UHH Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Human-Robot Interaction in Augmented Virtuality: Perception, Cognition and Action in 360° Video-based Robotic Telepresence Systems

This dissertation is submitted for the degree of
Doctor of Philosophy
at the Faculty of Mathematics, Informatics and Natural Sciences

Jingxin Zhang

Human-Computer Interaction
Department of Informatics
Universität Hamburg

November 2021

First reviewer:	Prof. Dr. Frank Steinicke
Second reviewer:	Prof. Dr. Jianwei Zhang
Chair of the examination commission:	Prof. Dr. -Ing. Timo Gerkmann
Deputy chair of the examination commission:	Prof. Dr. Simone Frintrop
Date of the disputation:	08.11.2021

To My Loving Parents . . .

Declaration

I hereby declare, on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids.

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Jingxin Zhang
November 2021

Acknowledgements

First of all, I would like to acknowledge my Ph.D. supervisor, Prof. Dr. Frank Steinicke, who led me into the virtual world. As an internationally renowned scientist in the area of VR/AR, his knowledge and scientific attitude have influenced me a lot. A communication or discussion with him can always spark my thoughts, answer my doubts and enrich my research. I am very grateful for his patient guidance, helpful suggestions, and great support throughout the work on this dissertation. It was really an honor and a pleasure to finish my Ph.D. degree under his supervision.

Furthermore, I want to thank all the reviewers, their perspectives and feedback are very instrumental for this dissertation. And thanks go to everyone who took the time to proofread this dissertation. In particular, I would like to thank all the people of our HCI group, without their help and support, this work would not have been possible. I have been extremely lucky to join a research group, in which everybody is very friendly, would like to cooperate and share, and never hesitate to help each other. Time passes, but nice memories will always stay. I will never forget those pleasant times that we have spent together. Special thanks to the voluntary participants in the experiments conducted throughout the work on this dissertation.

Moreover, I would like to thank my friends for their support. Living and studying in Hamburg for almost 5 years, I have gained a lot of friends, who make my life more pleasant and wonderful. Their warm encouragement and kind help will always remain in my heart.

Finally, I owe deep gratitude to my loving parents for their great support and infinite love in the last 31 years. When I was upset, they gave me the best encouragement and help me to keep patience; When I achieved something, they shared my happiness, reminded me to stay modest and focus on further targets. I am incredibly lucky to have them in my life. At last, I would like to thank my partner Xiaohui for her constant support, limitless giving, and great understanding. We met each other in Hamburg, and she let me know what is important in my life.

Abstract

Telepresence refers to a set of technologies, which enables humans to visit a remote environment (RE) and interact with objects and other people without the requirement for physical travel. Ideally, users would feel a sense of presence in the RE during the teleoperation tasks. However, current telepresence systems restrict this sensation as well as other relevant user experience severely. For example, single webcams for capturing the RE can provide only a limited illusion of spatial presence. Moreover, movement control of mobile platforms are often limited to traditional input devices like keyboards or mice, which lacks natural methods to map the movements of human users to motions of the robotic platform. In order to address these challenges, this dissertation proposes a novel solution by integrating 360° cameras and virtual reality (VR) head-mounted displays (HMDs) into telepresence systems and exploiting the human user's natural walking as locomotion method for platform control. In addition, in order to allow a human user to explore a RE, which is significantly larger than the tracking area, redirected walking (RDW) method is applied to compress the human user's locomotion. With this solution, the user receives an omnidirectional view of the RE with improved spatial information, immersion and sense of presence.

However, such re-design and improvement of telepresence solutions could also influence human's cognitive and perceptive capability as well as preference of interactive operations. For instance, due to the usage of 360° cameras, when users remotely visit an apartment and look around, the robotic platform located inside the real apartment does not need any camera rotations, since 360° cameras could provide users with an omnidirectional view of the apartment instead of some limited field of views in traditional webcam-based telepresence. This issue makes it very challenging for the host, who presents the apartment to the user and interacts with the telepresence robot, to understand where the user is looking at or interested in, which as a result reduces interactivity and efficiency. Furthermore, compared with a human agent, when interacting with a robotic surrogate, the host could only get very limited (social) cues during the interaction process to help understanding and predicting the user's intention.

The main focus of this dissertation is to solve these challenges of current telepresence systems, and to propose and evaluate potential solutions. Contributions of this dissertation could be summarized as follows:

(i) Development of an omnidirectional telepresence solution with 360° camera, VR HMD and mobile robotic platform;

(ii) Evaluation of global illumination and shadow effect of virtual objects on object localization tasks and user experience in 360° video-based mixed reality (MR) environments;

(iii) Psychophysical evaluations to quantify to what extent humans can be unknowingly redirected on virtual paths in a 360° RE;

(iv) Design and study of two potential solutions to indicate a user's gaze during 360° video-based telepresence with a focus on the effects of distance, display technique, yaw and pitch angle on human's localization accuracy;

(v) Introduction of a MR avatar upper body, which was designed for subtle communication of the velocity of the robotic platform it is attached to. Furthermore, the effectiveness of this method as well as the proxemics, when human users have dynamic interactions with the avatar-robot surrogate, was also explored and evaluated.

The findings of this thesis provide novel insights into the telepresence system design and the interaction between humans and telepresence systems.

Zusammenfassung

Als Telepräsenz klassifiziert man eine Reihe von Technologien, die es Menschen ermöglichen, eine entfernte Umgebung („entf. Umg.“) zu besuchen und mit Objekten und anderen Menschen zu interagieren, ohne physisch reisen zu müssen. Im Idealfall würden Benutzer/innen während Teleoperations-Aufgaben in der entf. Umg. ein Gefühl der Präsenz spüren. Aktuelle Telepräsenzsysteme schränken jedoch diese Empfindung sowie andere relevante Benutzererfahrungen stark ein. Beispielsweise können einzelne Webcams in der entf. Umg. nur eine begrenzte Illusion der räumlichen Präsenz vermitteln. Darüber hinaus ist die Steuerung mobiler Plattformen oft auf herkömmliche Eingabegeräte wie Maus und Tastatur beschränkt, die bei der Abbildung natürlicher menschlicher Bewegungen auf die Bewegungen einer Roboterplattform Defizite aufweisen. Um diese Herausforderungen anzugehen, beschreibt diese Dissertation eine neuartige Lösung durch die Integration von 360° Kameras und „Virtuelle Realitäts“ (VR) Brillen (HMDs) in Telepräsenzsystemen, bei der das natürliche Gehen menschlicher Benutzer/innen als Fortbewegungsmethode für die Plattformsteuerung verwendet wird. Außerdem wird es einem/einer menschlichen Benutzer/in ermöglicht, eine entf. Umg. zu erkunden, die signifikant größer ist als der Tracking-Bereich des Nutzers, indem die „Redirected Walking“ Methode zur Komprimierung der Bewegungen verwendet wird. Bei dieser Lösung erhält der/die Anwender/in eine omnidirektionale Ansicht der entf. Umg. mit verbesserter räumlicher Information, Immersion und Präsenz.

Eine derartige Neugestaltung und Verbesserung von Telepräsenzlösungen könnte jedoch auch die kognitiven Fähigkeiten und Wahrnehmung eines Menschen sowie den präferierten interaktiven Modus Operandi beeinflussen. Wenn ein/e Benutzer/in beispielsweise unter der Verwendung von 360° Kameras eine Wohnung aus der Ferne besucht und sich umschaut, muss die Roboter-Plattform in der entfernten Wohnung keine Kamerarotationen durchführen, da 360° Kameras bereits eine omnidirektionale Ansicht anstatt eines begrenzten (Webcam-) Sichtfelds präsentieren. Dies ist ein Problem für die Gastgeber, die mit dem Telepräsenzroboter interagieren und einem/er Benutzer/in das Apartment vorstellen, da es schwer ist zu vermitteln wo der/die Nutzer/in gerade hinschaut oder was sein/ihr Interesse geweckt hat, was wiederum zu einer reduzierten Interaktion und Effizienz führt. Darüber hinaus konnten Gastgeber im Vergleich zu einem menschlichen Agenten, bei der Interaktion mit einem Roboter-Ersatz, nur sehr begrenzt auf (soziale) Stimuli reagieren, die während des Interaktionsprozesses zum Verständnis beitragen und die Absicht der Interaktionspartner offenlegen.

Der Schwerpunkt dieser Dissertation liegt bei der Lösung dieser Herausforderungen aktueller Telepräsenzsysteme und darauf, mögliche Lösungen vorzuschlagen und zu bewerten. Beiträge dieser Dissertation können wie folgt zusammengefasst werden:

(i) Entwicklung einer omnidirektionalen Telepräsenzlösung mit 360° Kamera, VR HMD und einer mobilen Roboterplattform;

(ii) Bewertung der globalen Beleuchtung und des Schattenwurfs virtueller Objekte auf objektlokalisierungs-Aufgaben und Benutzererfahrungen in 360°-videobasierten Mixed Reality (MR)-Umgebungen;

(iii) Psychophysikalische Evaluation zur Quantifizierung, inwieweit Menschen unwissentlich auf virtuellen Pfaden in 360° entf. Umg. umgelenkt werden können;

(iv) Design und Evaluation von zwei potenziellen Lösungen, um den Blick eines/einer Benutzers/Benutzerin bei 360° videobasierter Telepräsenz anzuzeigen. Dabei liegt der Fokus auf die Auswirkungen von Distanz, Display Technologie, sowie Gier- und Nick-Winkel (yaw & pitch) auf die Lokalisierungsgenauigkeit des Menschen;

(v) Einführung eines MR-Avatar-Oberkörpers zur Befestigung an einer Roboterplattform um die Geschwindigkeit dieser subtil vermitteln zu können. Außerdem wurde die Wirksamkeit dieser Methode, sowie die Proxemik, wenn menschliche Benutzer/innen dynamische Interaktionen mit dem Avatar-Roboter-Ersatz durchführen, ebenfalls erforscht und ausgewertet.

Die Ergebnisse dieser Arbeit bieten neue Einblicke in das Telepräsenzsystemdesign und die Interaktion zwischen Menschen und Telepräsenzsystemen.

Table of contents

Abstract **ix**

Zusammenfassung **xi**

1 Introduction **1**

1.1 Motivation 1

1.2 Problem Statement 2

1.3 Application Scenario 3

1.4 Research Questions 4

1.5 Overview 7

1.6 Publications 8

2 Fundamentals **11**

2.1 Telepresence 11

2.1.1 Definition and Applications 11

2.1.2 Robotic Telepresence 13

2.1.3 Telepresence and VR 14

2.2 Immersion and Presence 15

2.2.1 Immersion 15

2.2.2 Presence 16

2.2.3 Related Research on Immersion and Presence 16

2.3 360° VR 18

2.3.1 Panorama Images 18

2.3.2 Related Research on 360° VR 21

2.3.3 360° VR in Telepresence 23

2.4 Human-Robot Interaction & Robotic Surrogate 24

2.4.1 Human-Robot Interaction 24

2.4.2 Robotic Surrogate 26

2.5 Perception 28
2.5.1 Human Perceptual System 28
2.5.2 Perception of Self-Translation 31
2.5.3 Perception of Self-Rotation 33
2.5.4 Perception of Social Cues 34

3 360° VR-based Telepresence System 37
3.1 System Design and Development 37
3.1.1 Motivation 37
3.1.2 System Design 38
3.1.3 Implementation 40
3.2 Global Illumination and Shadows in 360° MR Environments 42
3.2.1 Motivation 42
3.2.2 Related Work 43
3.2.3 User Study 44
3.2.4 Results 46
3.2.5 Discussion 48
3.3 Conclusion 49

4 Natural Exploration of Redirection in 360° Tele-mediated Space 51
4.1 Motivation 51
4.2 Related Works 52
4.2.1 Locomotion 52
4.2.2 Detection Thresholds 52
4.3 Concept and Challenges 54
4.4 RDW Gains in 360° Virtual Environment 55
4.4.1 Translation Gains 55
4.4.2 Rotation Gains 56
4.4.3 Other Gains 57
4.5 Experiment Fundamentals 57
4.5.1 Hardware Setup 57
4.5.2 Two-Alternative Forced-Choice Task 58
4.6 Experiment 1: Detection Thresholds for Translation Gains 59
4.6.1 Methods 59
4.6.2 Participants 61
4.6.3 Results 61
4.6.4 Discussion 62

4.7 Experiment 2: Detection Thresholds for Rotation Gains 63
4.7.1 Methods 63
4.7.2 Participants 65
4.7.3 Results 66
4.7.4 Discussion 67
4.8 Post-Questionnaires 68
4.9 General Discussion 69
4.10 Conclusion 71

5 Natural Interacting with Avatar-Robot Telepresent Surrogate 73
5.1 Rapid Localization of Teleoperator Gaze in 360° Hosted Telepresence . . . 74
5.1.1 Motivation 74
5.1.2 Related Works 76
5.1.3 Avatar Display and Prototype Setup 78
5.1.4 Hypotheses 80
5.1.5 Experiment 80
5.1.6 Results 84
5.1.7 Discussion 87
5.1.8 Conclusion 93
5.2 Evaluation of Proxemics in Dynamic Interaction with an Avatar Robot . . . 94
5.2.1 Motivation 94
5.2.2 Related Works 95
5.2.3 Hypotheses 96
5.2.4 Experimental Setup 97
5.2.5 Methods 99
5.2.6 Results 102
5.2.7 Discussion 104
5.2.8 Conclusion 107

6 Conclusion 109
6.1 Discussion 109
6.2 Outlook 112

References 115

Appendix A Questionnaires 141

Chapter 1

Introduction

1.1 Motivation

Human's history could be abstracted and deemed as a book consisting of exploration and discovery stories of the world surroundings. For thousands of years, the challenge of travelling distance has never weakened our ambition and curiosity to reach somewhere we have never been before or the places where we are interested in. Traditional traffic solutions can reduce the distance barrier, but require time and increase physical burdens.

With the advancement of information technology, more technical solutions like visual telephone, tele-conference system and various instant video-messaging applications have provided more available and sustainable options for us to interact with people in remote environments (REs). These technical solutions could help to overcome the constraints of geographical location and enable users to tele-communicate with people and visit distant environment in real-time without the requirements of physical travel. These technologies are often referred to as "*Telepresence*" [199] or "*Telexistence*" [299].

In recent decades, telepresence technology has become increasingly popular in our daily life, and such technology has provided solutions in different application domains ranging from tourism [66, 222], business [283], education [30, 83] to health care [197, 31] and is gradually changing our life style, in particular, in times of global crisis such as pandemics. Theoretically, ideal telepresence should create highly realistic sensation of existence in REs for the users without the requirement of actual travel [299]. However, in currently available telepresence systems, the sensation of existence in a RE is always undermined to some extent due to different design or technical reasons. In addition, a lack of immersive illusion during telepresence constrains users' exploration of REs. As a result, users could not get a natural and comfortable experience as well as an efficient task performance during telepresence, which may even cause extra mental burden or motion sickness [116].

All of these problems and limitations motivated the work of this dissertation, which aims to improve current telepresence technology and evaluate novel solutions regarding usability, perception and interaction performance.

1.2 Problem Statement

Since the concept of *telepresence* was originally proposed in 1980 by Marvin Minsky [199], one of the biggest challenges to develop telepresence system is to achieve the sense of "being there", which nowadays is more widely known as *the sense of presence* [264]. Slater et al. [279] described this as *place illusion*, having a sensation of being in a real place. Appropriate sensation of presence during telepresence tasks could help users reducing the error rate during teleoperation and improving the efficiency of cooperation during the interaction with remote people and environments [102].

However, current telepresence systems can not provide users with sufficient immersive experience to support efficient interaction or natural exploration of REs. In particular, the usage of single webcams for capturing remote people and environments, as well as limited screens for displaying remote scenes in current telepresence systems usually restrict user's illusion of being in a RE, and constrain their interactive behaviour inside a smaller bounded space. Furthermore, in order to provide the webcam with the ability to move freely, the webcam is often attached on top of a mobile robotic platform to travel within the RE, this combination is usually referred to as *telepresence robot* [308]. Nevertheless, the movement control of the mobile platform in today's telepresence robots are often restricted to some traditional interactive methods such as web interfaces [56], joysticks [300] and keyboards [24], which lack natural and intuitive mapping between user's instinctive moving behavior and motions of the robotic platform. Both of these reasons reduce the quality of user experience during telepresence [72].

In recent decades, the development and advancement of intelligent reality technologies like *virtual reality* (VR), *augmented reality* (AR), *mixed-reality* (MR) and *omnidirectional videos* provide potential possibilities to solve these inherent limitations. Compared with a classical screen display, virtual reality (VR) head-mounted displays (HMDs) could provide users with a wider field of view (FOV) which is closer to the human's visual field in natural state (slightly over 210° horizontally and around 150° vertically [159, 128]), and segregate the user's vision from the real environment to create an immersive experience in virtual scenario. These advantages make immersive HMDs becoming an ideal option for displaying a RE immersively to local users. In order to match a wider FOV on immersive HMDs, 360° cameras can be used for capturing live-stream from REs, which could be rendered in real-time and displayed on immersive HMDs.

1.3 Application Scenario

In this dissertation, solutions to the inherent limitations of current telepresence robots are proposed and evaluated. In order to describe the research questions in this dissertation, an application scenario of telepresence is presented first, before the specific questions are proposed and explained.

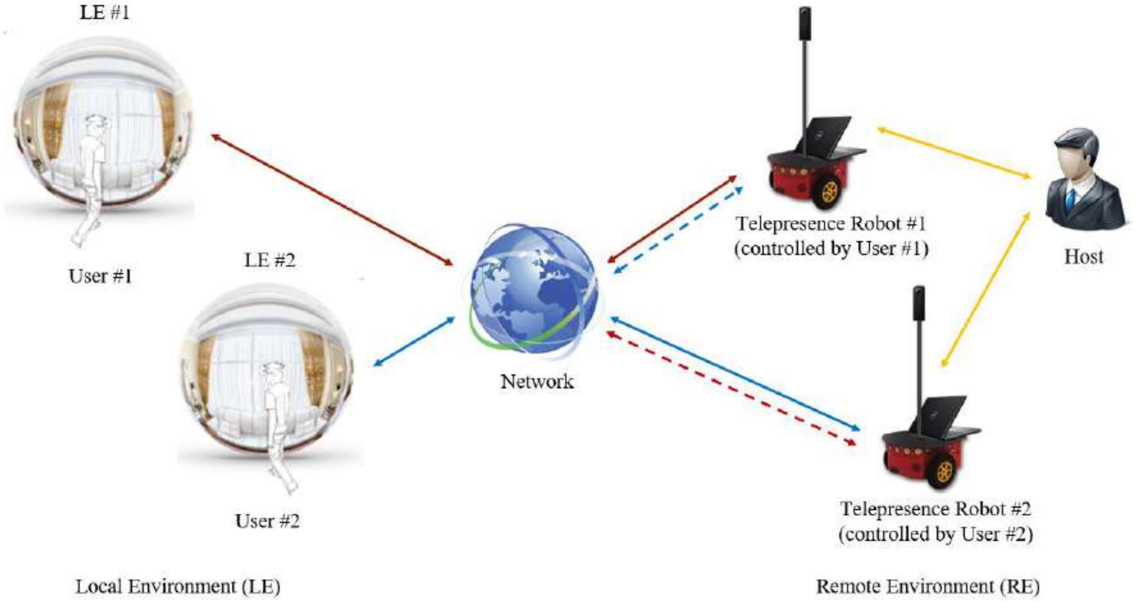


Fig. 1.1 Application scenario of telepresent interaction with 360° VR-based telepresence system.

The illustration shown in Figure 1.1 demonstrates a telepresence application in which two *users* (which will be occasionally referred to as *teleoperators* or *guests* as well throughout the dissertation depending on the context) are using two 360° video-based telepresence robots (which will be also occasionally called as *robotic platform* or *robotic surrogate* in this dissertation) to visit and review several remote apartments from their own sites. The apartments are presented by a real-estate agent, called *host*. In this case, the environments where the users locate are referred to as *local environment* (LE), while the apartment that the users want to visit remotely is defined as *remote environment* (RE). Throughout this dissertation, the LE and the RE are both decided from the perspective of the users. The LE and the RE are separated from each other in geographically distance, the connection between them is based on a communication network only.

In the RE, two 360° video-based *telepresence robots* work as the surrogates of the users. The host interacts with the robotic surrogates and presents the apartments to the users remotely. Normally, each user steers his or her own surrogate, while in some situations, it is also possible for users to switch from their own surrogates to others to share a common view with other users. The omnidirectional live stream of the RE is captured by the 360° camera equipped on top of the telepresence robot and transmitted to the LE via the described communication network. The live stream captured from the RE will be processed on a workstation in the LE, and then rendered and displayed on the user's immersive HMD as spherical video texture, the virtual space constituted by this is defined as *360° video-based RE* or *360° spherical space* in this dissertation. And the relevant method, which uses 360° cameras and immersive HMDs for capturing and displaying the RE, is referred to as *360° VR* technology in the scope of this dissertation.

Meanwhile, the user's motion (translation, rotation and other behaviors such as gazing, nodding or head shaking) is also captured by a tracking system and transmitted to the RE to manipulate or control the motions of the telepresence robot. In this way, the user could steer and move the telepresence robot through the RE by natural walking inside the tracking area of the LE, which has the potential to be more intuitive and natural compared to other traditional interaction methods.

1.4 Research Questions

The preliminary design of such a 360° VR-based telepresence system as well as its potential application scenarios inspire the following scientific research questions of this dissertation:

HMDs provide immersive VR experiences by supporting motion parallax, a wide field of view and stereoscopic display. Although, HMDs are often used with computer-generated VR content, the technology also allows the possibility of viewing 360° immersive videos, which are typically captured with 360° panoramic cameras. Thereby, it is possible for humans to remotely visit a real-world scenario in an immersive way without physical travel. However, in such 360° video-based VEs, stereoscopic perception is limited, and therefore, depth cues for spatial presence might be hindered [28].

In MR environments, these immersive videos can be augmented by virtual objects such as buildings, cars, or avatars [198] to blend real content (from the immersive video) with computer-generated virtual objects. The combination of real and virtual objects in 360° video-based VEs, raises the challenge of consistent global illumination [218, 235]. Previous algorithms [254] allow for seamless composition of 3D virtual objects into a 360° video-based VE by using the input panoramic video as the lighting source to illuminate the virtual objects, and form a new scenario which is referred to as *360° MR environments*. However, the effect of global illumination and shadows of virtual objects on the object localization and user experience in 360° MR environments still remain poorly understood. And this motivates the following research question:

- *Q1: To what extent will the illumination and shadow effect of virtual objects influence object localization tasks and user experience in 360° MR environments?*

According to the design, the user will steer the 360° video-based telepresence robot to explore the RE by natural walking inside the tracking area of the LE. However, usually the size of the RE is significantly larger than the size of the tracking area of VR users in the LE. Moreover, the layouts between RE and LE are also usually different. Thus, in this situation, a one-to-one mapping from the local movement of a user to the remote movement of a telepresence robot is not feasible. In other words, in order to visit a RE which is significantly larger than the tracking area in the LE, the locomotions of users inside the tracking area have to be manipulated.

In this context, *redirected walking* (RDW) is a suitable approach to solve this problem. It allows humans to perform near-natural walking in an infinitely large virtual environment (VE) by manipulating the virtual camera with different gains [249]. Many previous researches on redirected walking in VEs [288, 287, 292, 39, 119, 169, 217] have been conducted so far, however, the application of RDW in 360° video-based RE has rarely been studied [260]. Specifically, in order to create a natural user experience during redirected walking in 360° video-based RE, it is essential to quantify human's sensitivity to the walking redirection in such kind of virtual space. This inspires the following research question:

- *Q2: How much can humans be unknowingly redirected in a 360° video-based RE?*

For the host, who presents the apartment to the user and interacts with the telepresence robot, understanding the current state and predicting the next move of telepresence robot (which could be also regarded as the intention of user) accurately is significantly important for establishing an efficient interaction with the user in LE via the surrogate of telepresence robot [46]. However, the visual cues that the host could obtain from the telepresence robot is very insufficient and hard to understand [124], which increases cognitive efforts of the host and limits active interaction and, therefore, acceptance of the technology.

In addition, compared with traditional telepresence robots, behaviors of 360° video-based telepresence robot will be even more challenging to perceive, comprehend and predict.

Usually, traditional telepresence robots have to perform and complete some motions (like rotation and translation) on mobile base so as to let the camera access the target that the user is interested in. However, a 360° video-based telepresence robot could simplify such motions due to the usage of 360° camera, which reduces easy perception of user's direction for the host. For example, when using traditional telepresence robot with a limited webcam, if the user wants to look around in RE, the mobile base of telepresence robot has to rotate as well following the user's focusing direction in order to capture corresponding views of RE. In this process, the motions of telepresence robots could provide the host with some visual cues to perceive the intention and focus of the user.

Furthermore, if the host can localize the user's gaze direction rapidly and accurately, he or she could proactively offer additional information or description to make the communication more interactive and efficient. However, for a 360° video-based telepresence robot, the mobile base does not need to rotate at all in this situation to follow the user's focus, because the live-stream from the RE captured by 360° cameras is already rendered onto a spherical space and displayed on immersive HMDs. In this case, no rotational visual cues from the motions of 360° video-based telepresence robot are provided to the host to judge where the user is focusing or interested, which may lead to some confusing and misleading during interaction. Thus, localizing the user's gaze direction in RE effectively is extremely important for guaranteeing an efficient and fluent interaction during telepresence [228].

Moreover, most of todays telepresence robots can only present the user's face on a 2D tablet screen, which loses additional spatial deictic cues about the user. In addition, many social cues of the user such as body language, gesture, posture as well as emotions are also missing during telepresence interaction compared to a face-to-face communication. All of these issues mentioned above make it challenging for the host to efficiently and correctly understand the user's intention and state. Thus, corresponding technical methods have to be proposed for 360° video-based telepresence robots to represent the user in the RE appropriately, in order to provide the host with adequate channels to perceive and understand the user's intention and behavior.

To address these limitations, one potential solution would be to display a stereoscopically rendered AR avatar on top of the 360° video-based telepresence robot to indicate the user's body and gaze direction, since humans have learned to quickly interpret body postures of other humans. The pose of the avatar could be synchronized with the pose of the user's HMD and is, therefore, always "looking" at the correct point. The application of such AR avatars will generate following research questions:

- *Q3: Compared with a traditional 2D tablet display, will humans be better at perceiving and localizing the user's gaze direction with an AR avatar?*

- *Q4: Which factors will influence the perception and localization on the user's gaze direction during a 360° video-based telepresence?*

When the user steers a 360° video-based telepresence robot to look through the RE, the host will also need to update his or her position appropriately to adjust to the user's new focusing or interest. In this case, the interaction between the 360° video-based telepresence robot and the host significantly changes from a static process to a dynamic interplay in which both sides can update their states. When interacting with a robotic surrogate in such a dynamic situation, comprehending the current state of telepresence robot can help the host to predict its further actions. In particular, correctly perceiving the moving speed of a robotic surrogate and selecting suitable proxemic preferences based on it in a dynamic situation like guidance or head-on encounters could reduce the host's discomfort to the invasion of private space caused by robotic surrogate. However, with current setups, the states of the 360° video-based telepresence robot is very challenging for human users to perceive and understand.

