necessary to expand the study population. Our studies were performed using a relatively small number of participants, university or postgraduate students. A wider community with more participants from different backgrounds should be considered for future experiments.

A second area of improvement is to consider newer hardware to implement the prototypes. Computer hardware is rapidly evolving, offering better image quality, processing power, and lighter equipment, etc. Thus, the results presented here should be re-evaluated or extended considering these hardware improvements.

A good example of this new hardware are the sensors that capture information of the real scene, solving the visibility problem between real and virtual objects (see for example `http://pmdtec.com/nimbleux` or `http://www.leapmotion.com`).

It is also important to evaluate more realistic scenarios for ODV, where more than a couple of users are interacting with the content. In this situation new problems arise. For example, what should happen when two users are watching a different portion of the ODV and try to interact at the same time, e.g. one performs a zoom in and the other a pan left action. What is the correct approach to implement floor control techniques that allow users to share resources without access conflicts. We refer the reader to Dommel and Garcia-Luna-Aceves (1997) for more information about floor control techniques.

The use of public displays is a potential application area for ODV. In this context, gesture-base interaction presents an extra challenge: how to teach casual—probably inexperienced—users how to interact with the system at the same time they observe the ODV content. It is an important and interesting aspect that we hope to continue investigating in the future.

9.2 List of Publications

The following scientific publications are the result of our research:

2015

- **Gustavo Rovelo**, Francisco Abad, M.-C. Juan, Emilio Camahort. Multimodal Alerting Cues in Augmented Reality. Sent to Journal of Multimodal User Interfaces. Under reviewing process.

2014

- **Gustavo Rovelo**, Francisco Abad, M.-C. Juan, Emilio Camahort. Studying the User Experience with a Multimodal Pedestrian Navigation Assistant. In Proceedings of the International Conference on Computer Graphics Theory and Applications (GRAPP), 2015.
  CORE A - Short paper

- **Gustavo Rovelo**, Davy Vanacken, Kris Luyten, Francisco Abad, Emilio Camahort. Multi-Viewer Gesture-Based Interaction for Omni-Directional Video. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI), 2014.
  DOI: http://dx.doi.org/10.1145/2556288.2557113
  CORE A* - Full paper

2012

- **Gustavo Rovelo**, Francisco Abad, M. C. Juan and Emilio Camahort. Assessing a Multimodal User Interface in a Target Acquisition Task. Proceedings of the BCS Human Computer Interaction, People & Computers XXVI, (BCS HCI), 2012, pages 165 – 174
  CORE A - Full paper

- **Gustavo Rovelo**, Francisco Abad, M. C. Juan and Emilio Camahort. Stereoscopic vision in Desktop Augmented Reality - User performance in the presence of conflicting depth cues. Proceedings of the International Conference on Computer Graphics Theory and Applications (GRAPP), 2012, pages 460 - 465
  CORE A - Short paper

- **Gustavo Rovelo**, Francisco Abad and Emilio Camahort. A Survey on Development Tools for Mobile Augmented Reality. Proceedings of

the XXII Spanish Conference on Computer Graphics (CEIG), 2012, pages 141 - 150
DOI: 10.2312/LocalChapterEvents/CEIG/CEIG12/141-150
Full paper

2009

- **Gustavo Rovelo**, Francisco Abad, M. C. Juan y Emilio Camahort. Sistema de Realidad Aumentada para la enseñanza de Geometría. Proceedings of the XXII Spanish Conference on Computer Graphics (CEIG), 2009, pages 27 - 36
DOI: 10.2312/LocalChapterEvents/CEIG/CEIG09/027-036
Full paper

Appendices

A. THE GOAL/QUESTION/METRIC METHOD

Table A.1 presents the structure of the experiments described throughout this thesis according to the Goal/Question/Metric method (Basili et al., 1994). In this table we only consider the chapters that describe an experiment to assess the benefits of multimodal feedback for Augmented Reality applications. Chapter 7 about Omni-Directional Video Interaction has a different structure, as it describes an observational study to define the 3DUI and not an experiment with it. It is also worth noticing that we did present the hypotheses of every experiment in the form of research questions.

The number in each cell is a link to the page number where that element of the experiment is described.

Table A.1: Experiments structure according to the Goal/Question/Metric method.