In order to solve this problem, one possible solution is to visualize an avatar's upper body with arm swing animation to reflect the current speed of the telepresence robot it is attached to. This is because human arm swing is usually regarded as a passive movement of human's gait [54] and has specific frequencies in different walking speeds [168]. With this method, the visualized avatar could augment the telepresence robot in the real world. In addition, human users could interact with either the computer-generated avatar or the real environment, or a mixture of both. For this reason, in the scope of this dissertation, such visualized avatar is specifically called *mixed reality* (MR) avatar. The application of MR avatar and corresponding arm swing animation for displaying the current states of telepresence robot raises following research questions:

- *Q5: Is the avatar's upper body visualization with arm swing animation effective for humans to perceive different moving speeds of a robotic surrogate?*
- *Q6: With an upper body visualization on top of the robotic surrogate, which factors will influence the proxemic preferences of humans during a dynamic interaction?*

This dissertation will concentrate on the research questions listed above and perform corresponding user studies in the following chapters to explore the answers.

1.5 Overview

The remainder of this dissertation is structured as follows: Chapter 2 introduces fundamental knowledge and related research about telepresence, 360° VR, human-robot interaction (HRI) and human perception. Chapter 3 presents a full description of system design, development and technical details of 360° VR-based telepresence system, which establishes the basis of the rest research studies in this dissertation. Furthermore, we evaluates the effect of global illumination and shadows of virtual objects on object localization tasks and user experience in 360° MR environments. Chapter 4 explores the perceptive ability of human users on the manipulations of translation and rotation in a 360° video-based RE. The results provide detection thresholds of RDW, which means, when the corresponding manipulated gains are within these thresholds, a natural and smooth interaction between the user and RE could be guaranteed. Chapter 5 investigates the human's perception on the social cues expressed by a MR avatar attached on top of a robotic surrogate and its corresponding animations. Two independent user studies are included in this chapter. The first study concentrates on the perception and localization on the user's gaze direction in RE, which is delivered via a visualized avatar attached on top of the 360° video-based telepresence robot. The second study focuses on a MR avatar arm swing technique, which subtly communicates the velocity of a robotic surrogate it is attached to. Furthermore, the proxemic preferences between the robotic surrogate and the human subject in the dynamic scenarios of walking following or towards a robotic surrogate are also evaluated. Finally, Chapter 6 summarizes the main contributions of this dissertation, putting the conclusions and remained open questions into a broader context for future research.

1.6 Publications

The main contributions of this dissertation have been published in peer-reviewed international journals and conferences, which are listed below:

Main Authorship

The following publications were mainly created by myself while co-authors contributed parts of the system implementation, writing of paper sections, or supervision.

Journal Articles

- ***Jingxin Zhang***, Nikolaos Katzakis, Fariba Mostajeran, Paul Lubos, Frank Steinicke. "Think Fast: Rapid Localization of Teleoperator Gaze in 360° Hosted Telepresence". International Journal of Humanoid Robotics (IJHR) (2019): 1950038. [doi:10.1142/S0219843619500385]

- ***Jingxin Zhang***, Eike Langbehn, Dennis Krupke, Nikolaos Katzakis, Frank Steinicke. "Detection Thresholds for Rotation and Translation Gains in 360° Video-Based Telepresence Systems". IEEE Transactions on Visualization and Computer Graphics (TVCG) 24.4 (2018): 1671-1680. [doi:10.1109/TVCG.2018.2793679]

Conference Papers

- ***Jingxin Zhang***, Omar Janeh, Nikolaos Katzakis, Dennis Krupke, Frank Steinicke. "Evaluation of Proxemics in Dynamic Interaction with a Mixed Reality Avatar Robot [1]". Proceedings of International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments (ICAT-EGVE). 2019, pages 37-44. [doi:10.2312/egve.20191278]

- ***Jingxin Zhang***, Eike Langbehn, Dennis Krupke, Nikolaos Katzakis, Frank Steinicke. "A 360° Video-based Robot Platform for Telepresent Redirected Walking". Proceedings of ACM Human-Robot Interaction (HRI) Workshop on Virtual, Augmented and Mixed Reality for Human-Robot Interaction (VAM-HRI). 2018, pages 58-62. [https://basilic.informatik.uni-hamburg.de/Publications/2018/ZLKKS18a/]

Others

- ***Jingxin Zhang***, Jannis Volz, Frank Steinicke. "Effects of Global Illumination and Shadows on Object Localization Tasks in 360° Video-based Mixed Reality Environments". (submitted to ICAT-EGVE 2021 Poster Presentation)

- ***Jingxin Zhang***, Nikolaos Katzakis, Fariba Mostajeran, Frank Steinicke. "Localizing Teleoperator Gaze in 360° Hosted Telepresence". Proceedings of IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (Poster Presentation). 2019, pages 1265-1266. [https://basilic.informatik.uni-hamburg.de/Publications/2019/ZKMS19/]

- ***Jingxin Zhang***. "Natural Human-Robot Interaction in Virtual Reality Telepresence Systems". Proceedings of IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (Doctoral Consortium). 2018. [doi:10.1109/VR.2018.8446521]

[1] This publication received the Best Paper Award of ICAT-EGVE 2019 in Tokyo, Japan.

Co-Authorship

The following publications were mainly created by someone else and are not part of this dissertation. However, as a co-author, I contributed critical parts of the system implementation, experiment design, or paper writing.

- Manuela Uhr, Joachim Nitschke, ***Jingxin Zhang***, Frank Steinicke. "Hybrid Decision Support System for Traffic Engineers". Proceedings of IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (Poster Presentation). 2018, pages 713-714. [doi:10.1109/VR.2018.8446141]
- Manuela Uhr, Joachim Nitschke, ***Jingxin Zhang***, Paul Lubos, Frank Steinicke. "Evaluation of Flick Gestures on Multitouch Tabletop Surfaces". Proceedings of ACM International Conference on Interactive Surfaces and Spaces (ISS). 2017, pages 324-329. [doi:10.1145/3132272.3132274]

Chapter 2

Fundamentals

In this chapter, we resume fundamentals and related research on the topics of telepresence, human-robot interaction (HRI), human perception, immersion and presence, and 360° VR.

2.1 Telepresence

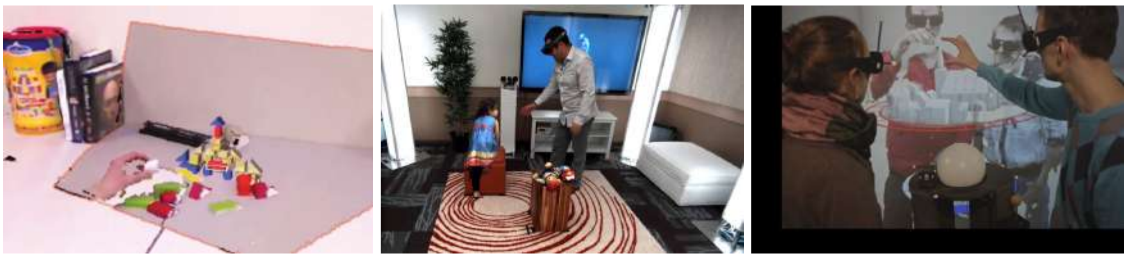
2.1.1 Definition and Applications

Telepresence refers to technologies, which enable people to visit remote environments and interact with people, objects and the surroundings there without physical travelling. This concept was originally proposed in 1980 by Marvin Minsky, who discussed the following application fields [199]:

- Safe and efficient nuclear power generation, waste processing, and land and sea mining.
- Advances in fabrication, assembly, inspection, and maintenance systems.
- The elimination of many chemical and physical health hazards and creation of new medical and surgical techniques.
- The construction and operation of low-cost space stations.

Besides the name "*Telepresence*", such kind of technologies could be also defined as "*Telexistence*", which was originally proposed by Susumu Tachi in 1980, who described it as "highly realistic sensation of existence in remote places without actual travel" [298].

Advanced telepresence technologies have enormous potential for people in complex, potentially dangerous or highly precise situations, since as far as we (as engineers) know, there is no better general purpose system as ourselves (as operators) [113]. With a fast



(a) (b) (c)

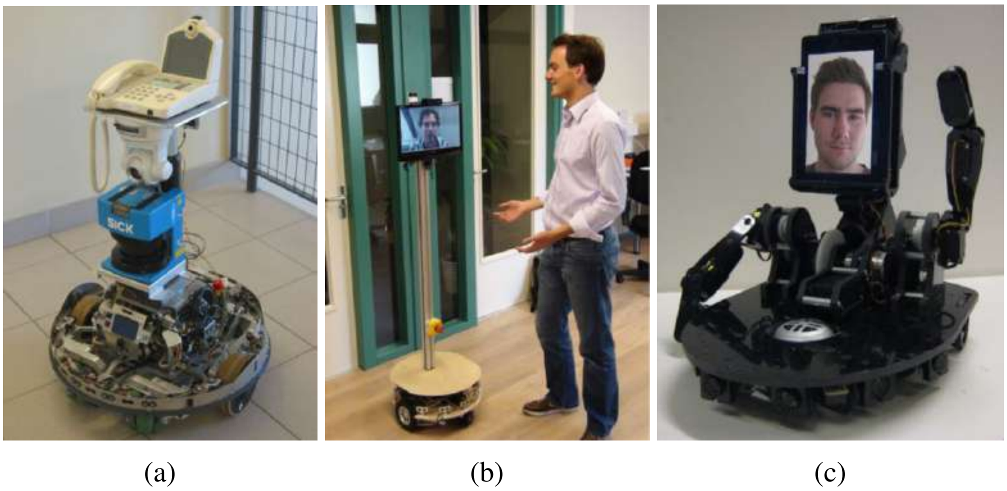
Fig. 2.1 Typical telepresence systems in remote collaboration: (a) BeThere (image from [282]); (b) Holoportation (image from [224]); (c) Immersive Group-to-Group Telepresence (image from [25]).

development in relevant fields in the last 40 years, telepresence nowadays has been widely used in our daily life and contributes in different domains. For example, in the aspect of medical treatment [19], the United States Department of Defense developed the da Vinci telerobotic surgical system [20] to enable combat surgeons to perform surgery for wounded soldiers from a remote and safe location so as to reduce the death rate on the battlefield. Besides of this, robotic telepresence system has been also adopted in the intensive care unit (ICU) of hospitals to promote rapid physician response to unstable patients and to reduce cost in neurointensive care [313]. Another typical application in medical treatment is using a haptic telepresence system for endoscopic heart surgery [186], by which the high visual fatigue of surgeons during operations was effectively decreased. Telepresence technologies also improve the way of traditional education by breaking the constraint of physical distance and establishing a communication between people from different locations via network and visualization devices. The education activities based on telepresence technology is referred to as "*Tele-education*" [51]. Today, tele-education exists in education events ranging from helping children in communicating with foreign language teachers [301] to college degree level [182].

As for the application in remote collaboration, telepresence provides new solutions to the remote cooperation and plays an important role in increasing efficiency and reducing operator workloads [95]. For example, TeleAR [319] provides a computer-mediated remote collaborative system to improve distributed cognition and to understand other remote team members' intentions. Similarly, Face2Face [74] makes it possible for remote users to perform collaborative multi-touch interaction in a remotely face-to-face situation. Furthermore, 3D helping hands [303], 3D-board [335], BeThere [282] (Figure 2.1a), Beaming [285], Streamspace [263], Holoportation [224] (Figure 2.1b), Room2Room [234] and Group-to-Group Telepresence [25] (Figure 2.1c) are all technologies using telepresence for remote communication and collaboration.

2.1.2 Robotic Telepresence

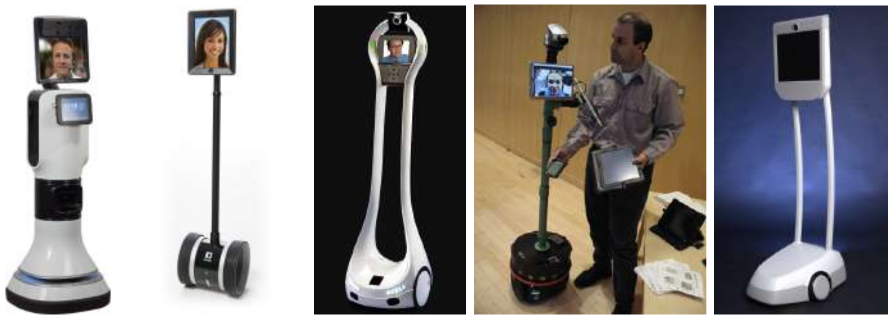
In the earlier years, most telepresence systems lacked mobility in the RE, which as a result provided only a limited view for the user. In order to solve this limitation, robotic telepresence systems were proposed [273], which combine video-based telepresence systems with robotic mobile bases. This combination allows users to explore the whole RE by controlling the robotic mobile base [165]. Michaud et al. [197] developed Telerobot (Figure 2.2a), a telepresence assistive robot consisting of a video-phone, a suspension mobile base and sensors for obstacles avoidance and behavior decision, which is designed and used for home-care assistance of elderly people. Moreover, Alers et al. [8] designed MITRO (Figure 2.2b), a telepresence robot for an augmented telepresence with assisted control. In addition, Adalgeirsson et al. [1] implemented the telepresence robot MeBot (Figure 2.2c) which supports social expressions. A user study verified that a socially expressive robot was more engaging and likable than a static one.



(a) (b) (c)

Fig. 2.2 Typical telepresence robots: (a) Telerobot (image from [197]); (b) MITRO (image from [8]); (c) MeBot (image from [1]).

Another primary target of robotic telepresence systems is to promote social interactions between humans [211, 58, 1]. For instance, Nakanishi et al. [211] performed a study to evaluate the degrees of social telepresence evoked by the motions of camera. The results showed that forward-backward movement of the camera significantly contributed to social telepresence, whereas rotations did not. In other studies, Shiarlis et al. [275] proposed a deep learning approach that learns social behaviour from demonstrations. Based on this approach, two challenging social tasks, i.e. group interaction and following, were tested on a semi-autonomous telepresence robot. The results illustrated a better performance of the proposed method compared with gradient boosting regression method. Coradeschi et al. [57]



(a) (b) (c) (d) (e)

Fig. 2.3 Commercial telepresence robots from different companies: (a) RP-Vita; (b) Double; (c) VGo; (d) PRoP; (e) BeamPro.

introduced the robotic telepresence system GiraffPlus, which combines social interaction and long-term monitoring for promoting independent living.

Meanwhile, rapid development of telepresence robots creates a large amount of potential market space, hence, many companies have developed and released telepresence robots (Figure 2.3) for commercial use, such as RP-Vita [1], Double [2], VGo [3], PRoP [4], BeamPro [5], etc.

2.1.3 Telepresence and VR

Before VR technologies were widely used as a media user interface, the user of telepresence systems in the LE could only rely on screen-based user interfaces to view the video stream of the RE. For the user, one of the major drawbacks of such methods is the lack of the sense of presence in the RE.

VR technologies provide an immersive solution to this restriction by using VR HMDs [125, 87, 320] or cave-like projection spaces [102, 267] to display rendered avatars of remote collaborators as well as the RE surrounding them. With a VR HMD or a cave-like projection space, the physical environment surrounding the user is isolated from the user's visual perception, as a result, only the information of rendered REs is received by the user. With such methods, the user's attention and perception on the LE when interacting through a screen-based user interface can be reduced effectively, which as a result could increase the

[1] https://intouchhealth.com/telehealth-devices/intouch-vita

[2] https://www.doublerobotics.com

[3] http://www.vgocom.com

[4] http://www.prop.org

[5] https://suitabletech.com

user's sense of presence in rendered REs and improve the user experience during remote interaction.

Due to the widespread usage of VR technologies and its huge potential for improving traditional telepresence, researchers start to combine VR and telepresence. With a focus on telepresence, Steuer [293] re-defined VR as "a real or simulated environment in which a perceiver experiences telepresence". In addition, Edwards [75] described telepresence as "virtual reality in the real world". Both of these definitions emphasize the strong connection between telepresence and VR.

2.2 Immersion and Presence

2.2.1 Immersion

When the concept of *telepresence* was proposed by Marvin Minsky in 1980, it has been pointed out that the biggest challenge in developing telepresence systems is to achieve the sense of "being there" [199]. Furthermore, he asked the question:

> *“Will we be able to couple our artificial devices naturally and comfortably to work together with the sensory mechanisms of human organisms?"*

According to the question above, the key to solve this problem is to coordinate artificial devices with human sensory and perceptual system naturally as well as comfortably in order to achieve the sense of "being there".

To reach this goal, researchers make their efforts and perform studies focusing on two aspects. One aspect is to improve artificial devices and relevant technologies to create more realistic and natural stimuli (which may include images, sound, smell, haptics, etc.) for stimulation of human sensory channels [97, 47], which induces the human perception on the virtual environment beyond the sensation to the physical environment. In this context, the concept of *"immersion"* is proposed to describe the ability of a VE as well as relevant systems or devices that tricks us in the illusion of being physically present in a non-physical world created with various computer-generated sensory stimulation. In other words, immersion denotes the technological degree to which a technology allows the user to experience a VE. Hence, it could be generalized that the sense of "being there" in telepresence is technically produced by the RE rendered with relevant technologies and perceived by the user during interaction.

2.2.2 Presence

However, even within a same immersive virtual environment, the user's perceived illusion of "being there" can vary significantly from each other. In other words, besides the immersion created by artificial devices from technical aspect, the capability of human users on its perception influences the sensation of "being there". Therefore, another research direction that scientists and researchers concentrate on is to understand the human sensory mechanisms as well as current perceptual limitations, so as to optimize the user's perceptive ability during telepresence and interaction. Such human's subjective sensation of being in a virtual environment depicted by a medium is referred to as *"presence"* [86].

Steuer et al. [293] pointed out that the presence was highly similar with the phenomenon of distal attribution or externalization, which refers to the referencing of human perception to an external space beyond the limits of sensory organs themselves. Many perceptual factors could have effects on the perception of presence, ranging from the input of sensory channels to attentional, perceptual and other mental process. For this reason, presence is a result which assimilates input sensory data with current attentions and past experiences. Similarly, Jerald et al. [139] provided an analogous conclusion that presence is a subjective perception, which depends on external immersion and user's current psychological state. The subjective feeling of presence in an immersive virtual environment can be measured by using questionnaires such as the Igroup Presence Questionnaire (IPQ) [268] or the Slater-Usoh-Steed (SUS) Presence Questionnaire [280]. Besides of questionnaires, presence can be also measured by employing physiological indices, behavioral feedbacks, and interviews [96].

From the definitions and description above, we summarize that immersion is a more objective term, which describes the amount and quality of technical sensory stimulation provided in a VE, whereas presence is a more subjective term that indicates how much a person perceives the illusion of being inside a VE. In other words, immersion is a technology-related, objective characteristic of a VE, while presence is a subjective perception of immersion in the aspects of psychology, perception and cognition.

2.2.3 Related Research on Immersion and Presence

As an important part of VR research, the effects of presence and immersion in VEs have been explored in many previous studies.

Banos et al. [21] designed and performed a study, which focused on the role of immersion and emotional content of VEs in the sense of presence. The results indicated that both immersion and emotional content of VEs have a significant impact on the presence. However, immersion was more relevant for non-emotional environments than for emotional ones.

Another study conducted by Riva et al. [258] confirmed the efficacy of VR as an affective medium. The research verified that an interaction with "anxious" and "relaxing" VEs could also produce a corresponding sense of anxiety and relaxation. Furthermore, the data also showed a circular interaction between presence and emotions: on one side, the sense of presence was stronger in "emotional" VEs, on the other side, the level of presence conversely influenced the user's emotional state as well. In addition, Schuemie et al. [270] summarizes a large body of literature and indicates that the consequences of presence in VEs could lead to effects on the subjective sensation, task performance, responses and emotions as well as simulator sickness.

McMahan et al. [189] provided an extra explanation on the immersion in VEs. An experience of being in an exquisitely generated VE is enjoyable in itself, which has little relevance with the fantasy content. This kind of experience is defined as the immersion in VEs. Immersion is a metaphorical term which comes from the physical experience of being submerged in water. When being in an immersive VE, the same feeling from psychological level is also expected, which can take over all of our attention and all the perceptual channels. In an immersive VE, immersion makes it possible for the users to explore new spaces, perform tasks, interact with medium and have entertainment. This is because in any medium, the human brain mechanisms could isolate the physical world around us when having experience in a VE with intensive stories and strong perceptual stimulus.

Meanwhile, human's subjective factors could influence the sensation of presence in VEs. For example, Kober et al. [161] conducted a user study in order to investigate the relationship between personality variables and the sense of presence. The results indicate that the presence correlates with people's personality factors such as absorption, mental imagination, perspective taking and immersive tendencies. Similar results were replicated by Alsina-Jurnet et al. [9]. The results show that a larger sense of presence in test anxiety environments was measured than in a neutral environment. In particular, high test anxiety students felt more presence than non-test anxiety students.

Usually, it is very challenging to design a VR experience which evokes the sense of presence and simultaneously inhibits simulator sickness. This is because the relationship between the user's susceptibility to VR sickness and a sense of presence, determined by velocity and visual angle of the visual information, involves a trade-off between the two [302]. In order to solve this problem, Tanaka et al. [302] proposed an optimal value search method that calculated the velocity and visual angle efficiently, which control VR sickness and do not impair presence by taking into account of a subject's characteristic. Hence, a trade-off between VR sickness and the sense of presence could be determined effectively during the design of VEs.

In some occasions, people may even experience a higher sense of presence when performing identical tasks in virtual and real environment. Villani et al. [314] confirmed this phenomenon from a scenario of simulated job interviews, in which the experienced presence was higher during a virtual interview than in a real-world simulation. However, this result was only supported by subjective anxiety scores.

Furthermore, Bowman et al. [36] proposed a Human-VE interaction loop and summarized that both display hard- and software can play a major role in determining the level of immersion. Components that could influence visual immersion include field of view (FOV), field of regard (FOR), display size, display resolution, stereoscopy, head-based rendering, realism of lighting, frame rate and refresh rate. By using Grounded Theory, Brown et al. [38] constructed a robust division of immersion into three levels: (i) engagement, (ii) engrossment and (iii) total immersion, which suggested new lines for investigating immersion and applying it in software domains. Moreover, Shin et al. [276] performed a study in order to investigate how story experiences in immersive storytelling was influenced by immersion. The results illustrate that the way users view and accept VR stories derives from the way they imagine and plan to experience them. This finding also suggests that the user's cognitive process will determine how they empathize with and embody VR stories. In addition, Schuchardt et al. [269] conducted an experiment to explore the benefits of immersion for spatial understanding in VR. The results demonstrate that for certain tasks, artificial systems with higher immersion could significantly improve the accuracy, speed, and comprehension in VEs.

2.3 360° VR

2.3.1 Panorama Images

Panorama images refer to wide-angle views or representations of some physical space [322]. Those images come in the forms of painting, photography, video or film. The word "panorama" was originally proposed in the 18th century by the English painter - Robert Barker to describe his panoramic paintings of Edinburgh and London (Figure 2.4). Moreover, Barker also created a novel presentation mode in which his panoramic painting was shown on a cylindrical surface and viewed from the inside. By this way, a vantage point was provided to the viewers to watch the entire circle of the city horizon, rendering the original scene with high fidelity.

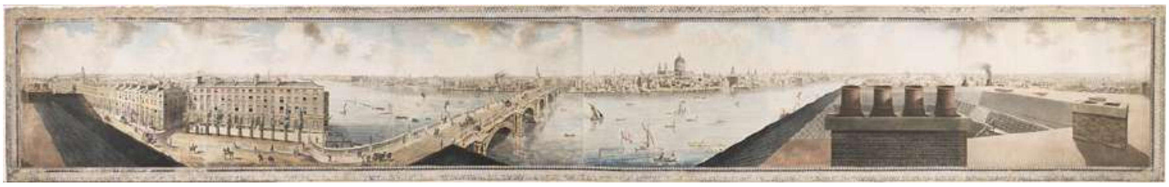


Fig. 2.4 A panorama of London by Robert Barker, 1792 [322].

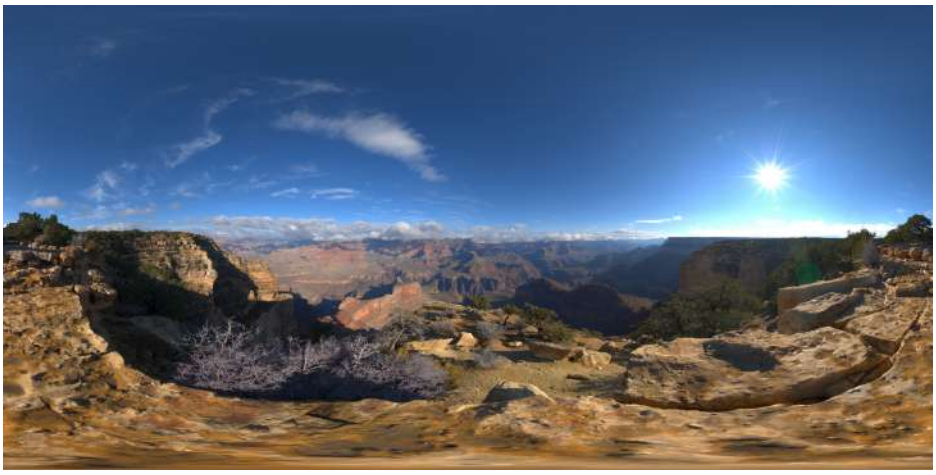


Fig. 2.5 A panorama of canyon [34].

Nowadays, rapid development of 360° (omnidirectional) camera makes the process of getting a panoramic photography or video much easier. Figure 2.5 shows an example of equirectangular panorama produced by a 360° camera with an image ratio of 2:1.

In addition, the display form of panoramas improved from traditional physical display media such as cylindrical surfaces, curved screens and IMAX cinemas to novel visualization technologies such as CAVE-like projection and 360° VR. In order to establish a connection between a 2D panorama (panoramic image or video) and corresponding compressed 3D information, panoramic projections are proposed for mapping a full or partial 3D scene onto a 2D surface [223]. Traditional cylindrical projections convey the scene visible in all directions except for the areas right above the viewers heads and under their feet. This will lead to the problem that the cylinder display space could not be "wrapped" totally, because the top and bottom area of such cylindrical mapping is missing.