	Experimental chapters			
Structure	Chapter 3	Chapter 4	Chapter 5	Chapter 6
Experiment planning	34	48	70	90
Hypotheses	42	57	74	97
Response variables	41	57	73	98
Factors	42	49	71	92
Experimental subjects	39	51	72	93
Objects of study	39	53	73	93
Experimental design	42	49	71	92
Experimental procedure	35	48	71	91
Analysis of results	42 and 43	59 and 62	75 and 80	100 and 102
Interpretation of results	45	65	84	104

REFERENCES

Ahmaniemi, T. T. and Lantz, V. T. (2009). Augmented Reality Target Finding Based on Tactile Cues. In *Proc. of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI 09, pages 335–342.

Aigner, R., Wigdor, D., Benko, H., Haller, M., Lindbauer, D., Ion, A., Zhao, S., and Koh, J. T. K. V. (2012). Understanding Mid-Air Hand Gestures: A Study of Human Preferences in Usage of Gesture Types for HCI. Technical Report MSR-TR-2012-111, Microsoft Research.

Axel, C., Ravindra, G., and Tsang, O. W. (2010). Towards Characterizing Users' Interaction with Zoomable Video. In *Proc. of the ACM Workshop on Social, Adaptive and Personalized Multimedia Interaction and Access*, SAPMIA, pages 21–24.

Azuma, R. T. (1997). A Survey of Augmented Reality. *Presence: Teleoperators and Virtual Environments*, 6(4):355–385.

Basili, V. R., Caldiera, G., and Rombach, H. D. (1994). The Goal Question Metric Approach. In *Encyclopedia of Software Engineering*. Wiley.

Baudel, T. and Beaudouin-Lafon, M. (1993). CHARADE: Remote Control of Objects Using Free-Hand Gestures. *Communications of the ACM*, 36(7):28–35.

Benko, H. (2009). Beyond flat surface computing: Challenges of depth-aware and curved interfaces. In *Proc. of the ACM International Conference on Multimedia*, MM, pages 935–944.

Benko, H. and Wilson, A. D. (2010a). Multi-point interactions with immersive omnidirectional visualizations in a dome. In *Proc. of the ACM Int. Conference on Interactive Tabletops and Surfaces*, ITS, pages 19–28.

Benko, H. and Wilson, A. D. (2010b). Pinch-The-Sky Dome: Freehand Multi-Point Interactions with Immersive Omni-Directional Data. In *CHI Extended Abstracts on Human Factors in Computing Systems*, pages 3045–3050.

Bier, E. A., Stone, M. C., Pier, K., Buxton, W., and DeRose, T. D. (1993). Toolglass and Magic Lenses: The See-through Interface. In *Proc. of the 20th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '93, pages 73–80.

Billinghurst, M. (2008). Usability Testing of Augmented/Mixed Reality Systems. In *ACM SIGGRAPH ASIA 2008 courses*, SIGGRAPH Asia 08, pages 1–13.

Blattner, M. and Glinert, E. (1996). Multimodal Integration. *IEEE MultiMedia*, 3:14–24.

Bleumers, L., den Broeck, W. V., Lievens, B., and Pierson, J. (2012). Seeing the bigger picture: A user perspective on 360° TV. In *Proc. of the European Conference on Interactive TV and Video*, EuroiTV, pages 115 – 124.

Bowman, D. A., Coquillart, S., Froehlich, B., Hirose, M., Kitamura, Y., Kiyokawa, K., and Stuerzlinger, W. (2008). 3D User Interfaces: New Directions and Perspectives. *Computer Graphics and Applications, IEEE*, 28:20–36.

Bowman, D. A., Kruijff, E., LaViola, J. J., and Poupyrev, I. (2004). *3D User Interfaces: Theory and Practice.* Addison Wesley Longman Publishing Co., Inc.

Bresciani, J.-P., Dammeier, F., and Ernst, M. O. (2008). Tri-modal Integration of Visual, Tactile and Auditory Signals for the Perception of Sequences of Events. *Brain Research Bulletin*, 75(6):753–60.

Burke, J. L., Prewett, M. S., Gray, A. A., Yang, L., Stilson, F. R. B., Coovert, M. D., Elliot, L. R., and Redden, E. (2006). Comparing the Effects of Visual-Auditory and Visual-Tactile Feedback on User Performance: A Meta-Analysis. In *Proceedings of the 8th International Conference on Multimodal Interfaces*, ICMI 06, pages 108–117.

Charoenchaimonkon, E., Janecek, P., Dailey, M., and Atiwong Suchato, A. (2010). A comparison of audio and tactile displays for non-visual target selection tasks. In *International Conference on User Science and Engineering*, i-USEr 2010, pages 238–243.

Clark, R. A., Richmond, K., and King, S. (2004). Festival 2 – build your own general purpose unit selection speech synthesiser. In *Proc. 5th ISCA Workshop on Speech Synthesis*.

Coutaz, J., Nigay, L., Salber, D., Blandford, A., May, J., and Young, R. M. (1995). Four easy pieces for assessing the usability of multimodal interaction: the CARE properties. *InterAct*, 95:115–120.

Diederich, A. and Colonius, H. (2004). Bimodal and trimodal multisensory enhancement: Effects of stimulus onset and intensity on reaction time. *Perception and Psychophysics*, 66(8):1388–1404.

Dommel, H.-P. and Garcia-Luna-Aceves, J. J. (1997). Floor control for multimedia conferencing and collaboration. *Multimedia Systems*, 5(1):23–38.

Dünser, A. and Billinghurst, M. (2011). *Evaluating Augmented Reality Systems*. Springer.

Dünser, A., Grasset, R., and Billinghurst, M. (2008). A Survey of Evaluation Techniques Used in Augmented Reality Studies. In *ACM SIGGRAPH ASIA 2008 courses*, pages 5:1–5:27.

El-Shimy, D., Marentakis, G., and Cooperstock, J. R. (2009). Tech-note: Multimodal Feedback in 3D Target Acquisition. In *Proceedings of the 2009 IEEE Symposium on 3D User Interfaces*, 3DUI 09, pages 95–98.

Ernst, M. O. (2006). A Bayesian View on Multimodal Cue Integration. *Perception*, 131(6):105–131.

Ernst, M. O. and Bülthoff, H. H. (2004). Merging the Senses Into a Robust Percept. *Trends in Cognitive Sciences*, 8(4):162–169.

Fikkert, W., van der Vet, P., van der Veer, G., and Nijholt, A. (2010). Gestures for Large Display Control. In *Proc. of the International Gesture Workshop*, GW, pages 245–256.

Freeman, W. T. and Weissman, C. D. (1995). Television Control by Hand Gestures. In *Proc. of the International Workshop on Automatic Face and Gesture Recognition*, FG, pages 179–183.

Fröhlich, P., Oulasvirta, A., Baldauf, M., and Nurminen, A. (2011). On the Move, Wirelessly Connected to the World. *Com. of the ACM*, 54(1):132–138.

Fuhrmann, A., Loffelmann, H., and Schmalstieg, D. (1997). Collaborative Augmented Reality: Exploring Dynamical Systems. In *IEEE Visualization*, pages 459–462.

Grandhi, S. A., Joue, G., and Mittelberg, I. (2011). Understanding Naturalness and Intuitiveness in Gesture Production: Insights for Touchless Gestural Interfaces. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, pages 821–824.

Grantcharov, T. P. and Funch-Jensen, P. (2009). Can everyone achieve proficiency with the laparoscopic technique? learning curve patterns in technical skills acquisition. *The American Journal of Surgery*, 197(4):447–449.

Haythornthwaite, C. and Wellman, B. (1998). Work, Friendship, and Media Use for Information Exchange in a Networked Organization. *Journal of the American Society for Information Science*, 49(12):1101–1114.

Henze, N., Löcken, A., Boll, S., Hesselmann, T., and Pielot, M. (2010). Free-Hand Gestures for Music Playback: Deriving Gestures With a User-Centred Process. In *Proc. of the Int. Conference on Mobile and Ubiquitous Multimedia*, MUM, pages 1–10.

Herbst, Iris, S. J. (2005). Comparing Force Magnitudes by Means of Vibro-Tactile, Auditory, and Visual Feedback. In *International Workshop on Haptic Audio Visual Environments and their Applications*, pages 67–71.