In contrast, spherical projection could solve this limitation by containing light data originating from all directions. In another words, spherical projection comprises a 180° viewing range in the vertical and a 360° viewing range in the horizontal direction. Thus, a panorama with spherical projection could be regarded as a 2D surface that comprises all the points of a sphere. For this reason, the panoramic medium in a spherical projection could be rendered onto a spherical surface to rebuild the scene. Figure 2.6 demonstrates an example of a spherical scene in Unity3D with the image of Figure 2.5 by using a visible-from-inside shader. Moreover, Unity3D also provides a technical solution for rebuilding a panorama by setting it as the Skybox of a virtual scene, which is presented in Figure 2.7.

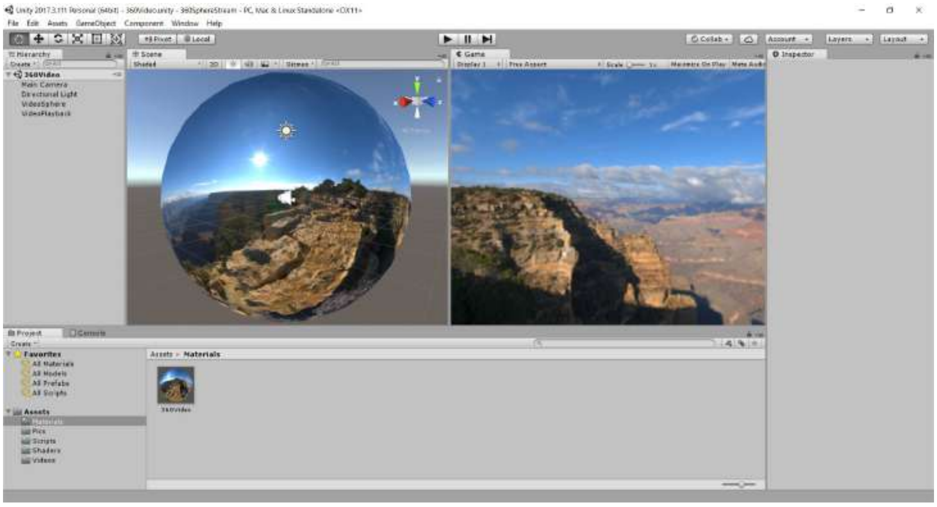


Fig. 2.6 Canyon panorama rebuilt in a spherical space in Unity3D.

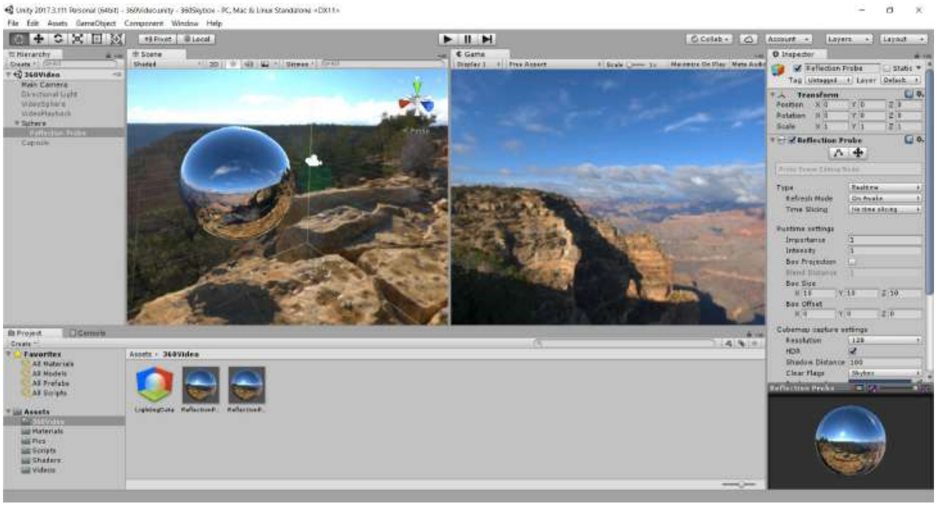


Fig. 2.7 Canyon panorama rebuilt as the Skybox of a virtual scene in Unity3D.

Since the spherical projection can compress a 3D scene into a 2D panorama from all directions, this projection format became very popular in social media, immersive panoramic movies, 3D graphics programs, and computer video games [253, 254].

When the rebuilt scene is set up with a virtual camera from inside and rendered on the HMD, the user with VR headset gets the illusion of being inside the environment presented on the panorama, and such setups are referred to as 360° VR [127]. The principle of 360° VR working inside a sphere-rendered virtual space is described in Figure 2.8.

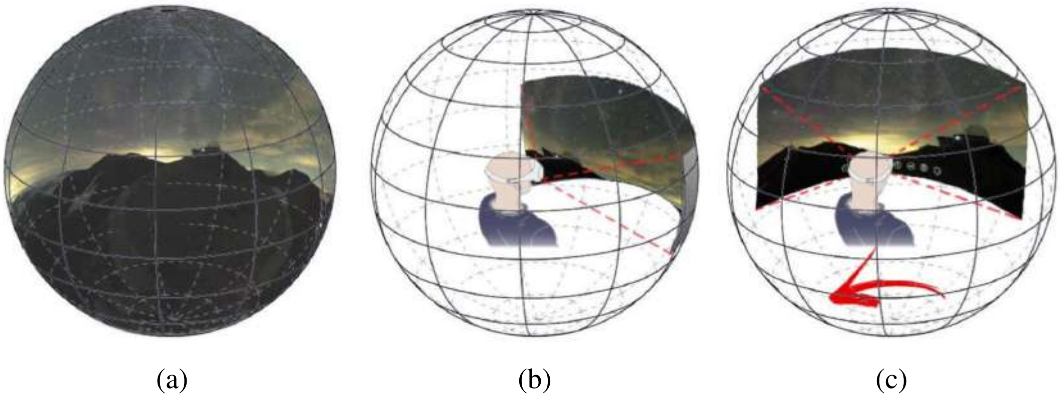


Fig. 2.8 The principle of *360° VR* (image from [238]): (a) A fully-rendered 360° video is projected on the surface of a spherical space. (b) The user's *point of view* (POV) in the VE is set in the center of the spherical space and the visible scene is determined by the user's *field of view* (FOV) and *line of sight* (LOS). (c) When the LOS changes, the user's visible area in the spherical space responds by following and turning in the same direction as the video is playing.

With the development of motion-sensing technology, nowadays 360° video as well as its rendering and display could respond in real-time based on the user's head motions (for instance, measured by the sensors on the HMDs), which makes it possible for the user to perceive place illusion in a dynamic scenario.

2.3.2 Related Research on 360° VR

Usually, panoramic videos for 360° VR require a lot of resources such as bandwidth and GPU to support a real-time interaction and provide high visual quality. However, these requirements can cause negative influences on the quality of user experience due to high latency and render artefacts in cases when resources could not be provided. In order to solve this issue, Petrangeli et al. [236, 126] proposed a method in which panoramic videos for 360° VR are divided into spatial tiles. During the interaction, only the tiles which are included inside the user's field of view are rendered with the best quality, whereas the rest parts of 360° videos are streamed at lower quality. In this way, essential resources for processing and rendering 360° videos can be significantly reduced without affecting the user experience. Furthermore, Petrangeli et al. also introduced an algorithm to predict the user's field of view, which can minimize the transition of visual quality during viewport changes.

Similarly, for the prediction of visible region in a 360° video, Fan et al. [80] developed fixation prediction networks, which predict the user's field of view watching 360° videos by using sensor- and content-related features based on HMDs. Compared with other existing methods, this solution consumes lower bandwidth and has shorter initial buffering time and running time. The prediction of user's field of view in 360° videos is not only important for improving the quality of rendering and display, but also for the delivery of 360° streaming data. For example, Qian et al. [246] considered the problem of optimizing the delivery of 360° videos in cellular network situation. In their proposed streaming scheme, only the user's visible segments in 360° videos are delivered according to the results of head movement prediction. By using the viewing behaviour collected from real users, they conducted a trace-driven simulation study and validated that their streaming scheme could compress bandwidth consumption by up to 80%. Afzal et al. [5] investigated the characteristics of 360° videos by examining thousands of YouTube videos from various categories. Based on their results, they summarized that the actual bit rate of 360° videos could be significantly reduced when only transmitting the region of the user's field of view. Compared with regular videos, 360° videos are less variable in terms of bit rate and contain less motion. This finding provides new expectation and requirements for 360° end-to-end streaming architectures.

Among the current 360° VR applications, monoscopic 360° videos are the most prevalent type of content. However, such type of 360° videos lack 3D stereoscopic cues. In order to solve this limitation, Huang et al. [130] proposed a novel warping algorithm, which can perform VR playback of input monoscopic 360° videos in stereoscopic. Based on this algorithm, novel views for each eye are synthesized for the user. Moreover, Ardouin et al. [12] presented a novel approach for stereoscopic rendering of VEs with a wide field of view up to 360°. Their method uses a new pre-clip stage specifically adapted to geometric approaches and can be integrated seamlessly with immersive VR systems because of its compatibility with stereoscopy, head-tracking, and multi-surface projections. Furthermore, Ardouin et al. [11] also investigated the effect of different visualization techniques on the navigation in 360° VEs. In a user study, a perspective projection was selected as baseline to evaluate different methods. The results demonstrate that omnidirectional rendering methods allows for more efficient navigation in terms of average task completion time, which suggested that omnidirectional rendering could be used in VR applications in which fast navigation and rapid visual exploration are required.

Nowadays, 360° VR is widely used in various applications ranging from cognitive training to immersive entertainment. For example, Grewe et al. [98] developed a novel 360° VR supermarket, which was displayed on a circular arrangement of 8 touch-screens. With this setup, they performed a cognitive user training in which users need to finish some specific shopping activities according to the provided list in some situations. The results illustrated that the task performance correlated significantly with participants' figural-spatial memory abilities as well as their sense of presence. In addition, 360° VR is widely used for educational VR content, which works as an alternative approach to fully computer-generated VR. Kavanagh et al. [148] discussed educational content creation for 360° VR setups with a focus on the point of view, video quality, video editing and directing attention. Another application of 360° VR in the field of immersive education was developed by Liao et al. [178]. They created a spherical projection system for scientific events. The system used OpenCL to combine multiple webcam images into a panoramic texture and mapped it on the surface of a spherical model. For immersive entertainment, McGinity et al. [188] created a versatile VR theatre (AVIE) by combining real-time 360° omni-stereocopic projection with surrounding audio and markerless motion tracking, which could provide a highly immersive and interactive VR theatre experience to maximum of 20 users.

2.3.3 360° VR in Telepresence

The application of 360° VR in the field of telepresence started in the last decade. Previous research such as Johnson et al.'s study [141] has verified that in remote collaboration and telepresence, wider field of views could improve task efficiency and lead to fewer collisions between collaborations. Due to the advantages of 360° VR regarding field of view, immersion and task performance compared with other relevant technologies, integrating 360° VR into telepresence and remote collaboration provides enormous potential. Hence, this technical route has attracted and motivated more and more attention and exploration.

A simple case of using 360° VR in telepresence was introduced by Singhal et al. [278]. They used Google Cardboard and attached 360° lens with a smartphone to build a 360° telepresence setup called BeWithMe. BeWithMe was designed to present the entire 360° view of a partner's location with an independent perspective to control the view. In another study, Heshmat et al. [115] explored the design space by investigating the benefits and challenges of using a telepresence robot to support leisure activities outdoors. The results of their user study showed that the 360° view can promote the sense of presence during remote collaboration. However, challenges related to a lack of environmental awareness, safety issues and privacy concerns due to bystander interactions were also identified in their user study. Thus, balancing safety and privacy issues is a major concern when developing 360° telepresence systems. Other examples that adopt 360° VR in telepresence are the human-to-human telepresence system Jackin head [145], panorama-based telepresence wheelchairs [215] and 360° VR-based teleoperated robot for virtual tours [221].

Moreover, to support stereoscopic display, many telepresence systems nowadays started to integrate binocular vision technique into 360° telepresence systems. Typical telepresence systems using such solution are TwinCam vision system [131] and MAVI robotic platform [17]. In addition, Aykut et al. [15, 16] also proposed a solution to reduce latency between final display and measured egomotion of the users based on generic delay compensation and deep-learning-based head motion prediction. With this proposed method, the mean compensation rate is significantly improved from 72.8% to 97.3% for investigated latencies (0.1s - 1.0s).

All of these previous researches provided valuable references for the system design and development work in this dissertation.

2.4 Human-Robot Interaction & Robotic Surrogate

2.4.1 Human-Robot Interaction

Human-robot interaction (HRI) describes the interactive activities between humans and robots. The original concept of HRI was proposed even before the appearance of robots [321]. In 1941, the famous science fiction author *Isaac Asimov* narrated the *Three Laws of Robotics* in his novel "*I, Robot*" [14] as:

0. A robot may not injure a human being or, through inaction, allow a human being to come to harm.

1. A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law.

2. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

These laws illustrated that safety is of primary concern [310]. With the development of relevant technologies, robot could gradually conduct more autonomous behaviors in unknown environments by detecting, perceiving and making decisions based on various sensors and corresponding algorithms, which makes it possible for people to have a closer interaction with robot under safe conditions. When safety issues are guaranteed, the interaction and collaboration become the main goals of HRI.

HRI is a rapidly evolving field, and has been applied in most robotic tasks nowadays, ranging from manufacturing [50], space exploration [233, 164], undersea research [45], agriculture [105], education [204] to medical surgery [129], rehabilitation [272, 81], package delivery [207] and military [152]. In addition to this, HRI also shows huge potential in some novel areas, such as autonomous vehicle, humanoid robots control in hazardous environments and human-robot social interaction [274]. Specifically, for robotic telepresence, HRI is always one of the most important issues that directly influence the user experience.

In the initial stage, the communication and interaction between teleoperator and robotic surrogate located in another working space usually depends on traditional hardware input devices such as keyboard, mouse or joystick. Such hardware-based interactive methods conform to people's daily operating habits, but lack more natural and intuitive means for the teleoperator to perceive the robot's state and explore the remote environment. For this reason, in recent years, novel HRI technologies and user interfaces have been proposed continuously in order to improve the naturalness and effectiveness of the HRI process.

Some researchers tried to create mind-control user interfaces for disabled people based on brain-computer interface [79, 174], by which the hands of operators could be released from traditional hardware-based interactive methods so as to reduce the physical and mental burden. In addition, head and gaze control of telepresence robots [109] as well as gesture-based telepresence [304] have been also explored by scientists to verify the availability and efficiency in practical applications. Furthermore, novel intelligent interactive devices like smartphones [6] or VR HMDs [332] are also adopted in the field of telepresence.

Besides the research on interactive methods, research concentrates on other issues of HRI to establish a deep connection between human and robot during interactive activities. For example, Mumm et al. [206] investigated the physical and psychological distance during HRI process and concluded that the proxemics between a human and a robot mainly depends on whether the human subjects like the robot or not. This conclusion indicated that when humans dislike robots, they prefer maintaining a larger physical distance from the robot. For the psychological distance, humans who dislike the robot also tend to disclose less to the robot. Hancock et al. [108] evaluated the effects of human, robot and environmental factors on the perceived trust in HRI. The results illustrated that factors related to the robot itself, especially the performance of the robot, had the largest association with trust, and environmental factors were moderately associated. However, there was no strong evidence for the effects of human-related factors.

Rich et al. [255] developed an initial computational model for recognizing engagement between human and humanoid robot according to the engagement process between humans. The model could recognize four types of connection events including directed gaze, mutual facial gaze, conversational adjacency pairs and back channels. Admoni et al. [4] summarized social eye gaze behavior for HRI and proposed three categories of gaze research in HRI: (i) a human-centered approach focusing on people's responses to gaze; (ii) a design-centered approach addressing the features of robot gaze behavior and appearance; and (iii) a technology-centered approach focusing on the computational tools for implementing social eye gaze in robots. Ikemoto et al. [132] proposed and implemented a computationally efficient machine learning algorithm, which concentrated on close physical interaction between robots and human being. Two human-in-the-loop learning scenarios inspired by human parenting behavior, which is an assisted standing-up task and an assisted walking task separately, were tested in order to evaluate the efficiency and performance of the proposed algorithm in HRI assignments. The results demonstrated that the proposed algorithm could improve the quality of interaction between a robot and a human caregiver.

Furthermore, Guadarrama et al. [104] proposed a system for HRI, which could parse the meaning of an input sentence into objects (nouns), spatial relationships (prepositions) and executing commands (actions) based on the visual and spatial information from the robot's sensors, and then send an appropriate command to a PR2 robot or respond to spatial queries. Thereby, robots are able to understand complex commands that refer to multiple objects and relations like: "Move the cup close to the robot to the area in front of the plate and behind the tea box". Wilcox et al. [323] developed a robotic scheduling and control capability, which could adapt to the changing preferences of a human co-worker or supervisor while providing strong guarantees for synchronization and timing of activities. Such robotic scheduling and control solution could promote manufacturing and assembly process, especially for industrial environments in which tight integration and variability required necessary cooperation and interaction between robots and human workers.

2.4.2 Robotic Surrogate

Surrogate refers to various forms of substitutes, which could replace a human to exist in a geographically different location during telepresence and interact with people and environment there. In some science fiction movies, the concept of surrogate was described as different entities based on the imagination of human beings. For example, in the movie *The Matrix*, the surrogate is related to a computer-generated avatar (Figure 2.9a) in a graphics-based illusion [59]. In the movie *Surrogates*, this concept is implemented with a real copy of human body (Figure 2.9b), which consists of titanium and fluid to impersonate the human's flesh and blood. And the surrogate could be controlled with people's mind to complete remote missions. Similarly, in the movie *Avatar*, a disabled soldier can teleoperate an alien body (Figure 2.9c) with his mind to take part in the military tasks in another planet.

When such illusions come to the real world from science fiction movies, the surrogate nowadays appear in our daily environment with the form of various intelligent robots, which is referred to as *robotic surrogate* specifically. A well-known example of robotic surrogate is the Mars Exploration Rover, which has worked on Mars as an alternate of human being to explore this neighbor planet of the earth [309]. Around our life, applications of robotic surrogate also become more and more common and popular. For example, by using robotic surrogate, homebound students, who have serious health issues that prevent them from

(a) (b) (c)

Fig. 2.9 Surrogates in movies: (a) The Matrix (image from [251]); (b) Surrogates (image from [73]); (c) Avatar (image from [10]).

attending school physically, are able to join the teaching events remotely. Meanwhile, Newhart et al. [213] emphasized that teachers and parents need adequate training on the technology and the provision of safe spaces both at home and school. Moreover, Lei et al. [175] explored prevailing patterns when students used robotic surrogates to engage in learning contexts. The results suggested that nonverbal communication with one's robotic surrogate is a dominant form of interaction and engagement during a synchronous learning process. In addition, the capabilities of robotic surrogate and the perception of students on the robotic surrogate as an extension of themselves could have effects on the engagement during educational contexts.

Mackey et al. [183] investigated the use of robotic surrogates as well as corresponding immersive control systems in the process of telecommunication, which could compensate high-level interactions in a face-to-face conversation compared with traditional telecommunication software like Skype. Nagendran et al. [210] explored the possibility of symmetric 3D telepresence via a collaborative task by using two identical humanoid robots located in UK and USA separately as surrogates. The results demonstrated that the symmetric 3D telepresence based on the double robotic surrogates made it possible for participants to use and understand gestures directly in cases where they have to describe their intention or actions non-verbally. Rosenthal et al. [259] tried to use robotic surrogates to navigate paths for visitors to go through buildings and provide useful information in indoor environment. Moreover, they also proposed a symbiotic interaction between robotic surrogates and humans so as to overcome the limitations of robot and let robotic surrogates service for human beings better. Grice et al. [99] developed a web-based augmented reality interface, which enabled people with profound motor deficits to control a PR2 mobile manipulator for some specific self-care and household tasks at home. Tidoni et al. [306] tried to help people with spinal cord injury conducting social interactions with virtual and robotic surrogates via brain-computer interface. The results of their study illustrated that people living with spinal cord injury could use brain-computer interface with good levels of performance (in the aspects of task accuracy, optimizations calls and information transfer rate) and perceived control of surrogates.

Kim et al. [154] explored the effects of physicality and gestures on the human perception and the social influence of surrogate via a large-scale user study in a public space. The results show that people were more likely to be engaged when interacting with a surrogate that was physically presented. However, the movements and gestures did not show expected benefits for the engagement of interaction. In order to use human cognitive capabilities to help guiding the autonomy of robotic surrogate, Hebert et al. [111] proposed a cooperative framework, which integrated high-level human cognition and commanding with the intelligence and processing power of autonomous robotic surrogate. With this architecture, high-level supervisory commands and intents from human users could be interpreted by robotic surrogates to perform whole-body manipulation tasks autonomously. Furthermore, through a series of studies, Kim et al. [155] concluded that augmented reality human surrogates have certain unique aspects due to the characteristics of AR technology, such as visually seamless connection between AR human surrogates and physical environment including real humans. Although, these features could provide some benefits in the aspect of visual plausibility, this uniqueness could also become a limitation, since AR human surrogates could not be physically influenced by their real environment, or be able to change the physical environment during interaction [266].

2.5 Perception

2.5.1 Human Perceptual System

According to Card et al. [49], the information processing on the human side, for example, during exploration of a VE, could be generalized as three sections: (i) perception, (ii) cognition and (iii) action. This partition is an analogy to the information processing on the computer side, which could be also summarized in three steps: (i) input, (ii) computing and (iii) output.

Figure 2.10 shows the information process during VR-based human-computer interaction. In such an interactive loop, the output of the computer (for example, audio-visually rendered virtual scene) is perceived by the user through the visual, auditory or other senses filter and selection. After that, the various cues coming from different channels will be integrated and analyzed in order to achieve a consistent perceptive result. Then, according to this result, the user will make decisions to perform some corresponding actions (such as clicking the button on the controller or some physical motions). On the computer side, the actions of

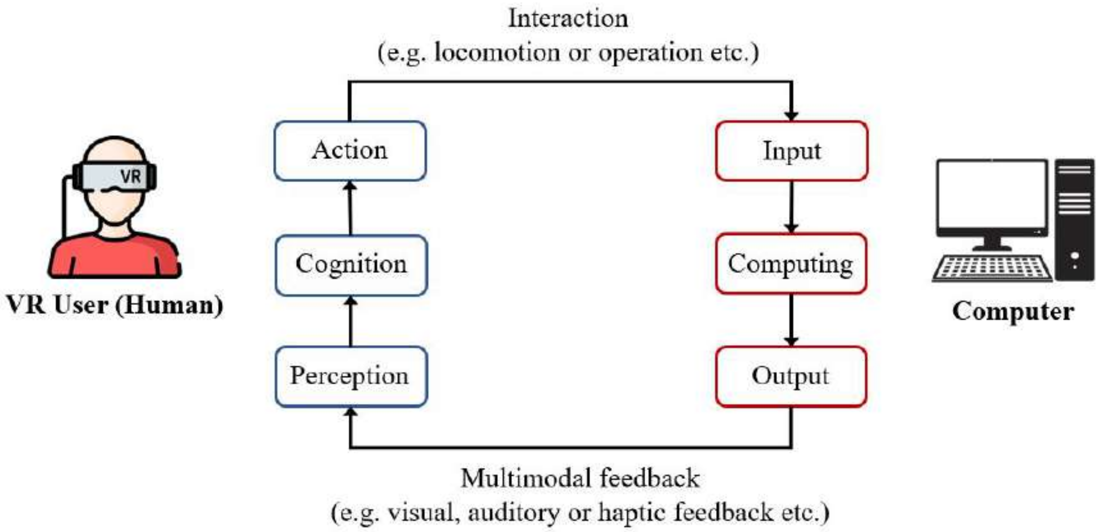


Fig. 2.10 Illustration of information processing in a VR-based human-computer interactive cycle.

the user will be sensed via different sensors, which converts these actions into some binary representation for the computer. And the results of computing will be the updated output of the computer for the next interactive cycle.

In VR systems, such an interactive cycle is usually repeated many times per second (often 90 frames per second (fps) is suggested [135, 220]), which could guarantee a fluent interaction between the user and the computer, and help avoiding disorientation, nausea, and other negative influences on the user experience.

Figure 2.11 presents a structural flow chart of the human perceptual system. Among all of the perceptual channels, vision is often considered as the most important one, which receives and processes most information from our surroundings [305]. The light from the environment is captured by the human eyes and projected on the retina. There are more than 130 million visual receptors distributed on the retina, which could transfer the light stimuli into specific electrical signals. The electrical signals then are transmitted to the visual cortex of the brain to be further processed [284, 77, 101].

In a similar way, the human ears perceive two kinds of input stimulus - auditory [43] and vestibular [281]. For the perception of auditory information, the sound waves are captured by the human ears, which in turn causes a series of vibration of the eardrums. Such vibration is then forwarded to the small bones in the middle ear to lead to their slight vibrations, which cause the fluid in the inner ear to flow and stimulate the auditory receptors on the cochlea.

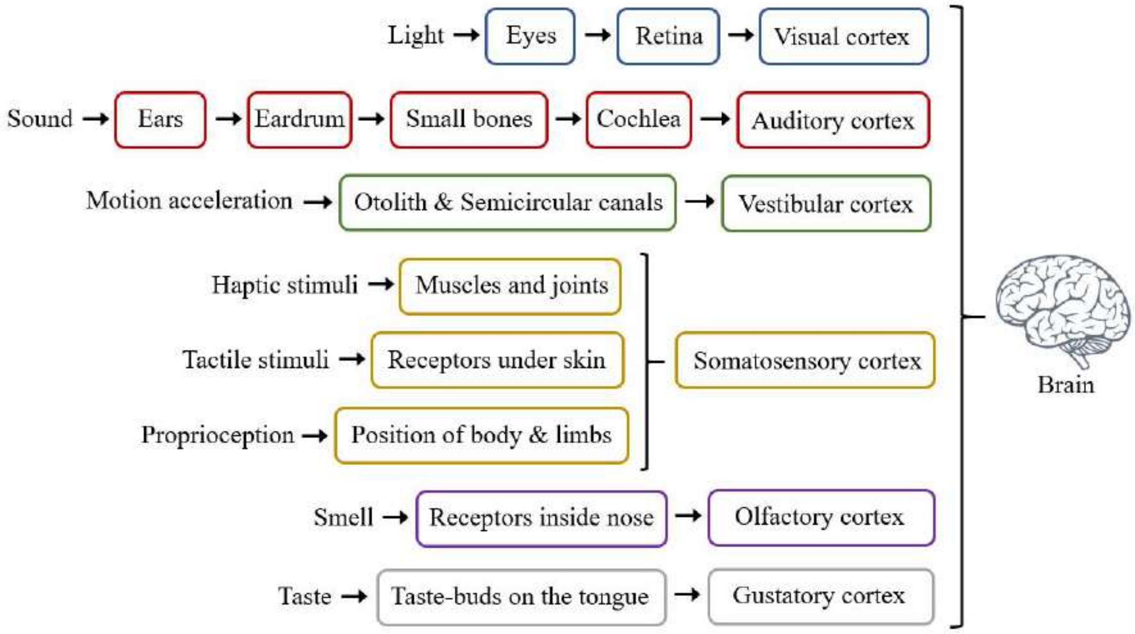


Fig. 2.11 Structural flow chart of human perceptual system.

After that, the electrical signals from the auditory receptors are sent to the auditory cortex on the brain to be processed into corresponding auditory information [331, 48, 209, 26].

The perception of vestibular stimuli mainly depends on the otolith [13] and the semicircular canals [27] in the inner ear. Because of the existence of the fluid in the inner ear, the movements of human as well as the gravity could cause corresponding movements of the fluid, which as a result stimulates the vestibular receptors. The otolith is responsible for transmitting the information about linear acceleration and gravity, while the semicircular canals deliver the information about angular acceleration [94, 247, 64]. Hence, the otolith could only sense linear translation and gravity, while the semicircular canals provide feedback about rotation. However, when motions of the human body come into a stable state (for example, a static state or a constant velocity), the fluid in the inner ear will also return to a steady state over time. Then, the vestibular perceptual system could not tell that the current state is a no or a constant velocity state.

The human's sense of touch [88] is another main channel for perceiving and interacting with external environments, which consists of three different perceptive forms: haptic, tactile and proprioception. Haptic perception [171, 103, 143] is normally caused by external force input to the human body. Such external stimuli could lead to relevant changes on the human muscles and joints like stretching. These changes then are captured by the receptors on the muscles and joints and passed to the somatosensory cortex on the brain. Tactile perception [63] is usually based on the receptors under the human skin and could perceive the stimulus such as pressure, pain and temperature. Proprioception [100] could reveal the perception of self-motions and tell a relevant position of body and limbs, which is generally believed to have an influence on the human's balance ability. Moreover, haptic and tactile perception could both be active or passive, which depends on the human's subjective aims or passive acceptance [139].

In addition, there are also other channels that are related to the human perceptual system such as taste and smell. In this dissertation, these two channels are not considered.

2.5.2 Perception of Self-Translation

Translation describes a process of movement from one place to another, which is regarded as one of the most fundamental locomotion methods. For the perception of self-translation, many previous researches found that a translation in VEs are consistently perceived compressed or underestimated compared to distances in real environment [177, 44, 40, 61, 60, 261]. Distance and speed of translation are the two main measurements in the relevant experiments. In some situations, these two measurements can be seen as equivalent, since speed is defined as distance divided by time. Hence, from the perspective of distance and speed, this phenomenon could be described more intuitively as: when performing a translation in VEs, people usually tend to walk a longer distance (or walk faster during the process) to achieve the target distance (or the target speed) in their mind.

In order to find reasons for this phenomenon, Willemsen et al. [324] explored two possible causes for the compressed distance perception in VEs. The results illustrated that display technology had a significant effect on distance perception in VEs, but the difference of distance judgment between a photographic panoramic environment and a rendered virtual environment is not obvious. Plumert et al. [243] verified this results via another user study and pointed out that the distance perception in VEs is more accurate when using large-screen immersive displays than using HMDs. Slightly different from Williemsen's conclusion, the results from Phillips et al. [237] illustrated that distance judgement is more accurate in a photorealistically rendered virual replica environment than in an non-photorealistically rendered (NPR) replica environment, however, people could improve their distance judgement in an NPR replica environment when given a first perspective avatar. Similarly, the study from Ries et al. [257] also presented a significant improvement on the egocentric distance judgment for the users equipped with a virtual avatar than those without self-embodiment.

Messing et al. [195] pointed out that the angle of declination from the horizon could be a strong cue to distance, and lowering the horizon line can produce "expansive" judgment of distance. Moreover, Steinicke et al. [286] presented that users can significantly improve their distance estimation skills when starting VR experience via a transitional environment. Kelly et al. [151] suggested that walking interaction as well as visual preview could both improve distance judgments significantly, but walking interaction may be more effective than visual preview for promoting distance perception in virtual environment. Bruder et al. [42] illustrated that optic flow manipulations could significantly affect the self-motion perception of users, and with such manipulations, the underestimation of travel distance could be compensated effectively.

Besides of the potential reasons mentioned above, there are also some factors that are likely not the source of distance underestimation and have been figured out. For example, Piryankova et al. [240] pointed out that the impact of stereoscopic depth cues in the flat large-screen immersive display is not effective to result in veridical distance perception. A similar conclusion was also got from the study of Willemsen et al. [325]. Their results illustrated that the limitations on the stereo imagery presentation, which is inherent in HMDs, are likely not the source of distance underestimation. In addition, Messing et al. [194] suggested that neither restricted field of view nor the level of graphical detail was the reason of distance compression in VR, which was also verified by the studies of Creem et al. [61] and Knapp et al. [160].

For speed perception in VR, Pretto et al. [245] explored the influence of the size of FOV on speed perception. They found that central FOVs smaller than 60° would lead to an underestimation on the visual speed. However, when occluding the central area (even only the central 10° of vision was occluded) and leaving only the peripheral visual information, the visual speed would be overestimated. Besides, another study from Pretto et al. [244] demonstrated that the speed perception relied on the spatial distribution of contrast over the virtual scenario more than the global level of contrast per se. From the study, they confirmed that the speed would be underestimated when the contrast is reduced uniformly for all objects in the visual scenario no matter with the distances from the viewer. However, when the contrast is reduced more for distant objects, which is like the case in real fog, the visual speed would be overestimated. Banton et al. [22] suggested that the change in speed perception from straight-ahead to side or down gaze is consistent with a shift from expanding of optic flow to lamellar flow. For this reason, lamellar flow is thought to be essential for accurate speed perception. However, during straight-ahead gaze, restricted FOVs usually eliminate this important cue. A similar effect of limited FOVs on speed perception has been verified by Mohler et al. [202] and Nilsson et al. [216] as well. They pointed out that visual flow is necessary for a correct speed perception, but this cue is usually cut off by a limited FOV when the user with HMD looks straight forward. Furthermore, a significant result was found showing that when participants perceived visual flow, they could gave more accurate speed estimation. In addition, Colombet et al. [55] found that the speed estimation would be significantly improved with the increase of visual scale factor, which is defined as the ratio between the geometric field of view (GFOV) and the real field of view (FOV).

In particular, Bruder et al. [41] analyzed the perception of three components of locomotion (speed, distance and time) during immersive walkthroughs. The results of the study illustrated that people usually significantly underestimate virtual distances, slightly underestimate the virtual speed and slightly overestimated the elapsed time.

2.5.3 Perception of Self-Rotation

Dodge et al. [69] described the perception of self-rotation as a process of multi-sensory information integration. The sensory data that contributes to the perception of self-rotation is derived not only from the vestibular sense, but also from vision, audition, kinaesthesis, muscular strains, articular and dermal sensations, eye-movements, and changing configuration of pressures. Among these sensory inputs, the visual and vestibular channels contribute the most on the perception of self-rotation. Mergner et al. [193] investigated the visual contribution to the human self-motion perception during horizontal body rotation. The results of the experiment illustrate that in normal situation, the visual information is usually regarded as a reference in the self-motion perception, while the vestibular input is used for checking the kinematic state. When being within a dynamic scene, the human visual channel will be suppressed more or less and the self-motion perception will be mainly based on the vestibular cues. Zacharias et al. [330] and Mergner et al. [192] suggested that the human estimation on the rotation velocity could be modeled as a function correlated with the vestibular and peripheral visual field motion cues simultaneously. In low-frequency situations, a parallel channel linear model is proposed, which has separated the visual and vestibular pathways summing in a complementary way. At higher frequencies, the vestibular cues would contribute more and dominate the human sensation. In addition, the proposed model and function could be supported and extended by the non-linear cue conflict model [330], in which the level of agreement between the visual and vestibular cues determined the cue weighting.

Furthermore, Riecke et al. [256] suggested that considerable navigation improvements can already be gained by supporting full-body rotations only. From a series of studies, Barraza et al. [23] found that human subjects could rely on a sensitive measurement of angular velocity. Moreover, they concluded from the results that there are separate mechanisms for human perceptions on the angular and linear velocity. In addition, Nooij et al. [219] pointed out that the sum of perceived rotational and translational components alone can not adequately explain the overall perceived motion. When shaping the perception, people may need additional information coming from the knowledge on the motion dynamics and familiar stimuli combinations.

2.5.4 Perception of Social Cues

Social cues [2] are defined as a series of verbal or non-verbal signals, which influence human impressions and responses to others during social interactions. Usually, social cues could be expressed through the face, voice, gaze, posture, proximity etc. In addition, social cues are important communicative channels as there are considerable social and contextual information released from them, which as a result facilitate social understanding and interaction [265, 67].

When having a face-to-face conversation, people often tend to focus on the face of each other. As one of the most intuitive social cues, facial expressions usually contain implicit information and self-state of the humans during a face-to-face interaction. Frith et al. [85] investigated the role of facial expressions in the social interactions. They pointed out that facial expressions could lead to rapid responses, and that people tend to imitate the emotions in the observed face. Moreover, facial expressions are not simply reflexive, but also include communicative components. For example, we usually tend to want people to know that we are empathetic, especially when realizing that we are being observed by other people, and hence, we even exaggerate emotional expression. They also concluded that among all of the facial expressions, ostensive gestures like eyebrow flash are especially important, because they indicate the intention to communicate.

Another important social cue during conversation is the vocal tone. A research from Simon et al. [277] indicated that the vocal tone of a talk could communicate emotions such as anger, fear, sadness, awe, compassion, interest, or embarrassment. In addition, they summarized from a series of studies that the vocal tone could be regarded as a rich modality for emotion display and could inform fundamental constructs about the emotion.

For the non-verbal communication, subject's gaze direction is usually an useful index for the observers to understand the focus or interest as well as the psychological state of the subject in the current environment, which could facilitate the communication between each other. For example, Kinsbourne et al. [158] found an interesting phenomenon that during verbal thought, human subjects always tend to look to the right, while during spatial thought they usually look up or to the left. This characterized expression in human behavior could be also regarded as an evidence of the cerebral lateralization of cognitive function. Moreover, Adams et al. [3] verified that the processing of facially communicated approach-oriented emotions like anger and joy could be facilitated by direct gaze, while the processing of facially communicated avoidance-oriented emotions like fear and sadness could be promoted by averted gaze. Such results suggested that there was a combination between gaze direction and facial expressions in the processing of emotionally relevant facial information. Besides, Marschner et al. [184] suggested that the direction of subject's body and head play a minor, but significant role in social communication, which could lead to attentional allocation as well as increased emotional responding. However, earlier findings suggested that mutual eye interaction is the main source for that.

Body language or posture is another major expression of social cue. Especially in the emotional communication and interaction, body language or posture always plays an important role for the observers to perceive and understand the emotional changes or mental activities of the humans. The human faces can hide and cover the real emotional activities when people are terrified to show emotions. However, from the bodily expression, the observers can easily get a particular emotion that is correlated with some specific actions without much need for interpretation of the cue [65]. In addition, body movement and posture could also convey specific emotional information [89]. Dael et al. [62] revealed that some patterns of body movement systematically happen in the portrayals of specific emotions, allowing emotions to be recognized. Even though some of the emotions were expressed by one particular pattern, most of them were variably expressed by multiple patterns, which could be explained as reflecting functional components of the emotions.

In addition, proximity [106] is another significant social index, which indicates the extent of interpersonal intimacy and trust between subjects. According to Mccall et al.'s study [187], some proxemic behaviors could also reveal the affective states (like interpersonal attitudes) and cognitive responses (like social attention) of interactants. All of these aspects could help people to determine suitable social behaviors during communication and interaction based on the proxemic feedback from social subjects. The perception of proxemics will be introduced and discussed in more details in Chapter 5.

Chapter 3

360° VR-based Telepresence System

This chapter presents a full description of the system design, development and other technical details on a prototype of a 360° VR-based telepresence system. The subsequent research studies in this dissertation are all based on this setup.

Furthermore, in order to enhance the space perception in 360° video-based VEs, virtual objects such as buildings, cars, or avatars can be displayed and blended with the real content (from the 360° immersive video). To generate global illumination and shadows for virtual objects in 360° video-based VEs, we implement Rhee et al.'s lighting and shadowing algorithm [254] in Unity3D, and perform a user study based on it, in which we explore the following research question:

- *To what extent will the illumination and shadow effect of virtual objects influence object localization tasks and user experience in 360° MR environments?*

3.1 System Design and Development

3.1.1 Motivation

As described in Chapter 1, current telepresence systems usually lack natural ways to support the interaction and exploration of REs. In particular, single webcams for capturing the RE could only provide a very narrow FOV and a limited illusion of spatial presence for the user. Furthermore, the movement control of mobile platforms in today's telepresence systems are often restricted to simple interactive devices, such as joystick, touchpad, mouse or keyboard. All of these issues may decrease the naturalness, task performance and overall user experience [299].

In order to address these limitations and challenges, this chapter introduces a prototype of a 360° VR-based telepresence system, which aims to provide the user in the LE with a more natural and intuitive way to explore and visit a RE with the help of a telepresence robot.

3.1.2 System Design

Figure 3.1 illustrates the basic components and structure of the 360° VR-based telepresence system. A mobile robot equipped with a 360° camera serves as a physical surrogate of the user in the RE. The mobile robot provides the 360° camera with mobility to be moved through the entire RE. The 360° camera captures a panoramic live stream from the RE and then transfers the captured scene to the LE in real-time. The robot control and data exchange in the RE are implemented on a laptop via the robot operating system (ROS).

In the LE, The VE is reconstructed and rendered in Unity3D based on the received live-stream data from RE. An HMD is provided to the user in the LE to display the reconstructed virtual representation of the RE in real-time. All reconstruction and rendering processes in the LE are implemented on a graphics workstation.

The user's movements in the LE are detected by a set of tracking systems. The user's position and orientation in the tracking space of the LE will be updated in real-time and mapped to the RE to control the robot's movements. This way, the user can steer the telepresence robot through the RE by means of natural walking in the LE. Compared with other methods of movement control for telepresence robots, walking in the LE is a more natural and intuitive way to travel in the RE from one location to another [288, 169]. Since the position of the 360° camera in the RE is determined and updated based on the state of the user in the LE, this approach provides the most consistent and intuitive perception of locomotion in the target environment, while releasing the user's hands for other potential interactive teleoperation tasks as well.

One major limitation of this approach is that the layouts of the LE and the RE should be identical. However, in most cases, the local tracking space is usually significantly smaller than the area in the RE, which the user wants to explore. Moreover, the LE and the RE typically have completely different spatial layouts. For VR scenarios, redirected walking (RDW) methods have been introduced [250] to solve this problem, which are adopted to the 360° VR-based telepresence system in the scope of this thesis. RDW is based on the real walking and guides the user on a path in the real world, which might vary from the path the user perceives in the VE. RDW can be implemented by manipulations applied to the virtual camera, which cause the user to unknowingly compensate for scene motions by repositioning and/or reorienting themselves [297].

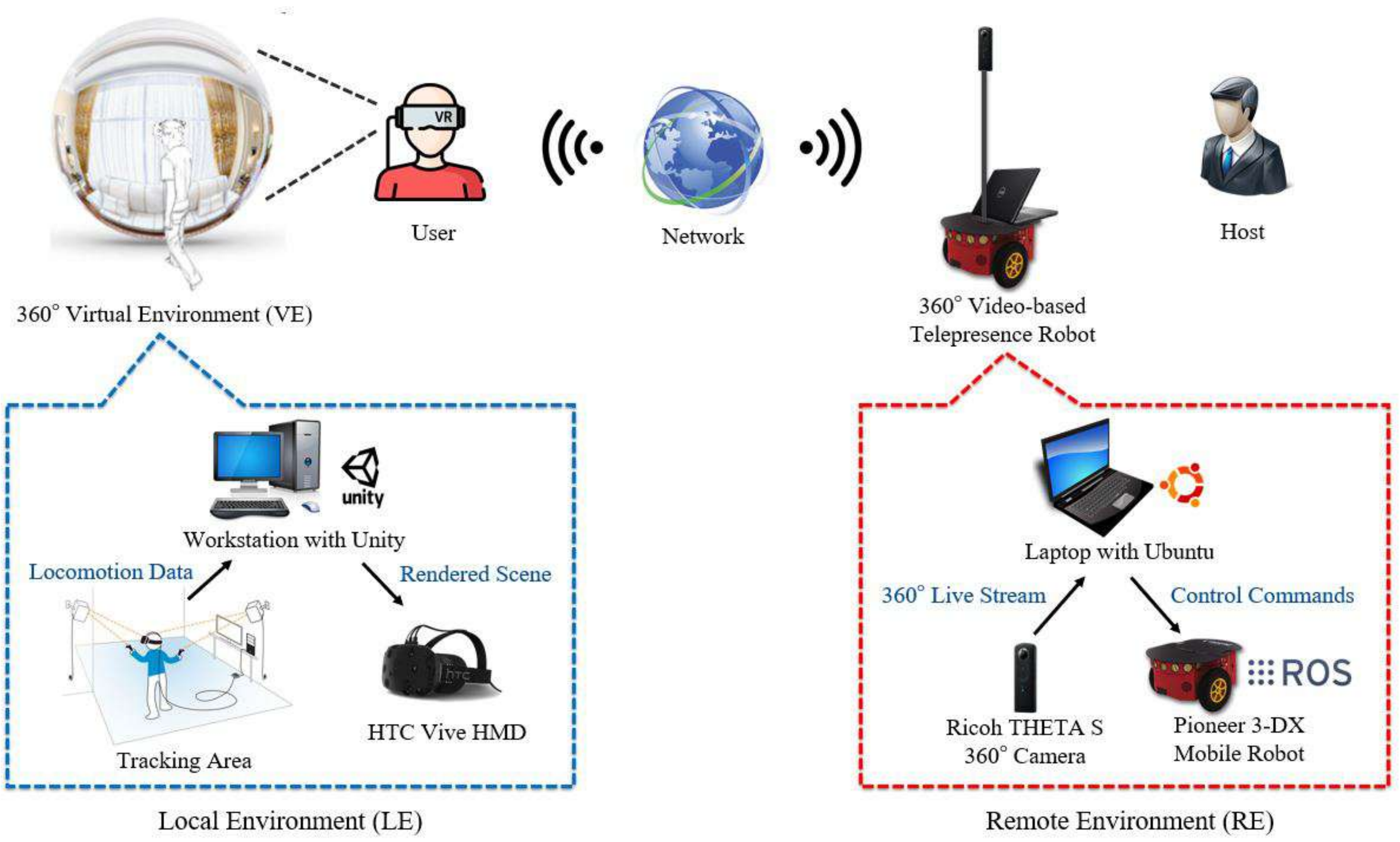


Fig. 3.1 Components and structure of 360° VR-based telepresence system: in the RE, a mobile robot equipped with a 360° camera serves as a surrogate of the local user and captures a 360° full-view live stream of the RE, and then transmits it to the LE in real time via a communicating network. In the LE, the received live stream is rendered and projected inside a spherical space and displayed on the user's HMD. The user wearing an HMD teleoperates the telepresence robot moving through the RE by means of natural walking in the local tracking space.

Such manipulations without the user's awareness is possible because the sense of vision often dominates proprioception [29, 68] as explained in Section 2.5. In other words, the visual feedback that the user perceives from the HMD corresponds to the motions in the VE, whereas proprioception and vestibular system are connected to the physical motions in the real world. When the discrepancy is small enough, it will be difficult for the user to detect the redirection, which could lead to the illusion of an unlimited natural walking experience [249, 287].

3.1.3 Implementation

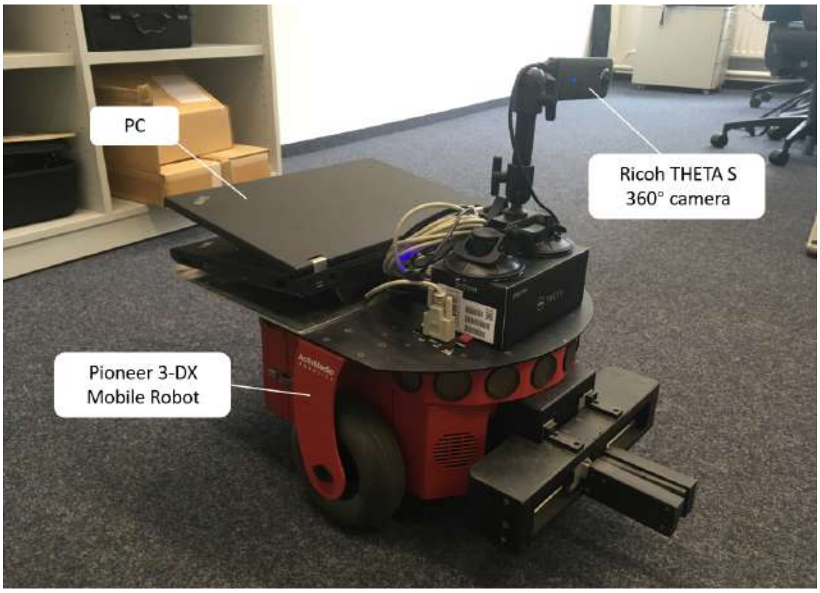


Fig. 3.2 Prototype of the 360° VR-based telepresense robot consisting of a Ricoh THETA S 360° camera, a Pioneer 3-DX mobile robot and a laptop running ROS Indigo.

Figure 3.2 illustrates a prototype of the 360° VR-based telepresence robot. A Pioneer 3-DX mobile robot working in a differential-drive way serves as the mobile platform of the telepresence system. In addition, a Ricoh THETA S 360° camera is equipped upon the mobile robot for capturing a 360° live stream from the RE, which works in a 1280×720 resolution and a 15fps frame rate. Both, the mobile robot and the 360° camera are connected with a laptop via USB cables. The laptop runs a ROS Indigo and serves as the processing core for robot movement control, device driving, remote communication as well as message publishing and subscribing. There are two nodes running under the ROS Indigo on this laptop, which are responsible for controlling the robot movement and capturing a 360° live stream separately. When the telepresence system is working, the ROS nodes publish the 360°

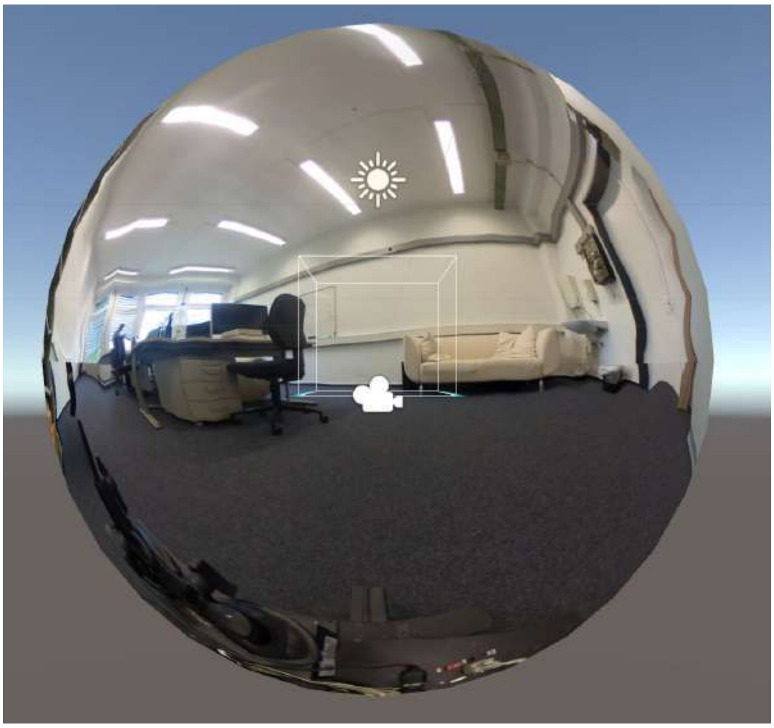


Fig. 3.3 Reconstructed virtual scene in Unity3D which presents the real-time rendered RE as the inner texture of a spherical space.

live stream ROS messages from the RE to the LE via communication network; while the mobile robot subscribes the movement control ROS messages from the LE simultaneously, and updates its position and orientation in the RE according to the information within the ROS messages.

In the LE, the user is provided with an HTC Vive HMD, which displays the 360° video-based RE with a resolution of 1080×1200 pixels per eye. The diagonal field of view is approximately 110° and the refresh rate is 90Hz. The tracking area is covered by a pair of lighthouse tracking stations to detect the update of user's position and orientation, by which the sensors on the HMD can be tracked in real-time. The tracking data of user's movements in the LE is packaged in the form of ROS messages, and transmitted to the RE, then interpreted into the movement control commands of the mobile robot in the RE. In this way, controlling a robot's movements in the RE by means of natural walking in the LE is successfully implemented in a one-to-one mapping. A graphics workstation with a 3.5GHz Core i7 processor, 32GB of main memory, and two NVIDIA Geforce GTX 980 graphics cards serves as the processing core of the LE, on which the remote scene reconstruction and rendering are performed. The connection between the HMD and the graphics workstation is based on an HTC Vive 3-in-1 (HDMI, USB and Power) cable with a length of 5 meters, so that the user could move within the tracking space. Furthermore, the reconstruction and rendering of virtual remote scene are implemented based on a spherical space modelled in Unity3D. The live stream from the RE is rebuilt and projected as a movie texture on the inner surface of the spherical model (Figure 3.3). A virtual camera is located in the center of the 360° VE in order to provide a perspective-correct view for the user to explore the RE. This way, with the frame update of live stream from the RE, the user could get a 360° telepresence view of the RE on the HMD in real-time. Moreover, the communication between the LE and the RE is implemented via ROSBridge between ROS and Unity3D [166]. In addition, RDW technology and corresponding manipulated gains [334] are applied to the telepresence system during remote interactions in order to allow the user to explore a much larger RE with different layouts compared to the LE (see Chapter 4).