Hespanhol, L., Tomitsch, M., Grace, K., Collins, A., and Kay, J. (2012). Investigating Intuitiveness and Effectiveness of Gestures for Free Spatial Interaction with Large Displays. In *Proc. of the International Symposium on Pervasive Displays*, PerDis, pages 1–6.

Hinrichs, U. and Carpendale, S. (2011). Gestures in the Wild: Studying Multi-Touch Gesture Sequences on Interactive Tabletop Exhibits. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, pages 3023–3032.

Hoggan, E., Williamson, J., Oulasvirta, A., Nacenta, M., Kristensson, P. O., and Lehtiö, A. (2013). Multi-Touch Rotation Gestures: Performance and Ergonomics. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, pages 3047–3050.

Irawati, S., Green, S., Billinghurst, M., Duenser, A., and Ko, H. (2006a). An Evaluation of an Augmented Reality Multimodal Interface Using Speech and Paddle Gestures. In *Proceedings of the 16th international conference on Advances in Artificial Reality and Tele-Existence*, ICAT 06, pages 272–283.

Irawati, S., Green, S., Billinghurst, M., Duenser, A., and Ko, H. (2006b). Move the Couch Where?: Developing an Augmented Reality Multimodal Interface. In *IEEE/ACM International Symposium on Mixed and Augmented Reality*, ISMAR 2006, pages 183–186.

Jacob, R., Mooney, P., and Winstanley, A. C. (2011). Guided by Touch: Tactile Pedestrian Navigation. In *Proc. of the 1st Int. Workshop on Mobile Location-based Service*, MLBS, pages 11–20.

Jameson, A. (2002). Usability Issues and Methods for Mobile Multimodal Systems. In *Proc. of the ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments.*

Jones, B., Benko, H., Ofek, E., and Wilson, A. (2013). IllumiRoom: Peripheral projected illusions for interactive experiences. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, pages 869–878.

Kato, H. and Billinghurst, M. (1999). Marker Tracking and HMD Calibration for a Video-based Augmented Reality Conferencing System. In *2nd IEEE and ACM International Workshop on Augmented Reality*, IWAR 99, pages 85–94.

Kortum, P. (2008). *HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces.* Morgan Kaufmann Publishers Inc.

Kruijff, E., Swan, J. E., and Feiner, S. (2010). Perceptual Issues in Augmented Reality Revisited. In *9th IEEE International Symposium on Mixed and Augmented Reality*, ISMAR, pages 3–12.

Lee, M. and Billinghurst, M. (2008). A Wizard of Oz Study for an AR Multimodal Interface. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, pages 249–256.

Levy, B. and Mobasheri, M. (2014). The principles of safe laparoscopic surgery. *Surgery (Oxford)*, 32(3):145–148.

Liljedahl, M., Lindberg, S., Delsing, K., Polojärvi, M., Saloranta, T., and Alakärppä, I. (2012). Testing Two Tools for Multimodal Navigation. *Advances in Human-Computer Interaction*, 2012.

Lindeman, R., Yanagida, Y., Sibert, J., and Lavine, R. (2003). Effective Vibrotactile Cueing in a Visual Search Task. In *Proc. of Human-Computer Interaction*, Interact 2003, pages 89–96.

Lindeman, R. W., Sibert, J. L., Mendez-Mendez, E., Patil, S., and Phifer, D. (2005). Effectiveness of Directional Vibrotactile Cuing on a Building-Clearing Task. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 271–280.

Loeliger, E. and Stockman, T. (2013). Wayfinding without Visual Cues: Evaluation of an Interactive Audio Map System. *Interacting with Computers.*

Looser, J., Grasset, R., and Billinghurst, M. (2007). A 3D Flexible and Tangible Magic Lens in Augmented Reality. In *Proc. of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, ISMAR '07, pages 1–4.

Looser, J., Grasset, R., Seichter, H., and Billinghurst, M. (2006). OSGART - A Pragmatic Approach to MR. In *Industrial Workshop*, ISMAR, pages 22–25.

Macq, J.-F., Verzijp, N., Aerts, M., Vandeputte, F., and Six, E. (2011). Demo: Omnidirectional Video Navigation on a Tablet PC Using a Camera-Based Orientation Tracker. In *Proc. of the ACM/IEEE International Conference on Distributed Smart Cameras*, ICDSC, pages 1–2.