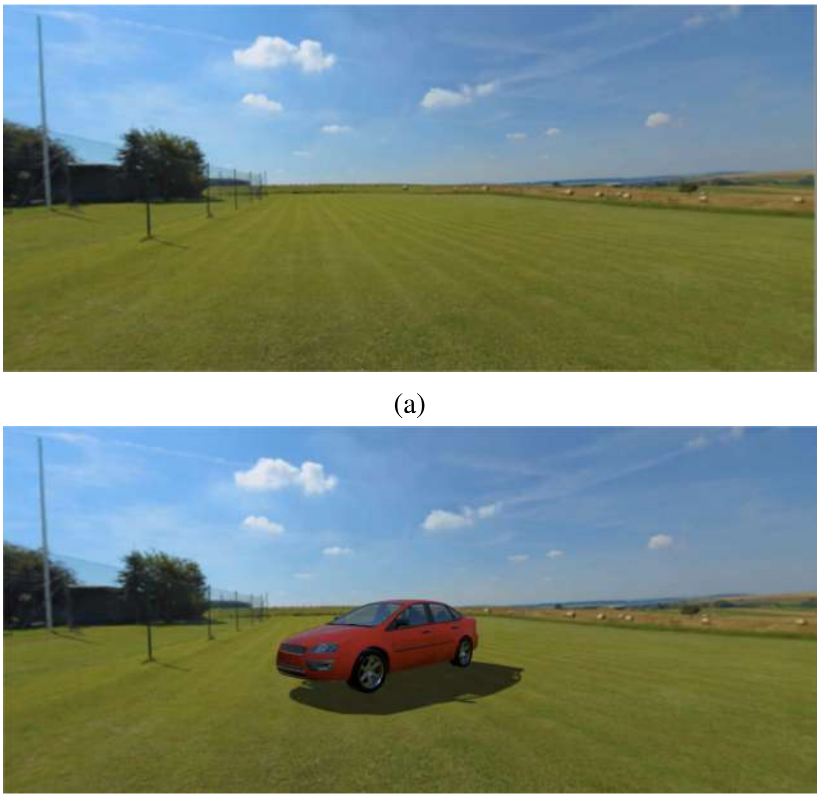
3.2 Global Illumination and Shadows in 360° MR Environments

3.2.1 Motivation

As introduced in Section 3.1, the traditional way to reconstruct a remote scenario using the input panoramic or omni-directional video is to attach it onto a spherical surface as movie texture [333] or to set as the wrappers of the skybox in VEs. As the content often displays scenes at larger distances, stereoscopic perception is limited, and therefore, depth cues for spatial presence might be hindered [28].

In mixed reality (MR) environments, these immersive videos can be augmented by virtual objects such as buildings, cars, or avatars [198] to blend real content (from the immersive video) with computer-generated virtual objects. Furthermore, the display of familiar objects inside the immersive video can also improve spatial perception [205]. However, the integration of real and virtual objects in 360° video-based VEs, raises the challenge of consistent global illumination [218, 235].

Illumination and shadows of virtual objects in MR can provide important depth cues for visual perception and spatial localization, and furthermore, could improve the overall user experience. Previous work has introduced MR algorithms, which allow for seamless composition of 3D virtual objects into a 360° video-based virtual environment (VE). For instance, for consistent lighting and shadows, the lighting source in the panoramic video can be detected and used to illuminate virtual objects [254, 253] (we implemented this lighting and shadowing algorithm for 360° video-based VEs in Unity3D, which could be seen in Figure 3.4, and the following user study was conducted based on it), or the panoramic video can be exploited for reflections on specular surfaces of virtual objects [291]. The new scenario consisting of 360° video-based VEs and 3D virtual objects is referred to as *360° MR environments*.



(a)

(b)

Fig. 3.4 (a) An in-game view of one portion of a 360° video-based VE; (b) A 3D virtual car was placed in the 360° video-based VE with merged lights and shadow.

While these MR algorithms provide visual consistent MR renderings, no research has been conducted yet to investigate to what extent such illumination and shadow effects could improve object localization and user experience in 360° MR environments. To address this research question, we conducted a user study in which we compared object localization with and without illumination and shadows of virtual objects in 360° MR environments. Furthermore, we evaluated the effects of illumination and shadows on user experience.

3.2.2 Related Work

Space perception is an important research topic and a vast body of literature has considered depth and size perception in virtual [133] and augmented reality [114, 142]. The majority of the previous works has shown that distances to virtual objects are often underestimated, in particular, when the objects are displayed at larger distances. It has also been shown that illumination and shadows can improve the depth perception in computer-generated VEs as well as in AR scenarios, in which users see the real world through an optical see-through HMD such as Microsoft Hololens.

Illumination and shadows of virtual objects in 360° MR environments can provide important depth cues for visual perception and spatial localization, and add a level of realism to virtual objects [107], which could help to integrate virtual objects into the rendered VE based on 360° videos smoothly and improve the overall user experience. For example, previous works by Steinicke et al. [291] introduced a MR setup in which a web camera was used to capture parts of the surrounding, which was in turn used to render lights and reflections of specular virtual objects. Furthermore, Rhee et al. [254] has proposed an algorithm which achieved seamless composition of 3D virtual objects into a 360° video-based VE by using the input panoramic video as the lighting source to illuminate the virtual objects.

Though, there is a large body of literature regarding global illumination and shadow effects of virtual objects in typical VR and AR environments, and the first algorithm to display virtual objects in 360° video-based VEs, the effects of virtual object illumination and shadows on the space perception and user experience in 360° MR environments still remains poorly understood, which motivates the user study in this section.

3.2.3 User Study

In our user study, we analysed the effects of global illumination and shadows of virtual objects on the object localization and user experience in 360° MR environments.

Hypothesis

We formulated the following hypotheses that we wanted to test in our study:

- *H1: Participants will have higher localization accuracy with illumination and shadow effects than without.*

- *H2: Participants will have a better localization ability at closer distances.*

- *H3: Participants will have better user experience with illumination and shadow effects than without.*

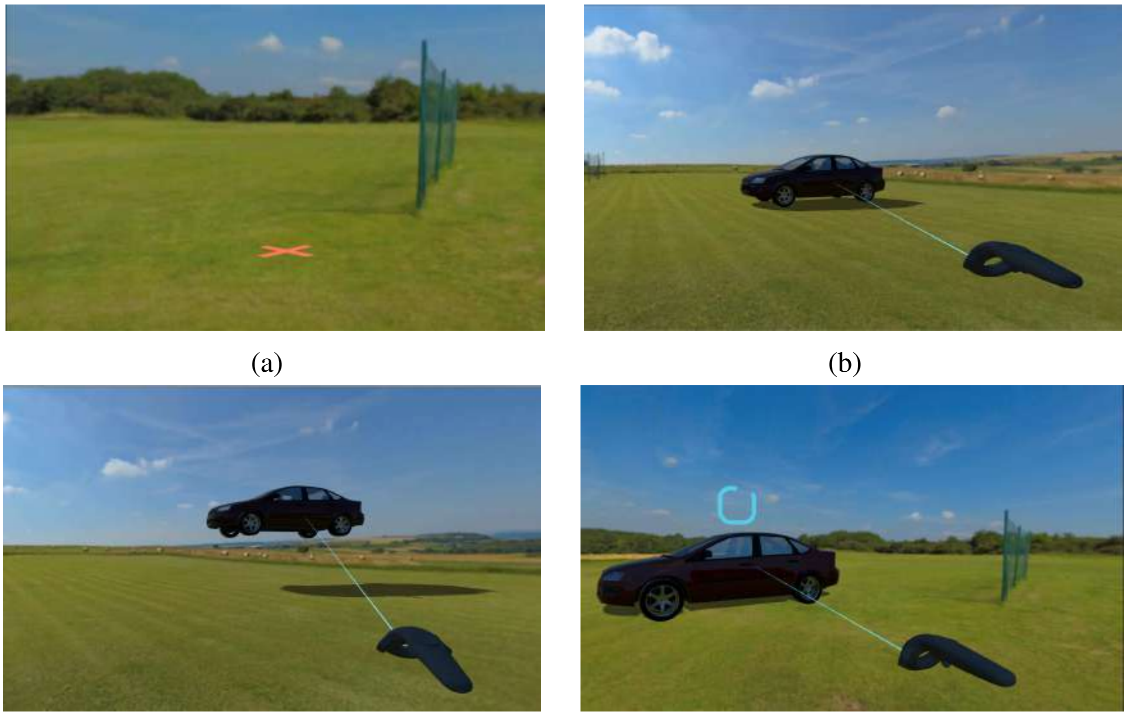

(a) (b) (c) (d)

Fig. 3.5 Screenshots of the task procedure with virtual object illumination and shadows: (a) the target position was marked with a red cross on the ground; (b) the virtual car was picked up from its default position using VR controllers; (c) the virtual car was being moved to the target position; (d) the placing operation was being confirmed.

Participants

15 participants (8 male, 7 female, average age 23.73 years, SD = 2.67) completed the user study, which were all students from our university. All participants had previous experience with VR HMDs.

Experimental Setups

As illustrated in Figure 3.5, we generated a 360° video-based VE showing a wide grass field with a resolution of 8192 x 4096 pixels. The VE was rendered with the Unity Engine on a workstation, which has an Intel Core i7-4930K CPU with 3.40 GHz, an Nvidia GeForce GTX 1080 Ti GPU and a 16GB main memory. The rendered VE was displayed on a HTC Vive HMD with a resolution of 1080×1200 pixels per eye. The diagonal field of view is approximately 110° and the refresh rate is 90Hz. The global illumination and shadows of virtual objects in the 360° video-based VE was implemented based on the research of Rhee et al. [254, 253].

Material and Methods

During the user study, participants were required to stand in the center of the 360° video-based VE (0,0,0) and move a polygon rendering car from its default position (8m away from the participants on the right side) to the target position (marked on the ground) using VR controllers. The task was designed with three factors: (i) *distance* (2 levels: 12m, 18m), (ii) *angle* (5 levels: 0°, 45°, 90°, 135° and 180° in the yaw axis of human body) and (iii) *illumination and shadow effects* (2 levels: with and without). The target position was determined by the factors of *distance* and *angle*, and marked on the ground until the moving operation started.

Prior the user study, participants were asked to fill out a demographic questionnaire as well as Kennedy's simulator sickness questionnaire (SSQ), and finish some training trials. Afterwards, participants signed in with their IDs and started the formal stage of the user study. Appropriate body rotation is allowed for participants to support them finishing the task. The task procedure can be seen in Figure 3.5.

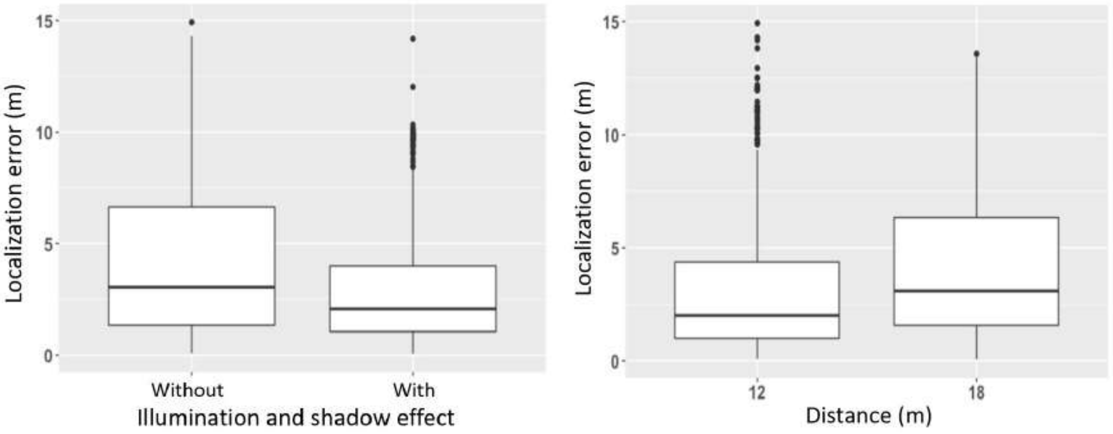
In summary, the user study was a within subject design with 2 distances × 5 angles × 2 illumination and shadow effects = 20 conditions, and each condition was repeated for 6 times. Thus, every participant needed to finish 20 conditions × 6 repetitions = 120 trials during the task. All the trials were shuffled and appeared in a randomized order. For the entire user study, 15 participants × 120 trials per participant = 1800 total trials were collected.

We measured the *localization error* for every trial after participants confirmed their operations, which indicates a bias between the target position and the center point of the car model on the x-z plane. After the user study, participants were asked to compare the illumination and shadow effects (with vs without) and rate their *preference* from 1 to 7 (higher is better). Moreover, the post-SSQ questionnaire as well as the igroup presence questionnaire (IPQ) were also required.

3.2.4 Results

The experimental data was analyzed with the *analysis of variance* (ANOVA). A normality assumption check was performed before the analysis using the Shapiro-Wilk test, which did not show a strong indication of normal distribution. However, as shown in previous research [110, 180], moderate deviations from normality can be tolerated by ANOVA.

Figure 3.6a shows the effect of illumination and shadows of virtual objects on the localization error. The average localization error is 4.296m (SD = 3.582) without illumination and shadows, and 3.094m (SD = 3.05) with illumination and shadows. This means that the average localization error with illumination and shadows was lower by 27.98% than without



(a) (b)

Fig. 3.6 Results of the user study: (a) illumination and shadow effects on the localization error; (b) distance on the localization error.

illumination and shadows. The ANOVA results show a significant effect of *illumination and shadows* of virtual objects on the localization error ($F_{1,14} = 4.688, p = 0.048, \eta^2 = 0.130$). These results illustrate that the global illumination and shadows of virtual objects could significantly help users to improve their localization accuracy in 360° MR environments, which verified the hypothesis *H1*.

Figure 3.6b shows the effect of distance on the localization error. The average localization error is 3.179m (SD = 3.156) at the distance of 12m, and 4.21m (SD = 3.516) at the distance of 18m. In other words, the average localization error at the distance of 12m is lower by 24.49% than at the distance of 18m. The ANOVA result shows that the *distance* has a significant effect on the localization error ($F_{1,14} = 38.58, p < 0.001, \eta^2 = 0.099$). These results demonstrate that the distance could significantly influence the localization error. In addition, when localizing a virtual object in 360° MR environments, human users usually have a better localization ability at closer distances. Thus, hypothesis *H2* was verified. No significant interaction effects between factors were found.

Figure 3.7 shows the preference rate of participants on the illumination and shadow effects, in which the likable scales are set ranging from 1 (not like it at all) to 7 (like it very much). For the situation with illumination and shadows, the average preference rates is 5.27 out of 7, which is significantly higher than 3.27 for the situation without illumination and shadows (confirmed by a one-sided Wilcoxon test with $p < 0.001$). This result illustrates that participants have better user experiences when interacting with a 360° MR environment, in which virtual objects are seamlessly integrated into the 360° video-based VE with global illumination and shadow effects. Thus, hypothesis *H3* was verified.

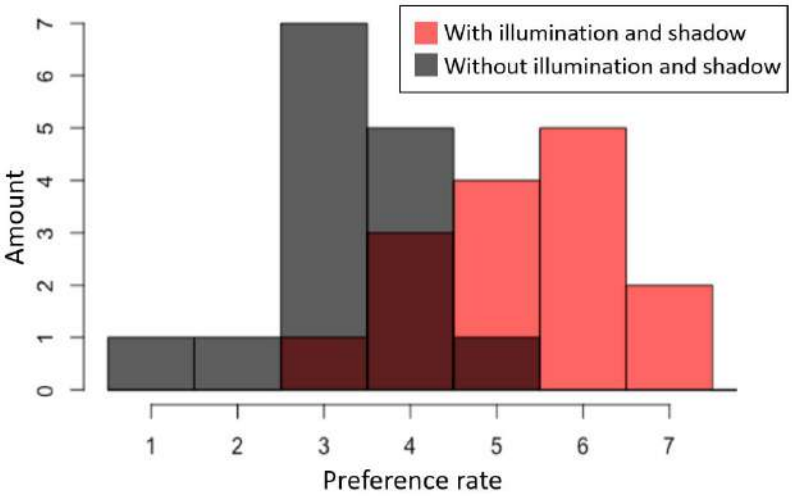


Fig. 3.7 Results of the preference rate on the illumination and shadow effects.

Table 3.1 Results of the IPQ.

	With illumination and shadows		*Without illumination and shadows*	
	Score	*STD*	*Score*	*STD*
Spatial presence	3.57	1.07	3.05	1.05
Involvement	3.32	1.47	2.93	1.53
Experienced realism	2.58	0.85	1.56	0.91

Furthermore, participants' sense of presence in 360° MR environments was evaluated with the data of the IPQ. The results (see Table 3.1) show that the illumination and shadow effects have a significant influence on the *experienced realism* ($p = 0.0018$). However, no significant influence was verified for the *spatial presence* and the *involvement*.

When participants were asked if they had any cognitive strategies during the task, 4 of them specifically reported that the illumination and shadows of virtual objects helped them to estimate the placing location in 360° MR environments.

3.2.5 Discussion

In this section, we reported a user study in which we evaluated the effects of illumination and shadows of virtual objects on the object localization and user experience in 360° MR environments. The results indicate that the illumination and shadows of virtual objects could significantly reduce the localization error and improve the user experience regarding the experienced realism. Participants also give higher preference evaluations on the experience with virtual object illumination and shadows.

3.3 Conclusion

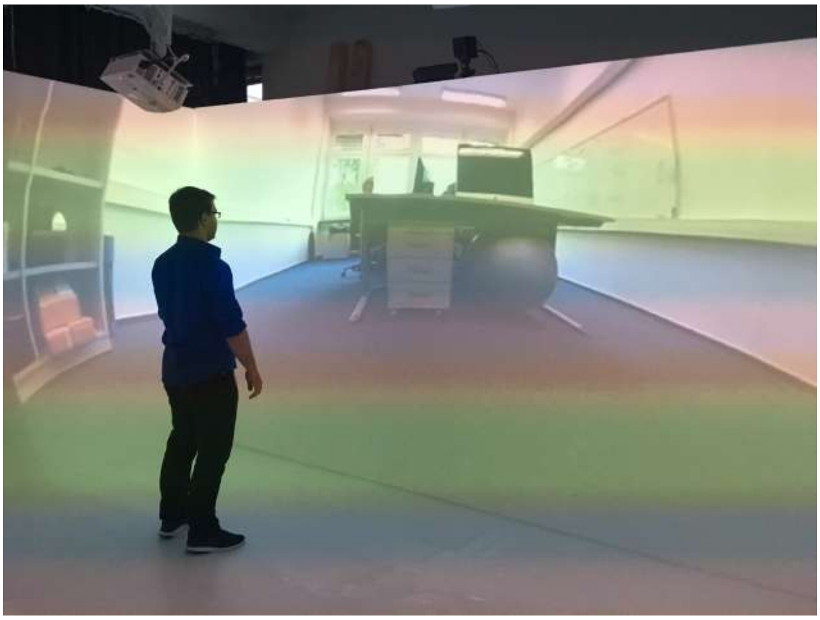


Fig. 3.8 Real-time 360° RE rendered and projected in a CAVE-like space.

In this chapter, a 360° VR-based telepresence system as well as its design and development was presented, which enables the user to explore and interact with a RE by means of natural walking in the local tracking space while perceiving a 360° immersive display of RE on the HMD. We also explore other VR setups in the LE to present the 360° video-based RE. In particular, a CAVE-like projected space (as illustrated in Figure 3.8) for displaying the 360° RE has been implemented and tested.

Furthermore, a user study was conducted in which we evaluated the effects of global illumination and shadows of virtual objects on object localization and user experience in 360° MR environments. The results show the importance of global illumination and shadows of virtual objects, which could significantly reduce the localization error and improve the user experience in 360° MR environments.

Chapter 4

Natural Exploration of Redirection in 360° Tele-mediated Space

In this chapter, the user's ability to detect redirected manipulations of translations and rotations in a 360° video-based RE is explored with a focus on the following research question:

- *How much can humans be unknowingly redirected in a 360° video-based RE?*

4.1 Motivation

Previous work has investigated the human sensitivity to redirected manipulations in computer-generated VE only, however, so far it is unknown how much manipulation can be applied to telepresence mobile platform, which transfers 360° videos of real-world RE rather than computer-generated VE. Furthermore, it seems reasonable to assume that there are significant differences in the perception of self-motions in computer-generated VE and 360° video-based RE, due to differences in visual quality, image distortion or stereoscopic disparity.

Therefore, on the basis of the prototype in the last chapter, two psychophysical experiments are conducted to investigate the amount of discrepancy between movements in the LE and the RE that can be applied without users noticing. More specifically, two experiments are designed in order to find the thresholds for two basic self-motions, i. e., translations and rotations, in 360° video-based RE. The results of these experiments will provide the basis for future immersive telepresence systems in which users can naturally walk around to explore REs when being in the LE that has a different layout.

To summarize, the contributions of this chapter include:

- a psychophysical experiment to identify detection thresholds for translation gains, and

- a psychophysical experiment to identify detection thresholds for rotation gains for natural interactions between human users and 360° VR-based telepresence system based on RDW manipulations.

4.2 Related Works

In this section, some previous works related to locomotion in general and detection thresholds in psychophysics are summarized.

4.2.1 Locomotion

In recent years, different solutions are used to make it possible for users to explore VEs, which are significantly larger compared to the available tracking space in the real world. Several of these approaches are based on specific hardware developments such as motion carpets [271], torus-shaped omni-directional treadmills [33, 32], or motion robot tiles [136, 138, 137]. As an cost-effective alternative to these hardware-based solutions, some techniques were introduced, which take advantage of imperfections in the human perceptual system. Examples include concepts such as virtual distractors [232], change blindness [297, 296], or impossible and flexible spaces [312, 311].

In their taxonomy [295], Suma et al. provide a detailed summary and classification of different kinds of redirection and reorientation solutions in a range from subtle to overt, as well as from discrete to continuous approaches. The solution adopted in this work belongs to the class of techniques that reorient users by continuous subtle manipulations. In this situation, when users explore a VE by walking in the tracked space, manipulations (such as slight rotations) are applied to the virtual camera [249, 287]. Based on these small iterative rotating manipulations, the user is forced to adjust the walking direction by means of turning to the opposite direction of the applied rotation. As a result, the user walks on a curvature in the real space while she perceives the illusion to walk along a straight path in the VE. In other words, the visual feedback that the user perceives from the HMD corresponds to the motions in the VE, whereas proprioception and vestibular system are connected to the real world. If the discrepancy between stimuli is small enough, it is difficult for the user to detect the redirection, which leads to the illusion of an unlimited natural walking experience [249, 287].

4.2.2 Detection Thresholds

Identifying detection threshold between motions in the real world and those displayed in the VE has been in the topic of several recent studies. In his dissertation, Razzaque [249] reported that a 1 deg/s manipulation serves as lower detection threshold. Steinicke et al. [288] described a psychophysical approach to identify discrepancies between real and virtual motions. Therefore, they introduced gains to map users' movements from the tracked space in the real world to camera motion in the VE. In this context, they use three different gains, i. e., (i) rotation, (ii) translation and (iii) curvature gains, which scale a user's rotation angle, walking distance and bending of a straight path in the VE to a curved path in the real world respectively. In addition, they determined detection thresholds for these gains through psychophysical experiments, by which the noticeable discrepancies between visual feedback in the VE on the side and proprioceptive and vestibular cues on the other side in the real world are identified. For example, to identify detection thresholds for curvature gains, participants were asked to walk a straight path in the VE, while in the real world they actually walked a path, which was curved to the left or right using a randomized curvature gain. Participants had to judge whether the path they walked in the real world was curved to the left or to the right using a *two-alternative forced-choice (2AFC)* task. Using this method, Steinicke et al. [288] found that users can not reliably detect manipulations when the straight path in the VE is curved to a circular arc in the real world with a radius of at least 22m. In recent work it has been shown that these thresholds can be increased, for instance, by adding passive haptic feedback [185] or by constraining users to walk on curved paths instead of straight paths only [169].

Several other experiments have focused on identifying detection thresholds for such manipulations during head turns and full body turns. For instance, Jerald et al. [140] suggest that users are less likely to notice gains applied in the same direction as head rotation rather than against head rotation. According to their results, users can be physically turned approximately 11% more and 5% less than the virtual rotation. For full-body turns, Bruder et al. [42] found that users can be physically turned approximately 30% more and 16% less than the virtual rotation. In a similar way, Steinicke et al. [288] found that users can be physically turned approximately 49% more and 20% less than the virtual rotation.

Furthermore, Paludan et al. [227] explored if there is a relationship between rotation gains and visual density in the VE, but the results showed that the amount of visual objects in the virtual space had no influence on the detection thresholds. However, other walk has shown that walking velocity has an influence on the detection thresholds [212]. Then, another study by Bruder et al. [39] found that RDW could be affected by cognitive tasks, or in other words, RDW induce some cognitive effort on users.

While the results mentioned above have been replicated and extended in several experiments, all of the previous analyses have considered computer-generated VEs only, whereas video-based streams of real scenes have not been in the focus yet.

4.3 Concept and Challenges

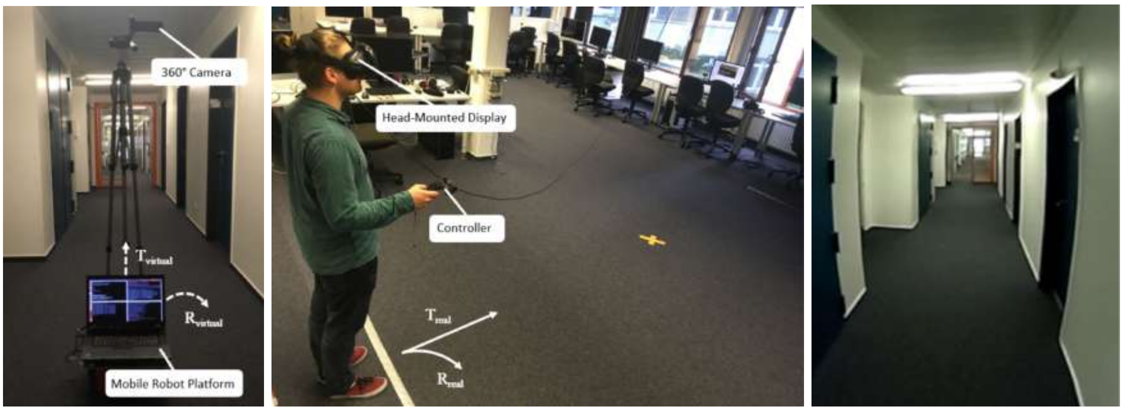
As described in Chapter 1, one of the main challenges of telepresence systems is to allow users to explore a RE by means of natural walking in the user's LE, and thus controlling the motion of the robot platform in the RE. However, usually the available tracking space in LE is significantly smaller than the RE that the user wants to explore, and furthermore, local and remote environments typically have dissimilar layouts.

For computer-generated VEs, RDW has been successfully used to guide users. Hence, RDW seems to be a very suitable approach to solve this problem also in the context of a 360° VR-based telepresence system. However, while in VEs, RDW is based on manipulating movement of the virtual camera, such approaches cannot be directly applied to manipulations of a real camera due to latency issues, mechanical constraints, or limitations in the precision and accuracy of robot control. Figure 4.1 illustrates a general concept of using RDW on a 360° VR-based telepresence system. It is supposed that both the tracking system in the LE and the coordinate system in the RE are calibrated and registered. When users wearing an HMD perform movements in the LE, their position change can be detected by a tracking system in real time. Such a change in position can be measured by the vector $T_{real} = P_{cur} - P_{pre}$, where P_{cur} and P_{pre} mean the current position and the previous position. Normally, T_{real} is mapped to the RE by means of a one-to-one mapping when a movement is tracked in the LE. With respect to the registration between the RE and the tracking coordinate system, the physical camera (attached to the robot platform) is moved by T_{real} distance in the corresponding direction in REs. One essential advantage of using a 360° video stream for a telepresence system is that rotations of the user's head can be mapped one-to-one without the requirement that the robot needs to rotate. This is due to the fact that the 360° video already provides the spherical view of the RE.

In a computer-generated VE, the tracking system will update multiple times every second (e. g., with 90Hz), and the VE is rendered accordingly. However, due to the constraints and latency caused by network transmission of 360° video streams and the robot platform, which needs to move, such constant real-time updates are not possible in telepresence setups. Instead, the current video data from the camera capturing the RE is transmitted and displayed with a certain delay to the HMD. However, the user can change the orientation and position of the virtual camera inside the spherical projection with an update rate of 90Hz, but needs to wait for the latest display from the RE until the robot platform has moved and re-sent an updated image again.

A prototype of 360° VR-based telepresence system is implemented based on the hardware and concept mentioned above. The prototype is shown in Figure 4.1a. The experiments described in Section 4.6 and Section 4.7 are both based on this prototype and exploit the

360° videos captured with it. However, currently the prototype is not suitable for a real-time use yet due to the latency of movement control and image update. But we assume that future telepresence setups will allow lower latency communication similar to what we have today in purely computer-generated VEs.



(a) (b) (c)

Fig. 4.1 Illustration of the concept of 360° VR-based telepresence system with RDW: (a) the mobile platform is equipped with a 360° video camera moving in the remote environment (RE). (b) the user wears a VR head-mounted display (HMD) walking in the tracking space of the local environment (LE). (c) the user's view of the RE on the HMD.

4.4 RDW Gains in 360° Virtual Environment

As described in Section 4.2, Steinicke et al. [288] introduced translation and rotation gains for computer-generated VEs. In this section, the usage of translation and rotation gains in the case of a 360° video-based VE is explained. Furthermore, how the application of such gains can influence user movements is also described.

4.4.1 Translation Gains

The camera motions, which are used to render the view to the RE, are referred to as *virtual translations* and *virtual rotations*. The mapping between real and virtual motions can be implemented as follows: a translation gain is defined as the quotient of the corresponding virtual translation $T_{virtual}$ and the tracked real physical translation T_{real}, i. e., $g_T = \frac{T_{virtual}}{T_{real}}$. When a translation gain g_T is applied to a real-world movement T_{real}, the virtual camera is moved by $g_T \cdot T_{real}$ in the corresponding direction in the VE. This approach is useful in many situations, especially, when the user needs to explore a RE, which is much smaller or larger than the size of the tracking space in the LE. For example, for exploring a molecular structure with a nano-scale robot by means of natural walking, the movements in the real world have to be compressed a lot with a $g_T \approx 0$, whereas the exploration of a larger area on a remote planet with a robotic vehicle by means of natural walking may need a translation gain like $g_T \approx 50$.

Translation gains can be also denoted as $g_T = \frac{v_{virtual}}{v_{real}}$, where v_{real} means the speed of physical movement in the LE and $v_{virtual}$ means the speed of virtual movements showing in the RE. In addition, the position changes in the real world can be actually performed in three orientations at the same time [289], which includes fore-aft direction, lateral and vertical motions. In the following experiments, only the translation gains in the direction of the actual walking direction is focused, which means that only the movements in fore-aft direction are tracked, whereas the movements in lateral and vertical directions are filtered [133].

4.4.2 Rotation Gains

In a similar way, a rotation gain can be defined as the quotient of the mapped rotation in a VE and the real rotation in the tracked space: $g_R = \frac{R_{virtual}}{R_{real}}$, where $R_{virtual}$ is the virtual rotation and R_{real} represents the physical rotation in the real world. When a rotation gain g_R is applied to a real rotation R_{real} in the LE, the user will perceive a resulting virtual rotation of the RE given by $g_R \cdot R_{real}$. That means, when $g_R = 1$ is applied, the rendered view to the RE remains static during the user's head orientation since the 360° RE already provides the spherical view. However, if $g_R > 1$, the displayed 360° RE that the user perceives on the HMD will rotate against the direction of user's head rotation and, therefore, the virtual rotation of the RE will appear faster than normal. In the opposite case $g_R < 1$, the view to the RE rotates with the direction of user's head rotation, and will appear more slowly. For example, when a user rotates her head in the LE by 90°, a gain of $g_R = 1$ will be applied in a one-to-one mapping to the virtual camera, which makes the virtual camera also rotate 90° in the corresponding orientation. For $g_R = 0.5$, the user rotates 90° in the real world while she views only a 45° orientation change in the VE displayed on the HMD. Correspondingly, for the gain $g_R = 2$, a physical rotation of 90° in the real world is mapped to a rotation by 180° in the VE.

Again, rotations can be performed in three orientations at the same time in the real world, i. e., yaw, pitch and roll. However, in the following experiments, only the rotation gain for yaw rotation is focused, since yaw manipulations are used most often in RDW as it allows to steer users towards the desired directions, for instance, in order to prevent a collision in the LE [290, 231, 162, 140].

4.4.3 Other Gains

In principle, all other gains introduced for RDW such as curvature gains [288] or bending gains [169] are possible with 360° video-based VE as well. Nevertheless, the focus of the work in this chapter is on evaluating the user's sensitivity to detecting thresholds for rotation and translation gains, those gains will not be discussed in more detail.

4.5 Experiment Fundamentals

In this section, two psychophysical experiments are described in which the detection thresholds for translation and rotation gains in 360° video-based VE are analyzed. Since both experiments used similar material and methods, the setup and procedure will be described firstly, and then each experiment will be explained in detail.

4.5.1 Hardware Setup

The experiment was performed in a $12m \times 6m$ laboratory room (see Figure 4.1b). During the experiment, all participants are required to wear an HTC Vive HMD, which displays the 360° video-based RE with a resolution of 1080×1200 pixels per eye. The diagonal field of view is approximately 110° and the refresh rate is 90Hz. For tracking the user's position, a pair of lighthouse tracking stations delivered with the HTC Vive HMD are adopted. The lighthouse tracking system was calibrated in such a way that the system provides a walking space of $6m \times 4m$. During the experiments, the lab space was kept dark and quiet in order to reduce interference with the real world. Experimental instructions were shown to the participants by means of slides displayed on the HMD only. Participants used an HTC Vive controller as the input device to perform the operations described below and answer questions after each trial. For rendering the RE and controlling the system, an Intel computer was used, which had a 3.5GHz Core i7 processor, 32GB of main memory, and two NVIDIA Geforce GTX 980 graphics cards. Furthermore, participants answered questionnaires on an iMac computer. The 360° experimental video-based RE was recorded by the RICOH THETA S camera, which was attached upon the robot platform (see Figure 4.1a). It has a still image resolution up to 5376×2688 pixels and a live streaming resolution up to 1920×1080 pixels. The rendering work for 360° video-based RE was conducted in Unity3D Engine 5.6. The HMD was connected with the link box using a HTC Vive 3-in-1 (HDMI, USB and Power) 5m cable, in such a way that participants could move freely within the tracking space during the experiment. Considering the constraints and latency caused by network transmission of video streams, the 360° video of RE for the experiments was recorded with a 1280×720

resolution and a 15fps frame rate, which is consistent with the prototype of 360° VR-based telepresence system as introduced in Chapter 3.

4.5.2 Two-Alternative Forced-Choice Task

In order to identify the amount of deviations between physical movements in the LE and the virtual movements as shown from the RE, which are unnoticeable to users, a standard psychophysical procedure was adopted based on the method of constant stimuli in a two-alternative forced-choice (2AFC) task. In this method, the applied gains are presented randomly and uniformly distributed instead of appearing in a specific order [288, 169].

After each trial, participants have to choose one of two possible alternatives such as in this case *"smaller"* and *"larger"*. Even though, in several situations, it is difficult to correctly identify the answer, participants would need to choose the answer randomly, and will be correct in 50% on average. The *point of subjective equality* (PSE) is defined as the gain for which the participants answer "smaller" in 50% of the trials. At the PSE, participants perceive the translation or rotation in the RE and in the LE as identical. When the gain decreases or increases from the PSE, it becomes easier to detect the discrepancy between movements in the RE and in the LE. Typically, this results in a psychometric curve. When the answers reach a chance level of 100% or 0%, it is obvious and easy for the participants to detect the manipulations. A threshold can be described as the gain at which participants can just sense the difference between physical motions in the LE and virtual motion displayed on the HMD. However, stimuli at values close to thresholds could be often perceptible. Hence, thresholds are determined by a series of gains where the participants can only sense the manipulations with some probability. Typically for psychophysical experiments, the point where the psychometric curve reaches the middle between the 0% chance level and 100% is regarded as a detection threshold (DT). Thus, the lower DT for gains smaller than the PSE value is defined as the gain where the participants answered in 75% of all trials with "smaller" on average. Similarly, the upper DT for gains larger than the PSE value is the gain where participants have just answered in 25% of all trials "smaller" on average.

In this chapter, the range of gains for which users are not able to reliably detect the discrepancy as well as the gain at which users perceive motions in the LE and in the RE as equal is analyzed. The 25% to 75% DTs shows a gain interval of potential manipulations, which can be applied for RDW in 360° video-based REs. Moreover, the PSE values indicate how to map the user motions in the LE to the movements of the telepresence robot in the RE, such that the visual information displayed on the HMD appears naturally to the users.

4.6 Experiment 1: Detection Thresholds for Translation Gains

In this experiment, the participant's ability to discriminate whether a physical translation in the LE was slower or faster than the virtual translation displayed in the 360° video-based RE was investigated. The participants were instructed to walk a fixed distance in the LE and mapped their movements to a pre-recorded 360° video-based RE.

4.6.1 Methods

A 360° video was pre-recorded with the prototype of the 360° VR-based telepresence system described in Section 3 showing a forward movement in the RE with a natural walking speed of 1.4m/s [292]. The height displayed in the 360° video was recorded at a height of 1.75m.[1] The playing speed of the video was manipulated based on the user walking speed measured in the LE by applying the described translation gains, in such a way that the forward movement speed in the displayed RE was manipulated accordingly. This means that when the user walked with a speed of 1.4m/s in the LE, the 360° video of the RE was displayed in normal speed, whereas when the user decreased the speed and stopped, the video was slowed down with the gains until it was paused. The 360° video of the RE showed a movement in the fore-aft direction in the RE, and all other micro head movements were implemented as micro motions of the virtual camera inside a 360° video-based spherical space. Changes of the head orientation were implemented using a one-to-one mapping.

Figure 4.2 illustrates the setup for Experiment 1. Before starting the experiment, participants were guided to the start line and held an HTC Vive controller. When participants were ready, they clicked the trigger button on the controller to enable the 360° video, which presented the RE on the HMD, and started to walk in fore-aft direction in the LE. The play speed of the 360° video during walking was adjusted to the participant's physical speed in real-time. For instance, if the participants stopped, the scene of the RE displayed on the HMD would also pause. The walking speed was determined by movements along the main direction of the corridor shown in the 360° video (Figure 4.3). During the experiment, different translation gains were used to control the play speed of the 360° video. For example, when walking with the translation gain g_T, the 360° video would be played in the speed of $g_T \cdot v_{real}$, where v_{real} is the participant's real-time physical speed along the fore-aft direction in the LE. When participants traveled 5m in the LE and crossed the end line, the 360° RE displayed on the

[1] We could not find any significant effect of the deviation from the user's actual eye height and the recorded height on the estimation of the detection thresholds.

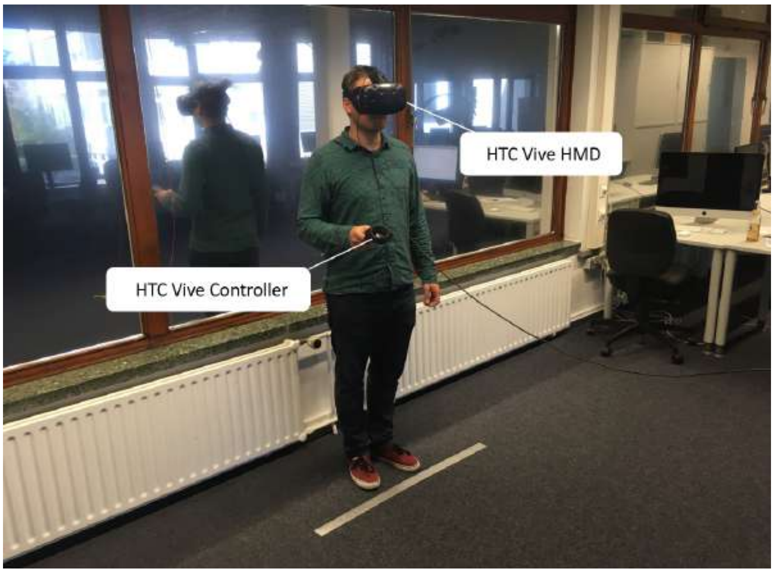


Fig. 4.2 Illustration of the experimental setup: A user is walking straightforward in the LE to interact with the 360° video-based RE. Translation gains are applied to change the speed of displayed virtual movement on the HMD.

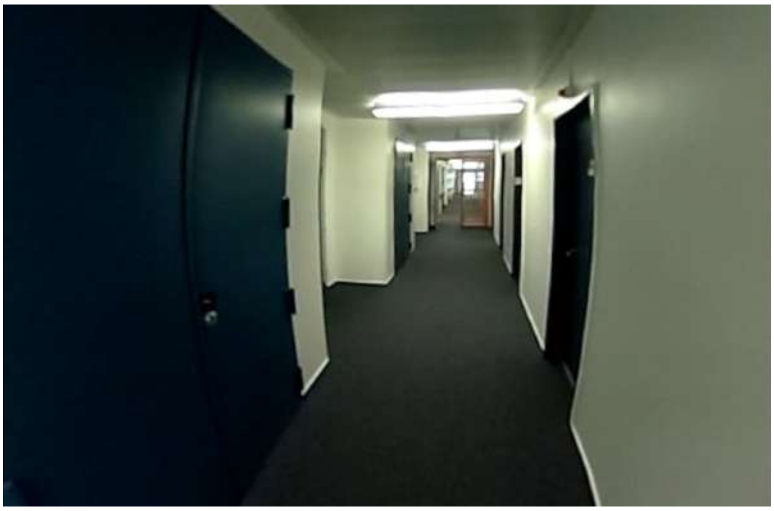


Fig. 4.3 The user's view to the 360° video-based environment, which shows a corridor from the RE (image captured from HTC Vive HMD).

HMD would automatically disappear. Then, participants had to estimate whether the virtually displayed motion was faster or slower than the physical translation in the LE (in terms of distance, this corresponds to longer or shorter). Participants had to provide their answers by using the touch pad on the HTC Vive controller. After that, participants walked back to the start line, while they were guided by visual markers displayed on the HMD, and then clicked the trigger button again to start the next trial. For each participant, 9 different gains were tested in the range of $\{0.6, 1.4\}$ in steps of 0.1 and each gain was repeated for 6 times. Hence, in total, each participant performed 54 trials in which they walked a 5m distance in the LE, while they viewed virtual distances within a range of $\{3m, 7m\}$ for each trial. All of the trials appeared in randomized order. After each trial, participants returned to the start line with the help of the markers displayed on the HMD, and clicked the trigger button again to continue with the next trial.

4.6.2 Participants

16 participants (14 male and 2 female, age 19-37, M=26.4) participated in E1, in which the participant's sensitivity to translation gains was explored. 1 participant could not complete the experiment because of cyber sickness. All data from the remaining participants was included in the analyses. Most of the participants were members or students from the local department of computer science. All of them had normal or corrected to normal vision. Five of them took part in the experiment with glasses. None of the participants suffered from a disorder of equilibrium. Four of the participants reported dyschromatopsia, strong eye dominance, astigmatism and night blindness separately. There are no other vision disorders reported by the participants. The experience of the participants with 3D stereoscopic displays (such as cinema or games) was M = 2.4 within the range of 1 (no experience) to 5 (much experience). 14 participants have worn HMDs before. Most of the participants had experiences with 3D computer games (M = 3.2, with 1 corresponds to no and 5 to much experience). On average, they played 4.4 hours per week. The body heights of the participants varied between 1.60m - 1.90m (M = 1.80m).

The experimental process for each participant included pre- and post-online-questionnaires, instructions, training trials, experiment, and breaks, the total time for each participant was about 40 - 50 minutes. The participants needed to wear the HMD for around 25-30 minutes. During the experiment, the participants were allowed to take breaks at any time.

4.6.3 Results

Figure 4.4 shows the mean probability over all participants that they estimate the virtual straightforward movement shown on the HMD as faster than the physical motion in the LE for different translation gains. The error bars show the standard errors. Translation gains g_T lead to faster virtual straightforward movements (relative to the physical movements) if $g_T > 1$. Then, participants would feel that they move a larger distance in the RE than in the LE. A gain of $g_T < 1$ results in a virtual translation movement, which is slower

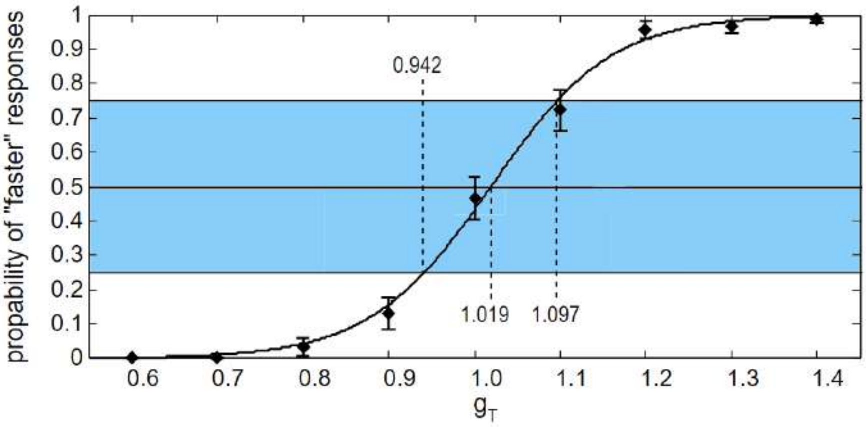


Fig. 4.4 Pooled results of the discrimination between movements displayed from the RE and movements performed in the LE. The x axis shows the applied translation gain g_T, the y axis shows the probability that participants estimated the virtual straightforward movement displayed as 360° video faster than the actually performed physical motion.

than the physical walking speed, resulting in a shorter distance displayed from the RE. A psychometric function was fitted in the form $f(x) = \frac{1}{1+e^{a \cdot x+b}}$ with real numbers a and b.

From the psychometric function, a slight bias for the PSE was determined at $PSE = 1.019$. In order to compare the found bias with the gain of 1.0, a one sample t-test was performed, which did not show any significant difference (t=1.271, df=14).

The results for the participant's sensitivity to translation gains show that gains from 0.942 to 1.097 (25% and 75% DT) cannot be reliably detected. This means that with manipulations in this range, participants were not able to reliably discriminate whether a physical translation in the LE was slower or faster than the corresponding virtual translation perceived from the 360° video-based RE.

4.6.4 Discussion

The results show that participants could not discriminate the difference between physical translation performed in the LE and virtual translation perceived from the RE, when the movement is manipulated with a gain in a range from 5.8% slower to 9.7% faster than the real movement. From the definition of translation gains, a $PSE = 1.019$ indicates that the virtual translations displayed from the 360° video-based RE are slightly faster than the physical translations in the LE [84, 133, 134, 181]. A translation gain $g_T = 1.019$ appeared natural to the participants, which means that walking a distance of 4.91m in the LE felt like traveling 5m in the RE. Therefore, participants tended to travel a shorter physical distance in the LE when they tried to approach the same expected virtual distance in the 360° video-based REs.

In addition, a PSE larger than 1 is consistent with the results from previous research on translations in fully computer-generated VEs [287, 288]. However, the bias reproduced in this experiment was not statistically significant.

According to the results from previous research for computer-generated VEs [287], a slight shift of the psychometric function and detection thresholds towards the larger gains are expected also for 360° video-based REs. Visual analysis from Figure 4.4 shows such a slight shift, and both the 25% and 75% detection thresholds are slightly shifted towards the larger gains. However, this shift is smaller than the ones reported in previous work. Furthermore, an interesting observation is that the 25% and 75% detection thresholds for the translation gains are both closer to the *PSE* value in the 360° video-based REs compared to the results from previous research in fully computer-generated VEs [287, 288]. This might be due to the fact that participants perceive the RE displayed on the HMD as more accurate and realistic, and therefore, are better in estimating movements.

4.7 Experiment 2: Detection Thresholds for Rotation Gains

4.7.1 Methods

In this experiment, the ability of participants to discriminate between virtual rotations displayed from the RE and physical rotations performed in the LE was analyzed. Figure 4.5 shows the setup for the experiment. During the experiment, the participants wore an HMD and were placed in the center of the tracking space. The were instructed to perform rotations in the LE, which were tracked and displayed as virtual rotations in the 360° video-based RE.

A 360° panoramic image was adopted to create a spherical projection space in Unity3D, which presented a 360° outdoor RE. Rotation gains were applied to the yaw rotation only. Again, the view height in the RE was adjusted to 1.75m. At the beginning of the experiment, participants were instructed to stand in the center of the tracking space and hold an HTC Vive controller. When everything is ready, participants could start the new trial by clicking the trigger button on the controller. Then, they could see the 360° RE (Figure 4.5) to which a randomized rotation gain was applied, when they started to turn. Participants saw a green ball in front of their view at the eye-level that marked the start point for the rotation. An arrow showed the rotation direction that participants were required to follow. The participants were told to rotate in the corresponding direction until a red ball appeared in the front of their view, which indicated the end point of the rotation. The angle between the start point (green ball) and the end point (red ball) was adjusted to 90°. Hence, the virtual rotation shown on the HMD from the RE was always 90°, but the physical rotation participants performed in the LE was different according to the corresponding rotation gains. During the experiment, different rotation gains were applied to the virtual rotations showing the RE. A rotation gain $g_R = 1$ shows a one-to-one mapping between physical rotation in the LE and virtual rotation displayed from the RE. However, for example, when a rotation gain satisfies $g_R < 1$, the virtual scene on the HMD rotates with the direction of the real physical rotation in the LE and slowed down the change in the RE. In the opposite case, for a rotation gain $g_R > 1$, the scene in the RE rotates against the direction of the real physical rotation in the LE, and accelerated the change of view in the 360° RE.

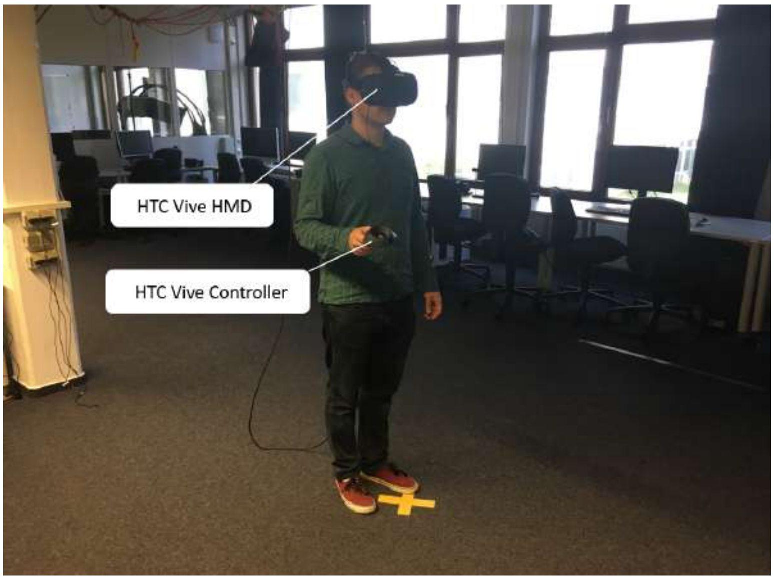


Fig. 4.5 Illustration of the experimental setup: a user is performing rotations in the LE to interact with the 360° video-based RE. Rotation gains are applied in the experiment to change the speed of virtual rotations displayed from the RE.

For each participant, 9 different gains were tested in a range of $\{0.6, 1.4\}$ by steps of 0.1 and each gain was repeated for 6 times. Hence, each participant performed a series of physical rotations in the LE with a range of $\{64.29^\circ, 150^\circ\}$ to achieve a 90° virtual rotation in the 360° RE. In order to study the effects of different rotation directions, rotations to the left and to the right were considered. Therefore, in total, there were 108 trials for each participant. All of the trials appeared in randomized order. Then, participants had to choose whether the perceived rotation from the RE was smaller or larger than the physical rotation performed in the LE. Again, responses had to be given via the touch pad of HTC Vive controller. After each trial, the participant turned back to the start orientation with the help of the markers displayed on the HMD, and clicked the trigger button again to continue with the next trial.

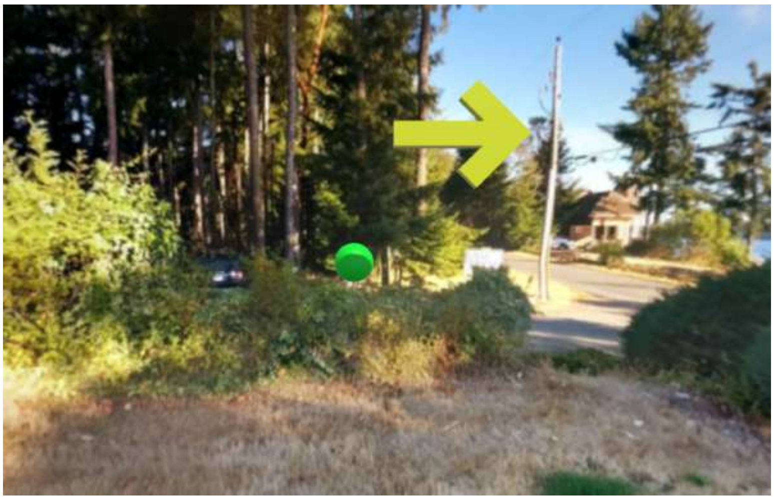


Fig. 4.6 The user's view to the 360° RE, in which a start point and a directional arrow are displayed (image captured through HTC Vive HMD).