Magnusson, C., Molina, M., Rassmus-Gröhn, K., and Szymczak, D. (2010). Pointing for Non-visual Orientation and Navigation. In *Proc. of the 6th Nordic Conf. on Human-Computer Interaction: Extending Boundaries*, NordiCHI, pages 735–738.

Maltz, M. and Shinar, D. (2007). Imperfect In-Vehicle Collision Avoidance Warning Systems Can Aid Distracted Drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, 10(4):345–357.

Milgram, P. and Kishino, F. (1994). A Taxonomy of Mixed Reality Visual Displays. *IEICE Transactions on Information Systems*, E77-D(12):1321–1329.

Morris, M. R., Wobbrock, J. O., and Wilson, A. D. (2010). Understanding Users' Preferences for Surface Gestures. In *Proc. of Graphics Interface*, GI, pages 261–268.

Nancel, M., Wagner, J., Pietriga, E., Chapuis, O., and Mackay, W. (2011). Mid-Air Pan-and-Zoom on Wall-Sized Displays. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, pages 177–186.

Neng, L. A. R. and Chambel, T. (2010). Get Around 360° Hypervideo. In *Proc. of the International Academic MindTrek Conference*, MindTrek, pages 119–122.

Nielsen, M., Störring, M., Moeslund, T. B., and Granum, E. (2004). A Procedure for Developing Intuitive and Ergonomic Gesture Interfaces for HCI. *Gesture Based Communication in Human-Computer Interaction*, 2915:409–420.

Oron-Gilad, T., Downs, J. L., Gilson, R. D., and Hancock, P. A. (2007). Vibrotactile Guidance Cues for Target Acquisition. *IEEE Transactions on Systems, Man and Cybernetics. Part C, Applications and Reviews*, 37:993–1004.

Oulasvirta, A., Tamminen, S., Roto, V., and Kuorelahti, J. (2005). Interaction in 4-second bursts: The fragmented nature of attentional resources in mobile hci. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 919–928.

Park, W. and Han, S. H. (2013). Intuitive Multi-Touch Gestures for Mobile Web Browsers. *Interacting with Computers*, 25(5):335–350.

Peltonen, P., Kurvinen, E., Salovaara, A., Jacucci, G., Ilmonen, T., Evans, J., Oulasvirta, A., and Saarikko, P. (2008). It's Mine, Don't Touch!: Interactions at a Large Multi-Touch Display in a City Centre. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, pages 1285–1294.

Pielot, M. and Boll, S. (2010). Tactile Wayfinder: Comparison of Tactile Waypoint Navigation with Commercial Pedestrian Navigation Systems. In *Proc. of the Int. Conf. on Pervasive Computing*, Pervasive, pages 76–93.

Pielot, M., Heuten, W., Zerhusen, S., and Boll, S. (2012a). Dude, Where's My Car?: In-situ Evaluation of a Tactile Car Finder. In *Proc. of the 7th Nordic Conf. on Human-Computer Interaction: Making Sense Through Design*, NordiCHI, pages 166–169.

Pielot, M., Poppinga, B., Heuten, W., and Boll, S. (2012b). PocketNavigator: Studying Tactile Navigation Systems In-situ. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI, pages 3131–3140.

Pierno, A. C., Caria, A., Glover, S., and Castiello, U. (2005). Effects of Increasing Visual Load on Aurally and Visually Guided Target Acquisition in a Virtual Environment. *Applied Ergonomics*, 36(3):335–343.

Prewett, M., Elliott, L., Walvoord, A., and Coovert, M. (2012). A Meta-Analysis of Vibrotactile and Visual Information Displays for Improving Task Performance. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(99):123–132.

Prewett, M., Yang, L., Stilson, F., Gray, A., Coovert, M., Burke, J., Redden, E., and Elliot, L. (2006). The Benefits of Multimodal Information: a Meta-Analysis Comparing Visual and Visual-Tactile Feedback. In *Proc. of the 8th international Conference on Multimodal Interfaces*, pages 333–338.