4.7.2 Participants

17 participants (13 male and 4 female, age 24 - 38, M=29.5) took part in this experiment analyzing the sensitivity to rotation gains. Two participants stopped the experiment because of suffering from motion sickness. The data from the remaining 15 participants were included in the analyses.

Most of the participants were students or members from the local department of computer science. All participants had normal or corrected to normal vision. 3 participants wore glasses during the experiment, and 1 participant wore contact lenses. 1 participant reported to suffer from a disorder of equilibrium. 2 participants reported strong eye dominance and night blindness. No other vision disorders have been reported by the participants. Most of the participants had experiences with 3D stereoscopic displays before with M = 2.76 in a range of 1 (no experience) to 5 (much experience). 14 participants had experiences using HMDs before, and 13 of them had experiences with 3D computer games with M = 2.71, and played games with an average time of 5.26 hours per week. The body height of the participants was in a range of 1.60m - 1.92m with M = 1.75m.

The total time of the experimental procedure for each participant including pre-online-questionnaires, instructions, a few training trials, experiment, breaks and post-online-questionnaires, took almost 40 - 50 minutes. The participants wore the HMD for about 25 - 30 minutes. During the experiment, the participants were allowed to take breaks at any times.

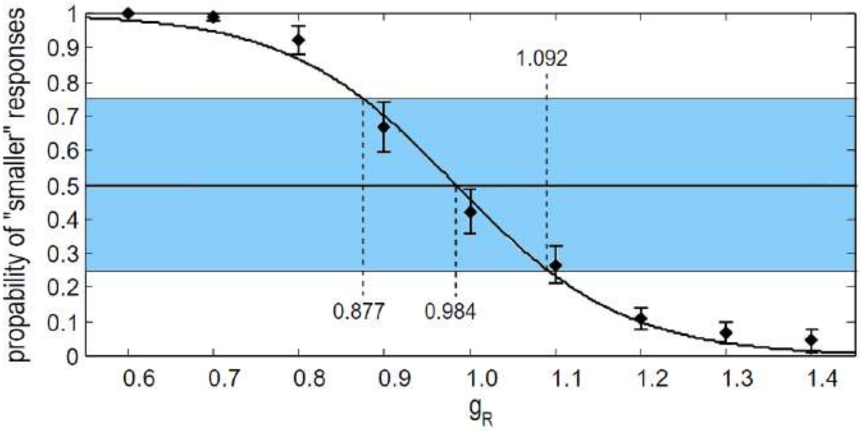


Fig. 4.7 Pooled results of the discrimination between remote virtual and local physical rotations towards left. The x axis shows the applied rotation gain g_R, the y axis shows the probability that participants estimated the virtual rotation in the RE as smaller than the physical rotation in the LE.

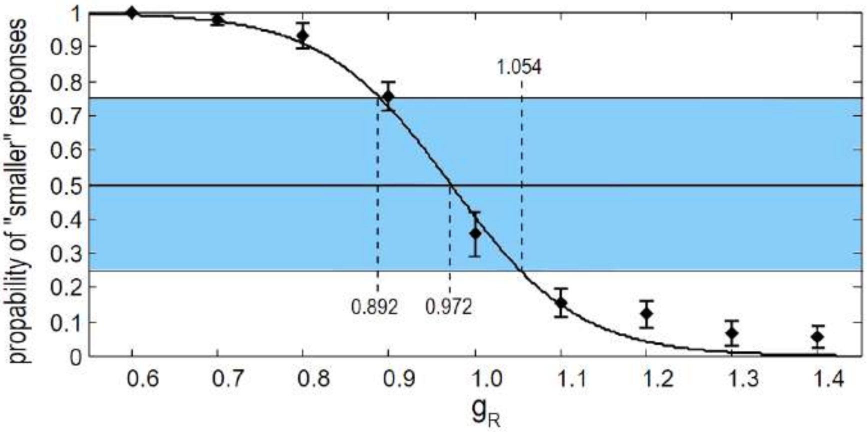


Fig. 4.8 Pooled results of the discrimination between remote virtual and local physical rotations towards right. The x axis shows the applied rotation gain g_R, the y axis shows the probability that participants estimated the virtual rotation in the RE as smaller than the physical rotation in the LE.

4.7.3 Results

To verify the influence of different rotation orientations, the data of rotations to the left (cf. Figure 4.7) and rotations to the right (cf. Figure 4.8) was analyzed separately. In the experiment, a rotation gain g_R results in a smaller physical rotation than the virtual rotation if $g_R > 1$. This means that participants rotate less in the LE than in the RE. A rotation gain leads to a larger physical rotation than the virtual rotation if $g_R < 1$. In other words, in this case participants would rotate more in the LE compared to the rotation they view in the RE. The data was fitted with the same psychometric function as in Experiment E1.

Figure 4.7 presents the mean probability over all participants that they estimated the virtual rotations to the left in the RE smaller than the physical rotations in the LE with different applied rotation gains. The error bars show the standard errors. The psychometric function determined a bias for the PSE at $PSE = 0.984$. The 25% and 75% detection thresholds for rotation gains were found at 0.877 and 1.092. Within this range of rotation gains, participants were not able to reliably discriminate whether a physical rotation to the left in the LE was smaller or larger than the corresponding virtual rotation displayed from the 360° RE. Figure 4.8 presents the situation in which rotations were performed to the right. The PSE value was derived at $PSE = 0.972$, and the rotation gains between the detection thresholds of 25% and 75% were from 0.892 to 1.054.

In order to compare the found bias from the gain of 1.0, a one sample t-test was performed, which did not show any significant difference for rotations to the left (t=-0.429, df=14) or rotations to the right (t=-1.466, df=14). Furthermore, there was also no significant differences between rotations to the left and rotations to the right (t=0.472, df=14).

4.7.4 Discussion

For a physical rotation to the left, participants could not discriminate the difference between physical rotations in the LE and perceived virtual rotations from the 360° video-based RE when rotation gains were within a range of $\{0.877, 1.092\}$. This means that the virtual rotation in the RE is 12.3% less and 9.2% more than the physical rotation in the LE. A rotation gain $g_R = 0.984$ appeared most natural, indicating that participants have to rotate for 91.46° in the LE to perceive the illusion that they already rotated by 90° in the RE.

For a physical rotation to the right, the range of rotation gains that participants could not reliably detect as manipulation between physical rotations in the LE and virtual rotations in the 360° RE is $\{0.892, 1.054\}$. In other words, for a virtual rotation in the RE, participants could accept a 10.8% smaller or 5.4% larger physical rotation in the LE without noticing the discrimination. The most natural rotation gain for rotations to the right is $g_R = 0.972$, which indicates that participants need to rotate for 92.59° in the LE to feel that they rotate 90° in the RE.

As described above, independent from the direction of rotations, the most natural rotation gains for the participants are slightly smaller than 1, which suggests that the participants need to rotate more in the LE to perceive the illusion that they have already rotated the same expected angle in the 360° video-based REs. However, this bias was not statistically significant. These results show an opposite effect to the results from the translation experiment, but a PSE smaller than 1 appears to be consistent with the results from previous research on the rotations in fully computer-generated VEs [140, 76, 288]. Moreover, the results in this study indicate that the range of gains, which can be applied to 360° video-based REs and be unnoticeable to the participants, are narrower than the results reported in earlier work for purely VEs. Hence, the results suggest again that participants have a better discrimination ability for manipulations of rotations in a 360° video-based RE compared to rotations in a purely computer-generated VE. This point will be discussed further in the general discussion.

Furthermore, the results indicate that the interval of detection thresholds for manipulations of rotations to the right is smaller than the manipulations of rotations to the left in the 360° video-based RE. This means that participants have provided more accurate estimations for rotation to the right than to the left. Such a finding has not been reported in earlier work. One possible explanation of the observed phenomenon might be related to the structure of the brain and hand dominance; since most of the participants were right-handed, however, this has to be verified in further research.

In summary, there is a range of rotation gains, in which participants could not reliably discriminate between physical rotations in the LE and virtual rotations in the 360° video-based REs.

4.8 Post-Questionnaires

After the experiments, participants answered further questionnaires in order to identify potential drawbacks of the experimental design. Participants estimated whether they feel that the 360° RE surrounded them (0 corresponds to fully disagree, 7 corresponds to fully agree). For the translation experiment E1 the mean value was 4.4 (SD = 1.76), and for the rotation experiment E2 the average value was 5.2 (SD = 1.29). Hence, most of the participants agree that when using the 360° VR-based telepresence system, they perceived a high sense of presence. Furthermore, participants were asked how confident they were that they chose the correct answer (0 corresponds to very low, 4 corresponds to very high). The average value for answers to this questions was 2.53 (SD = 0.83) for the translation experiment E1 and 2.29 (SD = 1.06) for the rotation experiment E2. Even though the design of the experiment based on a 2AFCT probably has some influence on the accuracy of final results, which makes the participants have to guess when facing some situations that really hard to judge, but from the feedback it could be seen that most of the participants are sure that they have given the correct answers for most of the situations. So the results derived from the experiments in this chapter is reliable.

After both experiments, simulator sickness were also measured by means of Kennedy's Simulator Sickness Questionnaire (SSQ). For the translation experiment, the average Pre-SSQ score for all participants was 6.23 (SD = 9.34) before the experiment, and an average Post-SSQ score was 26.68 (SD = 27.80) after the experiment. For the rotation experiment, the average Pre-SSQ score for all participants was 9.23 (SD = 21.06) before the experiment, and the average Post-SSQ score was 55.60 (SD = 68.03) after the experiment. The results show that the average Post-SSQ score after the rotation experiment was larger than after the translation experiment. This finding can be explained by the sensory-conflict theory, since continuous rotations provide more vestibular cues than constant straightforward motions. Hence, manipulations during such rotations induce more sensory conflicts [170].

4.9 General Discussion

The results of the translation experiment show that participants cannot distinguish discrepancies between physical translations in the LE and perceived virtual translations in the 360° video-based RE when the virtual translation is down-scaled by 5.8% and up-scaled by 9.7%. A small bias for the PSE was determined in $PSE = 1.019$ indicating that slightly up-scaled virtual translations in the RE appear most natural to the users, which means that users believe they have already walked a 5m distance in the 360° video-based RE after only walking a 4.91m distance in the LE. These results are consistent with most previous findings in the fully computer-generated VEs [84, 133, 134, 181]. However, the strong asymmetric characteristic of the psychometric function, which was found in previous research on RDW in VEs could not be replicated in this experiment in which realistic 360° video-based environments were used.

The rotation experiment results show that when virtual rotations in the 360° RE are applied within a range of 12.3% less or 9.2% more than the corresponding physical rotation in the LE, the users cannot reliably detect the difference between them. For rotations to the left, a rotation gain of $PSE = 0.984$ appears most natural to the participants, meaning that they have to rotate 91.46° in the LE to have the illusion that they have already rotated 90° in the RE. The most natural rotation gain for the rotation to the right is $PSE = 0.972$, which means that participants need to rotate 92.59° in the LE to have the impression that they have already rotated 90° in the RE. These results also confirm previous findings to some extent [140, 76, 288]. Again, the asymmetric characteristic of the psychometric function is not so obvious for this experiment results in 360° video-based REs compared to previous findings for fully computer-generated VEs.

Before the psychophysical experiments were performed, a hypothesis has been proposed that the detection thresholds for translation and rotation gains in 360° video-based REs would be different to previous results, found in fully computer-generated VEs. There are some essential differences between showing a VE which is computer-generated in contrast to a RE which is captured with a 360° video camera. Compared with fully computer-generated VEs, 360° videos lack some depth cues like stereoscopic disparity and motion parallax. Especially, in a fully computer-generated VE, the scenes are consisting of actual 3D objects, while in a 360° video-based RE all the environmental information is framed within a spherical texture which may cause some spatial cognition disorders.

The data as well as the analysis presented in Section 4.6 and 4.7 suggest that manipulations in 360° video-based REs have similar influence on users as manipulations in fully computer-generated VEs [288], i. e., users tend to travel a slightly shorter distance but rotate a slightly larger angle in the LE when they try to approach the same expected motion in the 360° video-based REs. However, some differences in PSE values and distribution of detection thresholds between 360° video-based REs and computer-generated VEs should be also noted. On the one hand, the PSE value for translations in 360° video-based REs is 1.019 which is much closer to a one-to-one mapping compared to previous results in VEs. The same situation can be also found from the results of the rotation experiment with a PSE value of 0.984 to the left and 0.972 to the right in 360° video-based REs and 0.9594 in fully computer-generated VEs. Conversely, the ranges between 25% and 75% detection thresholds for translations and rotations in 360° video-based REs are both smaller than the results in VEs, which indicates a smaller range for users in which they are not able to reliably discriminate the difference between the motions in a 360° video-based RE and in the real world. All these differences suggest that users have a more accurate ability to judge the difference between physical motions in the LE with corresponding virtual motions in 360° video-based REs than in fully computer-generated VEs. However, future work is required to explore these differences in more depth.

There are a few possible explanations for these findings. Firstly, the scenes shown to the participants during the experiments are 360° videos of REs in the real world rather than computer-generated VEs. What the telepresence system implemented is to break the limitation of distance by transmitting the video information from the RE to the LE and displaying it to the user instead of creating a totally new space fully generated by the computers. Therefore, a 360° video-based RE, which is displayed in the LE might appear more realistically and accurately with respect to size, texture or other characteristics in contrast to a fully computer-generated VE. Furthermore, since the objects in 360° video-based REs are projected as spherical textures on the HMD, there was a lack of stereoscopic disparity and motion parallax in this case. Hence, users might have perceived translations and rotations based on the whole environments rather than on one or two specific objects in the REs. For this reason, a lack of stereoscopic disparity and motion parallax has little influence on the users concerning distance and angle perception in 360° video-based REs.

Moreover, the resolution of the virtual scenes presented on the HMD can also lead to different results. Usually, human perceive the angle of a physical rotation via two sensory channels: visual information from the environment as well as the proprioception and vestibular information. When people perform a rotation in a VE displayed with high resolution, their sensory system will likely weight the visual information as more reliable in the multi-sensory integration process compared to scenarios with low visual resolution [78, 147]. Cues from vision will take a leading role in the perception of rotation angle, which has been shown in the literature [78, 147]. In contrast, when the scenes displayed on the HMD have low resolutions, the reliability of vision will be reduced to a certain extent, whereas the vestibular and proprioceptive cues might be evaluated with a higher weight in the multi-sensory integration process [78, 147]. The extreme situation occurs when participants close the eyes. In this case, no visual information from the environment provide cues about rotation angles, and users could only perceive information about the rotation using the vestibular system and proprioception. Hence, it is reasonable that the resolution of scenes on the HMD can also be a possible explanation that caused different results between a fully computer-generated VE and a 360° video-based RE.

Similar to RDW in computer-generated VEs, in most applications a greater range of gains can also be accepted by users without noticing that they are manipulated. The detection thresholds of the experiments in this chapter are conservatively estimated, since for most actual telepresence systems, the users will not be able to easily recognize the discrimination between real motions and remote virtual motions, because they need to focus on other tasks in the RE like object selections, manipulations etc. Hence, the upper and lower detection thresholds for translation and rotation gains found in this experiments could serve as lower and upper bounds in the actual telepresence process.

4.10 Conclusion

In this chapter, the user's ability of recognizing RDW manipulations for translation and rotation was evaluated in two separate experiments. The results show that participants were not able to reliably discriminate the difference between physical motions in the LE and perceived virtual motions from the 360° RE when virtual translations are down-scaled by 5.8% and up-scaled by 9.7%, and virtual rotations are about 12.3% less or 9.2% more than the actual physical rotations. These findings provide interesting implications for future implementations of 360° telepresence systems in which users can explore REs by controlling a remote robotic surrogate by means of natural walking in the LE.

As described above, a telepresence system based on remote robotic systems and 360° video camera would introduce larger latency compared to typical VR environments. Furthermore, in the experiment, micro movements of the user's head was implemented by micro movements of the virtual camera inside the spherical space on which the 360° videos or images are projected as movie textures. While these movements were considerably small in the experiment, for real-world applications it might be required to implement larger movements, which will then get noticeable. Again, it is an interesting question how much deviation from the projection center of the spherical space can be reliably detected by users. In addition, it is also attractive to explore other VR setups in the LE. For example, a CAVE-like projection space has been tested to display the 360° REs instead of VR HMD, and it is interesting to see if the detection thresholds and the biases of the PSEs can be replicated in such setups. Moreover, different REs and application domains should be analyzed such as exploration of hallways, cooperation in business meeting rooms or inspections of outdoor scenarios to explore if the thresholds proposed in this chapter can be generalized to different scenarios.

Chapter 5

Natural Interacting with Avatar-Robot Telepresent Surrogate

In this chapter, the human's perception of social cues expressed by a mixed reality avatar displayed on top of a robotic surrogate and its corresponding animations are investigated. Therefore, we conducted two independent user studies: the first study analyzes the perception and localization of the user's gaze direction in RE, which is delivered via a visualized avatar attached upon the 360° video-based telepresence robot. The following research questions are explored through the user study:

- *Compared with a traditional 2D tablet display, will human subjects have a better performance on perceiving and localizing the user's gaze direction with a stereoscopically rendered AR avatar?*
- *Which factors will influence the perception and localization on the user's gaze direction in a 360° hosted telepresence?*

The second study focuses on a mixed reality avatar arm swing technique, which naturally communicates the velocity of a robotic surrogate it is attached to. Based on this, the proxemic preferences between robotic surrogate and human users in the dynamic scenarios of walking following or towards an avatar-robot surrogate are evaluated to address the research questions below:

- *Is the avatar arm swing technology effective for human subjects to perceive different moving speeds of robotic surrogate?*
- *With a mixed reality avatar displayed on top of the robotic surrogate, which factors will influence the proxemic preferences of human subjects during a dynamic interaction?*

5.1 Rapid Localization of Teleoperator Gaze in 360° Hosted Telepresence

5.1.1 Motivation

As explained in Chapter 1, traditional telepresence platforms consist of a display that depicts the face of the teleoperator so that bystanders feel an enhanced sense of teleoperator presence [225, 229, 326, 176]. In some cases, the teleoperator is accompanied by a host. e.g. A telepresent doctor is making a house call accompanied by a family member; a telepresent building site inspector or industrial safety inspector accompanied by the locally present foreman; an apartment viewing, where a real-estate agent is attending to show the property to a remotely telepresent client. All of these examples could be referred to as *hosted telepresence*. With a traditional telepresence platform, because the camera is mounted facing forward, when teleoperators wish to change their view direction, they rotate the platform manually. This change of heading by the platform is an important gaze cue for the bystanders. However, this means that in hosted telepresence there are occasions when the display is facing away from the host, and as such, the head or face of the guest[1] is not visible. As a result, this will cause a loss of visual cues (especially in vertical dimension) for the host to localize gaze directions of the guest (Figure 5.1a).

Moreover, the usage of single webcam on the telepresence platform for capturing the RE can only provides the guest with a very narrow field of view and a limited illusion of immersive display and spatial presence. For this reason, 360° VR-based telepresence systems are increasingly adopted for creating an immersive full-viewed user experience [221, 144, 115, 278]. When the guest uses a 360° VR-based telepresence system, they are normally required to wear a HMD to obtain an immersive virtual exploration [333], and the scene of RE on the HMD is projected and displayed on the inner-surface of a spherical space. Because of the HMD, it is technically complicated [224] to capture an accurate reconstruction of the teleoperator's head and face. In addition to this, since the teleoperator has a 360° video sphere surrounding them, the robotic platform in the RE does not need to rotate when he or she looks around thereby conserving battery and avoiding extra noise and collision. The platform would rotate only to move to a new heading. If the locomotion mechanism is holonomic/omnidirectional [239], the rotation of the platform is not necessary at all. In this case, the host has no visual cues from the platform for the gaze directions of the guest in both horizontal dimension and vertical dimension (Figure 5.1b).

[1] the teleoperator will be occasionally referred to as "guest" throughout the chapter depending on the context

Nevertheless, if the host can effectively and rapidly localize the gaze directions of the guest like in an aforementioned apartment viewing scenario, he or she could proactively offer additional information or description to make the communication more interactive. “This vase comes from 17^{th} century Qing dynasty...” or “... don’t worry about that damaged part on the wall, we will have it repaired before you move in”.

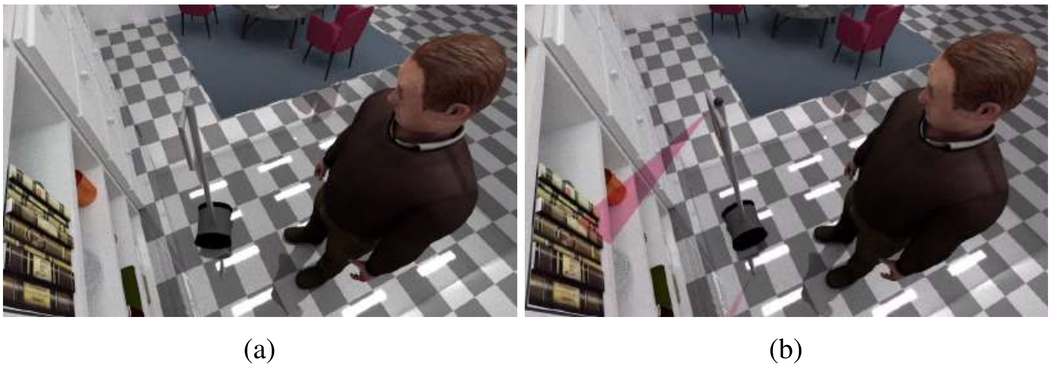
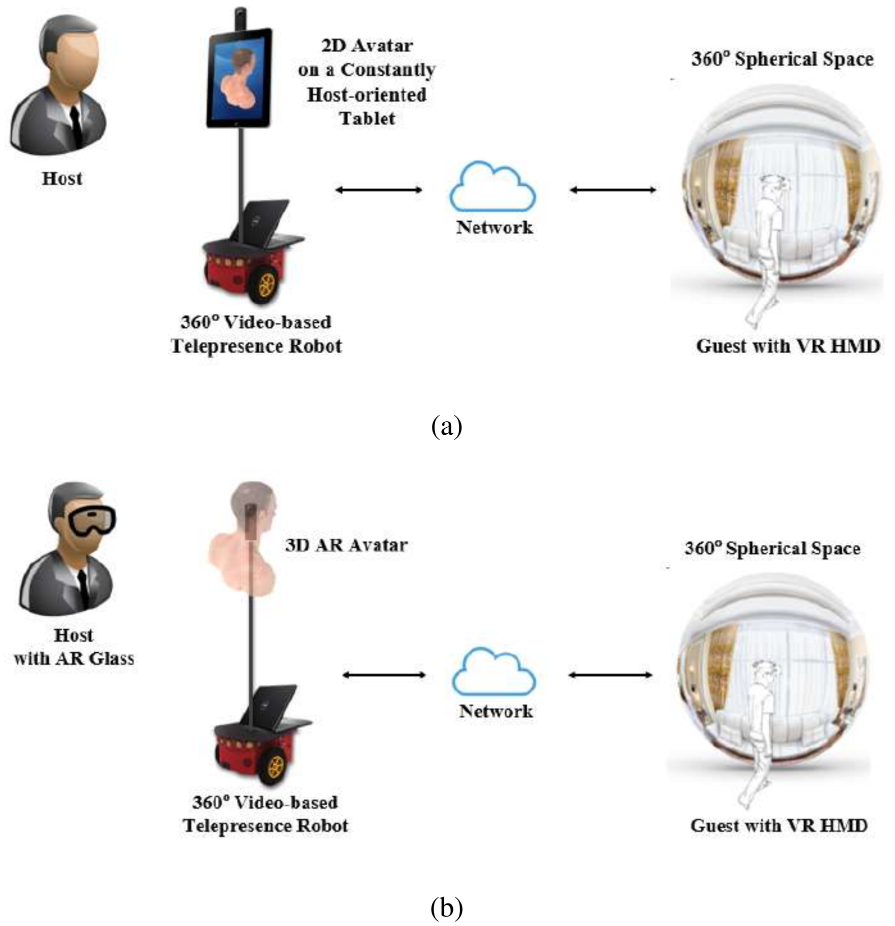


Fig. 5.1 Comparison of two telepresence systems in the perspective of bystanders: (a) Traditional telepresence robot: the host could localize a guest’s gaze directions by the rotation of platform, but the head or face of the guest is invisible in some situations. (b) 360° telepresence robot: in 360° video-based telepresence, the teleoperator (guest) has a 360° video sphere surrounding them on the HMD, so the robotic platform does not need to rotate when he or she looks around.

All of the aforementioned problems (i.e., display occasionally facing away from the host, teleoperator’s face obscured due to HMD, platform not rotating to give heading cues) impede gaze localization by the host, hinder his or her ability to offer proactive information and devalue the hosted telepresence experience for both parties.

Two potential solutions were proposed to the current problem: In the first solution, after the robot completes travel following the teleoperator’s command, the platform rotates so that the display faces the host. On the display, an avatar head is depicted to indicate the guest’s gaze direction (Figure 5.2a). The pose of the avatar is synchronized with the teleoperator’s HMD pose and is therefore always “looking” at the correct point. The second solution is to display a stereoscopically rendered avatar on top of the robotic platform (Figure 5.2b). In both cases it is assumed that the robot is equipped with sensors that can track the host and, in the latter case, that the host is wearing an AR glass to view the avatar.

The goal of the study in this chapter is to determine the performance improvement of these two potential solutions, if any, relative to a traditional telepresence robot. In order to test this, a formal user study is conducted in a simulated AR environment. Furthermore, it is hypothesized that in a rapid localizing case the second solution, a stereoscopically rendered avatar, will allow the host to localize the guest’s gaze directions in the RE with



(a)

(b)

Fig. 5.2 Two proposed solutions to the current limitation in this chapter: (a) Solution I: present a 2D avatar on a constantly host-oriented tablet and synchronize guest's gaze directions after the robot completes its travel movement. (b) Solution II: display a stereoscopically rendered 3D avatar on top of the robotic platform using AR technology to synchronize guest's gaze directions.

more accuracy. In addition, it is assumed that there will be a "sweet spot" for the distance that the telepresence robot must be positioned from the host to enable the highest localization accuracy of the host and that this sweet spot is different for the two solutions. The study results in this chapter will serve as baseline reference for augmented reality telepresence system designers.