Rach, S. and Diederich, A. (2006). Visualtactile integration: does stimulus duration influence the relative amount of response enhancement? *Experimental Brain Research*, 173(3):514–520.

Rach, S., Diederich, A., and Colonius, H. (2011). On quantifying multisensory interaction effects in reaction time and detection rate. *Psychological Research*, 75(2):77–94.

Raisamo, R., Nukarinen, T., Pystynen, J., Mäkinen, E., and Kildal, J. (2012). Orientation Inquiry: A New Haptic Interaction Technique for Non-visual Pedestrian Navigation. In *Proc. of the 2012 Int. Conf. on Haptics: Perception, Devices, Mobility, and Communication - Volume Part II*, EuroHaptics, pages 139–144.

Rovelo, G., Vanacken, D., Luyten, K., Abad, F., and Camahort, E. (2014). Multi-viewer gesture-based interaction for omni-directional video. In *Conference on Human Factors in Computing Systems*, CHI, pages 4077–4086.

Smith, C., Clegg, B. A., Heggestad, E. D., and Hopp-Levine, P. J. (2009). Interruption Management: A Comparison of Auditory and Tactile Cues for Both Alerting and Orienting. *International Journal of Human-Computer Studies*, 67(9):777–786.

Stellmach, S., Jüttner, M., Nywelt, C., Schneider, J., and Dachselt, R. (2012). Investigating Freehand Pan and Zoom. In *Mensch & Computer*, pages 303–312.

Sun, M., Ren, X., and Cao, X. (2010). Effects of Multimodal Error Feedback on Human Performance in Steering Tasks. *Journal of Information Processing*, 18:284–292.

Szymczak, D., Magnusson, C., and Rassmus-Gröhn, K. (2012). Guiding Tourists Through Haptic Interaction: Vibration Feedback in the Lund Time Machine. In *Proc. of the 2012 Int. Conf. on Haptics: Perception, Devices, Mobility, and Communication - Volume Part II*, EuroHaptics, pages 157–162.

Trevisan, D. G., Vanderdonckt, J., and Macq, B. (2004). Conceptualising mixed spaces of interaction for designing continuous interaction. *Virtual Reality*, 8(2):83–95.

Unger, B. J., Nicolaidis, A., Thompson, A., Klatzky, R. L., Hollis, R. L., Berkelman, P. J., and Lederman, S. (2002). Virtual Peg-in-Hole Performance Using a 6-DOF Magnetic Levitation Haptic Device: Comparison with Real Forces and with Visual Guidance Alone. In *Proceedings of the 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, HAPTICS 02, pages 263–280.

Vainio, T. (2009). Exploring Multimodal Navigation Aids for Mobile Users. In *Proc. of the 12th IFIP TC 13 Int. Conf. on Human-Computer Interaction: Part I*, INTERACT, pages 853–865.

Van Erp, J. B. and Van Veen, H. A. H. C. (2004). Vibrotactile In-Vehicle Navigation System. *Transportation Research Part F: Traffic Psychology and Behaviour*, 7(4-5):247–256.

Van Krevelen, D. and Poelman, R. (2010). A survey of Augmented Reality Technologies, Applications and Limitations. *International Journal Virtual Reality*, 9(2):1–20.

Vermeulen, J., Luyten, K., van den Hoven, E., and Coninx, K. (2013). Crossing the Bridge over Norman's Gulf of Execution: Revealing Feedforward's True Identity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1931–1940.

Walter, R., Bailly, G., and Müller, J. (2013). StrikeAPose: Revealing Mid-air Gestures on Public Displays. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, pages 841–850.

Wickens, C. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3(2):159–177.

Wickens, C. D., Goh, J., Helleberg, J., and Talleur, D. A. (2002). Modality Differences in Advanced Cockpit Displays: Comparing Auditory Vision and Redundancy for Navigational Communications and Traffic Awareness. Technical Report ARL-02-8/NASA-02-6, Nasa Ames Research Center.

Wobbrock, J. O., Aung, H. H., Rothrock, B., and Myers, B. A. (2005). Maximizing the Guessability of Symbolic Input. In *CHI Extended Abstracts on Human Factors in Computing Systems*, pages 1869–1872.

Wobbrock, J. O., Morris, M. R., and Wilson, A. D. (2009). User-Defined Gestures for Surface Computing. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, pages 1083–1092.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2000). *Experimentation in Software Engineering: An Introduction.* Kluwer Academic Publishers.

Wozny, D., Beierholm, U., and Shams, L. (2008). Human Trimodal Perception Follows Optimal Statistical Inference. *Journal of Vision*, 8(3):24.1–11.

Zhou, Z., Cheok, a., Yang, X., and Qiu, Y. (2004). An Experimental Study on the Role of Software Synthesized 3D Sound in Augmented Reality Environments. *Interacting with Computers*, 16(5):989–1016.

Zoric, G., Engström, A., Barkhuus, L., Hidalgo, J. R., and Kochale, A. (2013). Gesture Interaction with Rich TV Content in the Social Setting. In *CHI workshop on Exploring and Enhancing the User Experience for TV*, TVUX.

ONLINE REFERENCES

13th Lab AB (2012). PointCloud. Retrieved June 2nd., 2014, from: http://pointcloud.io/.

20th Century Fox (2002). Minority Report. Retrieved June 3rd., 2015, from: http://www.imdb.com/title/tt0181689/.

AirPano project (2014). Spherical 360 Video. Retrieved June 3rd., 2015, from: http://www.airpano.com/.

BBC (2014). Reef sharks in 360. Retrieved June 3rd., 2015, from: http://www.bbc.co.uk/oceans/360/reefsharks/.

GameLoft (2013). Tom Clancy's Rainbow Six: Shadow Vanguard. Retrieved June 3rd., 2015, from: http://www.gameloft.com/mobile-games/rainbow-6-shadow-vanguard-android/.

Geerds, J. (2014). Freedom360 - 360rig. Retrieved June 3rd., 2015, from: http://freedom360.us/.

Google Inc. (2013). Google Glass. Retrieved June 3rd., 2015, from: http://www.google.com/glass/start/.

Google Inc. (2014). Project Tango. Retrieved June 3rd., 2015, from: https://www.google.com/atap/projecttango.

ONLINE REFERENCES

Hewett, Baecker, Card, Carey, Gasen, Mantei, Perlman, Strong, and Verplank (1996). Acm sigchi curricula for human-computer interaction. In *ACM, Association for Computer Machinery.* Retrieved June 3rd., 2015, from: http://sigchi.org/cdg/cdg2.html.

Laforest, M. (2008). Wiiuse - The Wiimote C Library. Retrieved June 3rd., 2015, from: http://sourceforge.net/projects/wiiuse.

Leap Motion, Inc. (2014). Leap Motion. Retrieved June 3rd., 2015, from: https://www.leapmotion.com/.

Marvel Entertainment Inc. (2009). I am iron man. Retrieved June 3rd., 2015, from: http://www.iamironman2.com.

Microsoft Inc. (2013). Kinect for Windows Human Interface Guidelines v1.8.0. Retrieved June 3rd., 2015, from: http://msdn.microsoft.com/en-us/library/jj663791.aspx.

Novint Technologies Inc. (2012). Novint falcon. Retrieved June 3rd., 2015, from: http://www.novint.com/.

Occipital Inc. (2014). STRUCTURE. Retrieved June 3rd., 2015, from: http://structure.io/.

Oculus VR. (2014). Oculus rift. Retrieved June 3rd., 2015, from: http://www.oculus.com/.

OpenNI Organization (2010). OpenNI Library. Retrieved June 3rd., 2015, from: http://www.openni.org.

Samsumg Corporation (2013). All of the Smart TV Gestures. Retrieved June 3rd., 2015, from: http://www.samsung.com/global/microsite/tv/common/ guide_book_5p_vi/main.html.

Sony Computer Entertainment (2012). Wonderbook: Book of Spells. Retrieved June 3rd., 2015, from: http://us.playstation.com/games-and-media/games/wonderbook-book-of-spells-ps3.html.

Sony Computer Entertainment, E. (2009). Invizimals. Retrieved June 3rd., 2015, from: http://invizimals.eu.playstation.com/.

SPRX Mobile (2010). Layar. Retrieved June 3rd., 2015, from: http://www.layar.com/.

Wikitude GmbH (2011). Wikitude. Retrieved June 3rd., 2015, from: http://www.wikitude.com/.