5.1.2 Related Works

Gaze is an important cue for localizing another person's focus or interest in space and can help making a communication more active and effective [229, 326]. A number of researchers have already tried different solutions to support gaze awareness in video-based display communication. Gemmell et al. [90] developed a software approach for gaze awareness in video tele-conference, in which the head and eye movements of participants were tracked using computer vision techniques and then placed graphically in a 3D environment. Also, some works have tried to find solutions by creating novel hardware-based user interface. For example, Otsuki et al. [225] used a simulated eyeball display to create an interface for a video communication system to present gaze directions of remote user. Similarly, Misawa et al. [200] used a face-shaped screen to reflect the user's head gestures. They performed a series of studies on facial expressions, head gestures and perception of eye gaze based on it. Kawaguchi et al. [150] conducted a study on gaze perception of a face image presented on a flat display that could rotate. The results suggest that both the rotation of avatar in the display and the rotation of display device could influence people's estimation, and lead to an overestimation on the avatar gaze direction. Kawaguchi et al. [149] additionally investigated the effect of embodiment presentation on social telepresence and found that the embodiment could improve social presence, familiarity and directivity of a remote person. Moubayed et al. [203] presented a back-projected human-like robot head Furhat using state-of-the-art facial animation, and conducted a series of experiments on the measurement of perceptive accuracy of Furhat's gaze based on eye design, head movement and viewing angle. In addition, some works used a projector to highlight gaze from a remote user by using a spotlight-like method in GazeTorch [7] or a colorful arc [326]. Targeting perceived naturalness, Liu et al. [179] proposed a model for generating head tilting and nodding on humanoid robot. With this model, they verified that the proposed method could perform equally to directly mapping people's original motions with gaze information. All of these aforementioned methods bring something new to the teleoperation domain but can only depict teleoperator gaze in a limited range, with less than optimal social interactive cues.

Some researchers proposed to use virtual avatars to synchronize the guest's gaze directions, so as to help the host understanding and localizing the guest's focus more effectively. For example, Pan et al. [228] introduced a cylindrical videoconferencing system which could display a perspective-correct avatar for multiple viewpoint applications. The authors verified the effectiveness of gaze perception by multiple participants simultaneously. They also performed a study [229] with a random hole display system by measuring the accuracy of avatar gaze perception with different horizontal and vertical viewing angles in a multiple observer situation. Furthermore, Lee et al. [172, 241, 242] performed a series of studies on how to further improve shared live-panorama-based collaborative experiences by applying mixed reality technology. The results indicated that the view awareness cues and the collaborative gaze were helpful for understanding the remote collaborator's focus.

In summary, although many works exist about gaze perception and localization, however, there is no relevant research about perception and localization on avatar gaze directions in 360° video-based RE, and the study in this section attempts to fill a gap in the literature.

5.1.3 Avatar Display and Prototype Setup

As mentioned above, two display methods were adopted to show the avatar to the host. The first solution is showing a 2D avatar on a host-oriented tablet placed perpendicular to the host, which is similar with most commercial telepresence robots. But in our case, because the robot is equipped with a 360° camera and does not have to rotate when the guest with HMD seeing around, the direction of the avatar on tablet is no longer consistent with the direction of the robot. Instead, the avatar is consistent with the guest's HMD heading. And every time after the robot travels following the guest's command, the platform rotates so that the tablet display keeps facing the host during the guest's gazing process.

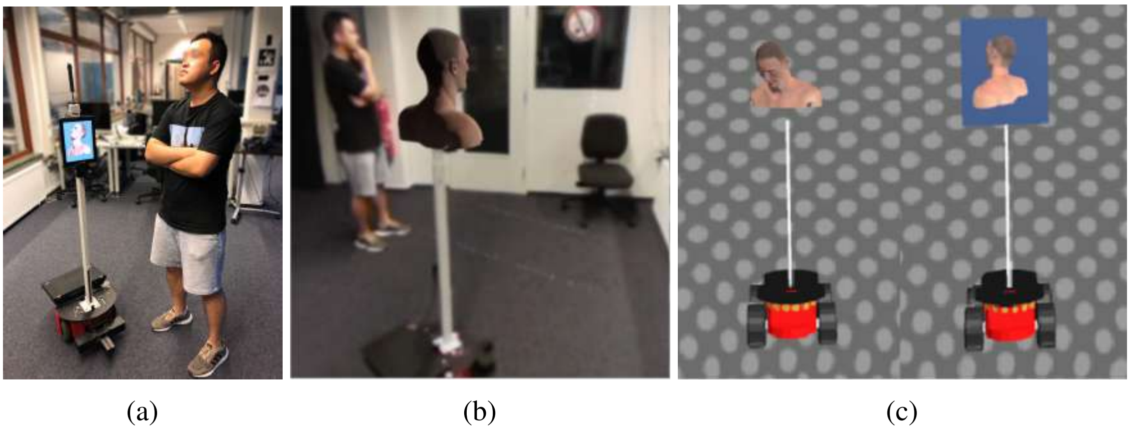


Fig. 5.3 Depiction of the robotic surrogate and simulated experimental environment in Unity. (a) The tablet variant of the telepresence robot. The host is looking towards the same direction according to the localization of the guest's gaze. (b) AR display: The 3D avatar of the guest is superimposed on top of the robot. Image captured through HTC Vive Pro HMD. (c) Simulated experimental environment in virtual reality space: AR display solution and tablet display solution.

Based on this method (Figure 5.2a), a prototype system was developed (Figure 5.3a). In this prototype, the telepresence robot consists of a Ricoh Theta S 360° camera for capturing a panoramic live stream of RE, an iPad used for showing a 2D torso avatar to the host and a Pioneer 3-DX mobile robot working as the mobile base. In the LE where the guest is located, the reconstruction and rendering of the RE where the host is situated is implemented based on a spherical space modelled in Unity3D. The live stream from the RE is rebuilt and projected as a video-texture on the inner surface of this spherical space. A virtual camera is positioned in the center of the spherical space to provide the guest with a perspective-correct view of the RE. Thus, the guest can get a real-time 360° telepresent view on the HMD with the live stream from the RE and see around flexibly.

In addition, the guest's gaze direction is available through the HMD orientation sensors. The orientation data is then transmitted to the RE via network, and used for synchronizing the avatar's head orientation displayed on the tablet. Meanwhile, it is assumed that the sensor on the telepresence robot is tracking the host consistently, and after the guest drives the telepresence robot to a new location or the host changes his or her position, the mobile base turns so that the tablet display always faces the host. The avatar displayed on the tablet, however, remains fixed in world coordinates. Unless the guest moves his or her head, the host sees the avatar maintain its gaze direction. This solution is designed such that it can be used with existing tablet-equipped telepresence systems, with the simple addition of a 360° camera.

The second solution (Figure 5.2b) is to display a 3D-rendered torso avatar with AR technology, superimposed on the robotic platform. Again, the avatar is consistent with the guest's heading and its heading is, similar to the tablet solution, completely detached from the heading of the robot. In this solution, because the avatar display would be in AR mode, the host is required to wear an AR glasses and the robotic surrogate does not need to rotate at all after travelling to a new location to keep the avatar in view for the host.

On the basis of the telepresence prototype system for the first solution, the thought of the second solution was prototyped with an HTC Vive Pro HMD which works in AR mode with two see-through cameras to display a 3D avatar together with the real environment (Figure 5.3b) instead of the tablet display. The AR 3D avatar is rendered in Unity3D and essentially a puppet, controlled by the guest's HMD orientation. In addition, a Vive tracker was attached on the robotic surrogate and used for locating the AR avatar at the correct position.

For the telepresence prototype systems above, the communication between the host's RE and the guest's LE was implemented by the ROSbridge node [307] which offered a websocket-based communication via the network and a ROS2Unity library [167] based on it.

In both solutions, a neutral expression adult torso avatar is displayed with the same size and purposely avoided an elaborate haircut to eliminate cues from the hair. Furthermore, the shoulders of the avatar torso are included because the shoulder line is an additional perspective cue as opposed to the neck which is cylindrical and offers no cues regardless of orientation. In addition to this, the torso acts as a canvas to cast a shadow from the avatar's jawline, further giving cues about the head orientation. Moreover, the avatar wore no clothes to avoid perspective cues from patterns on the clothes and to avoid attenuating the jawline shadow (clothes would make the shadow less salient).

5.1.4 Hypotheses

The goal of the user study was to evaluate the gaze localization differences between the 3D stereoscopically rendered avatar and the 2D flattened tablet avatar. This helps evaluate if the solution is possible to implement on current telepresence platforms and if the performance benefit of the 3D avatar vs. the 2D tablet merits mounting an AR headset for the host.

The experiment was designed in order to test the following hypotheses:

H1: *Participants will have better performance (lower error) in localizing the avatar gaze direction with the AR display rather than with a tablet display, as well as higher subjective evaluation.*

In order to verify this hypothesis, display *technique* and *distance* were selected as manipulated factors in the study.

H2: *Localization ability of the host will have an asymmetric nature when the avatar faces towards or away from the host, as well as for the pitch axis when the avatar gazes upwards or downwards.*

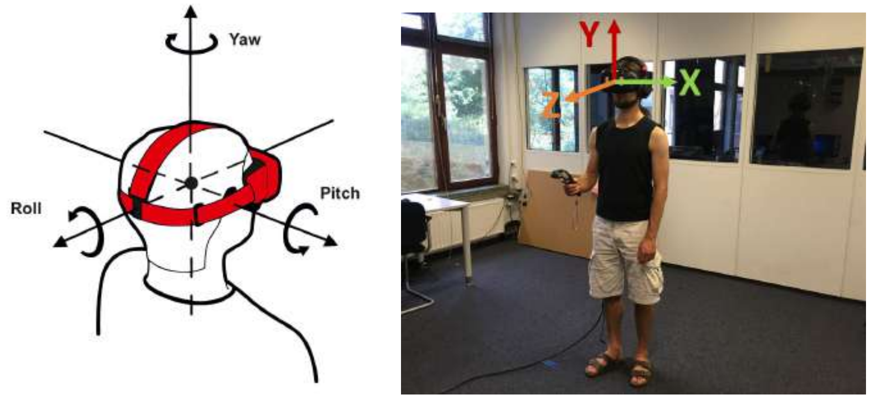
This hypothesis is assumed due to the asymmetric visual cues the host would receive from the avatar's facial structure in the respective perspectives. In order to verify this hypothesis, avatar gaze horizontal angles *yaw* and vertical angles *pitch* were introduced as manipulated factors in the study.

5.1.5 Experiment

Participants, apparatus and environment

15 participants were recruited from local department to take part in the experiment. 14 participants (ages 20-26, mean age 23.36, SD = 1.946) completed the experiment, 1 participant quit because of cybersickness. 16.7% of the participants were left-handed while 83.3% were right-handed. All participants had normal or corrected to normal vision and had prior experience with VR.

During the experiment, participants stood in the center of a 5 × 5 meter room, mounted an HTC Vive HMD, which has a resolution of 2160 × 1200 pixels (1080 × 1200 pixels per eye), a refresh rate of 90 Hz and a 110 degree field of view, and held an HTC Vive controller

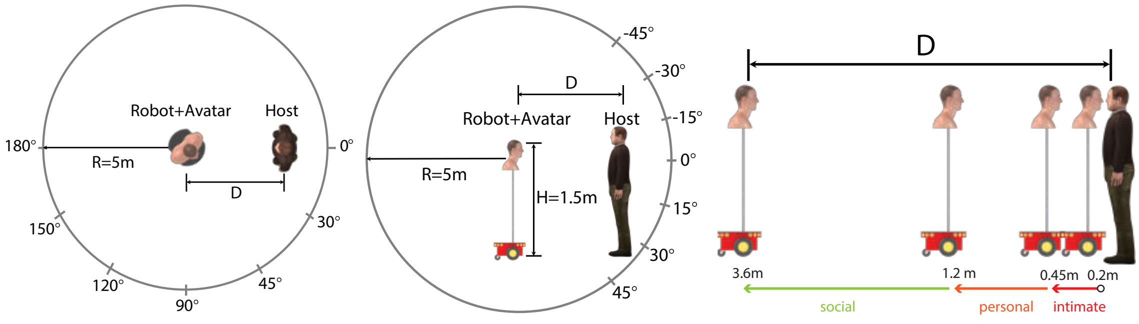


(a) (b)

Fig. 5.4 Avatar rotation axes and experimental setup: (a) Avatar rotation axes (only yaw and pitch were used in the experiment). (b) Participant and space coordinates during the experiment.

in their dominant hands (Figure 5.4). The real environment was kept quiet and participants were required to wear silencer earphones, such that no voice cues or communications is available during the process, because this user study focuses on a rapid gazing localization only based on visual cues. The experiment lasted around 45-50 minutes, and participants were free to have a break at any time during the process.

Design, stimuli and procedure



(a) (b) (c)

Fig. 5.5 Schematic layout of the experimental setup. (a) Top-down view showing yaw angles tested. (b) Side view showing pitch angles tested. (c) Distances tested based on Hall's Proxemic zones: intimate space (D = 0 - 0.45m); personal space (D = 0.45m - 1.2m); social space (D = 1.2m - 3.6m).

The experimental environment consisted of a spherical space with a radius of 5 meters. The sphere was always anchored to the avatar on the robot (Figure 5.5). A tiny polka dot texture was applied to the inside surface of the spherical space to allow participants to feel the size and bounds of the space (Figure 5.3c).

During the experiment, participants were asked to stand on a target spot. The distances between telepresence robot and participants in the virtual spherical space were chosen based on Hall's research in proxemics [106]. Specifically, 0.2m was adopted as the minimum possible distance in the intimate space, while 0.45m, 1.2m and 3.6m are the boundary distances between intimate space, personal space, social space and public space respectively (Figure 5.5c).

In addition, according to the hypothesis proposed above, the avatar attached on top of the telepresence robot were displayed with two *techniques*, which were 3D stereoscopically rendered avatar (*AR*) and 2D flattened tablet avatar (*tablet*) respectively. The height of the displayed avatar with both two techniques are fixed at 1.5m, which is also widely adopted by most of current commercial telepresence platforms in a standing situation such as Double Robotics[2]. The avatar's gaze was manipulated by rotating about the *yaw* (axis defined by the spine) and *pitch* angles (axis defined by the ears) (Figure 5.4a). Angles were 0°, 30°, 60°, 90°, 120°, 150°, 180° for clockwise yaw; -45°, -30°, -15°, 0°, 15°, 30°, 45° for pitch (Figure 5.5b and 5.5a). Figure 5.6 offers a combined overview of all the avatar poses used. The rotation angles were chosen based on the research of Youdas et al. [327] about normal motion range of the human cervical spine and Y.Pan et al.'s study [229] on head gaze estimation with a multi-view autostereoscopic display. Additionally, previous research [214, 226] suggest that because the human face is symmetric, the ability of participants to assess avatar gaze directions is independent of the side of the face and therefore there was no need to test counter-clockwise yaw rotations. i.e., the avatar head only turned clockwise to the right.

In summary, the experiment was a within subjects design with 7 horizontal angles (yaw) × 7 vertical angles (pitch) × 4 distances between avatar and participant × 2 avatar display techniques (AR and tablet) × 2 repetitions for each condition for a total of 784 trials. All trials were shuffled randomly. 14 participants × 784 trials per participant = 10976 total trials collected.

When the experiment started, an avatar attached to the telepresence robot appeared directly in front of the participants in four specific distances (D = 0.2m, 0.45m, 1.2m, 3.6m) with two display techniques (AR and tablet) respectively (Figure 5.3c). The avatar was displayed for 2 seconds and after that it disappeared. Specifically, this parameter was set because in pilot test, it was noticed that some participants used controllers as a reference by making the ray from the controller go through the avatar head. However, participants were hoped to finish the task only depending on localizing with visual cues and were prevented

[2] https://www.doublerobotics.com

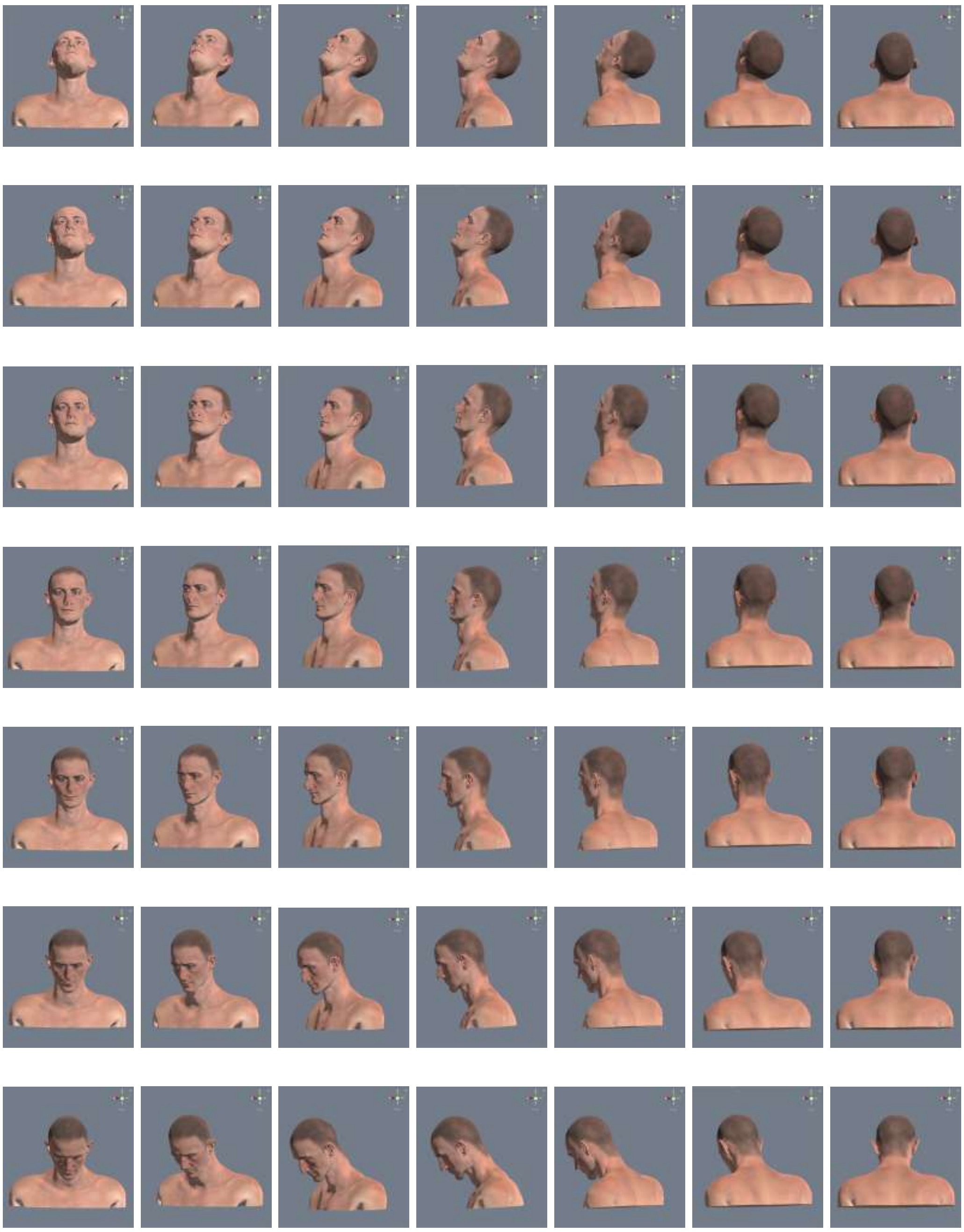


Fig. 5.6 Overview of all yaw angles (from left to right: 0°, 30°, 60°, 90°, 120°, 150°, 180°) and pitch angles (from up to down: -45°, -30°, -15°, 0°, 15°, 30°, 45°) for avatar gaze directions in a 0.45m distance. Here, the avatar torso was implemented based on the research of Hietanen et al. [117].

from using such strategies. In addition to this, controlling avatar display time could keep participant's localizing time in every gaze directions same and force them to concentrate on the avatar and localize rapidly, which is similar with some real scenarios like an sold apartment viewing where the gaze directions of the guest changes very frequently and the dwelling time is short. Participants then had to point to indicate, using the controller in their hands, which point on the sphere they thought the avatar was looking at. A pink sphere was displayed on the surface of the spherical space as cursor at the participant's controller intersection point. Participants would click the trigger button on the controller to confirm their estimation of the gaze point and one second later the avatar would appear for the next trial. Since some of the gaze points in the 0° - 30° yaw range were behind the participants, they were allowed to pivot in place to point.

Data was analyzed with a repeated-measures ANOVA. Display *technique*, *distance*, *yaw*, *pitch* were the manipulated factors. The effect of these factors on angular error was investigated, i.e. the angle between the line defined by the avatar's gaze direction and an avatar gaze line to the point that the participant pointed. Angular error will simply be referred to as *Error* (symbol *e*) for the remainder of this section.

Having described the experiment above, now the rationale could be explained why this experiment is being conducted in simulated AR [156, 252, 248, 315, 18] instead of Video-See-Through or Optical-See-Through AR with real setup: when the experiment is conducted in a standard laboratory room, even if the room had been completely clear of furniture, there would still be wall outlets, windows, doors, light fixtures etc. These would act as salience hot-spots and therefore when the avatar's gaze landed in their vicinity, participants would most likely be biased to think that the avatar is looking at those hot-spots and thus skew the results. Even if the space was completely covered with curtains or other partitions, the partition frames or curtain rings etc. would be salience hot-spots. In addition to this, if a non-spherical room had been used, some pitch/yaw angles would result in the avatar's gaze intersecting the room at different distances and thus participants would be tasked to make estimations at different distances.

5.1.6 Results

A normality check for all factor levels was performed using Shapiro-Wilk test [262] before the analysis, and the results did not show a strong indication of normal distribution. However, as shown in some previous research [92, 110, 180], a moderate deviations from normality is allowed by ANOVA, especially when considering a large sample pool of this study (10976 total trials).

The results of ANOVA for all factors are presented in Table 5.1. For the main effects, *Distance* had a significant effect on *Error* ($F_{3,39} = 4.77, p < 0.01, \eta^2 = 0.013$). A plot for distance can be found in Figure 5.7; *Technique* also had a significant effect on *Error*

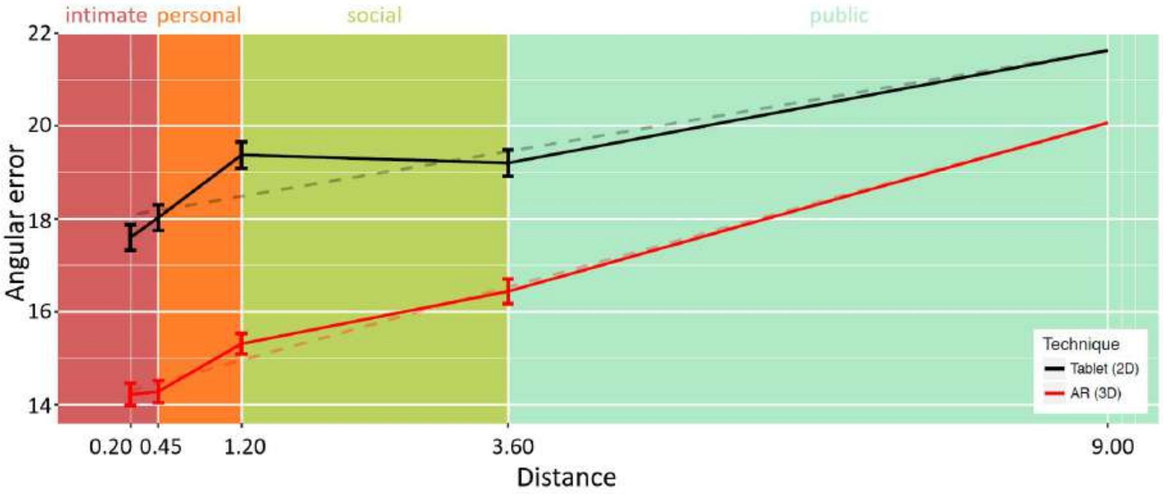


Fig. 5.7 Effect of *distance* on angular *error* (lower is better). Dotted line is linear regression results according to the measurement data at distances of 0.2m, 0.45m, 1.2m and 3.6m. Based on this, an extrapolation beyond distance of 3.6m could be predicted, which were not tested in the experiment.

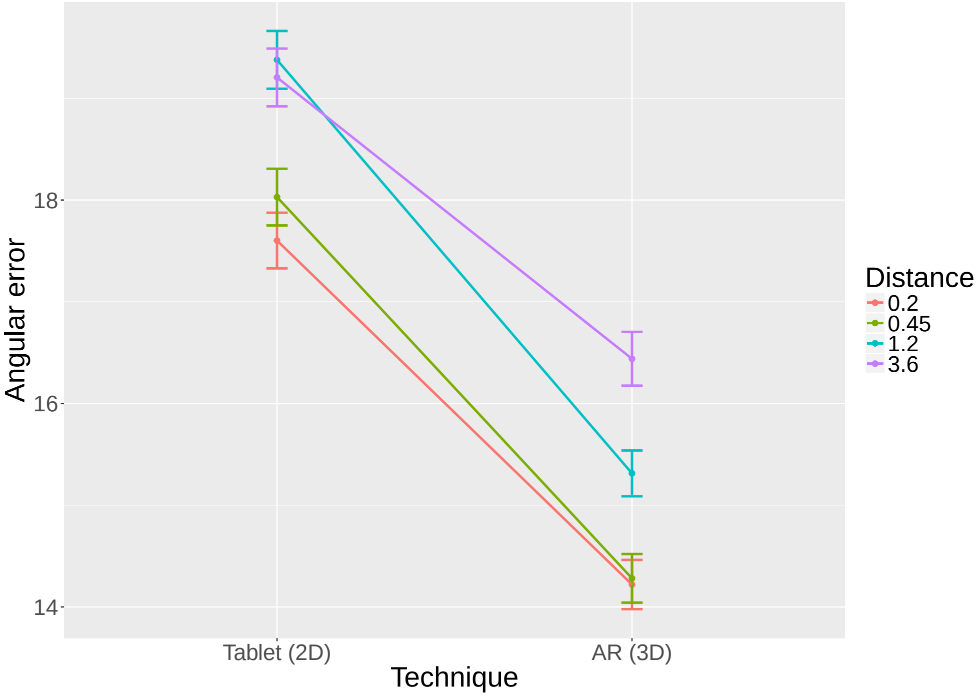


Fig. 5.8 Effect of display *technique* on angular *error* (lower is better).

($F_{1,13} = 26.47, p < 0.001, \eta^2 = 0.060$). A plot for *technique* can be seen in Figure 5.8; *Yaw* had a significant effect on *Error* ($F_{6,78} = 11.89, p < 0.001, \eta^2 = 0.080$), a plot can be seen in Figure 5.9; And *Pitch* also had a significant effect on *Error* ($F_{6,78} = 5.00, p < 0.001, \eta^2 =$

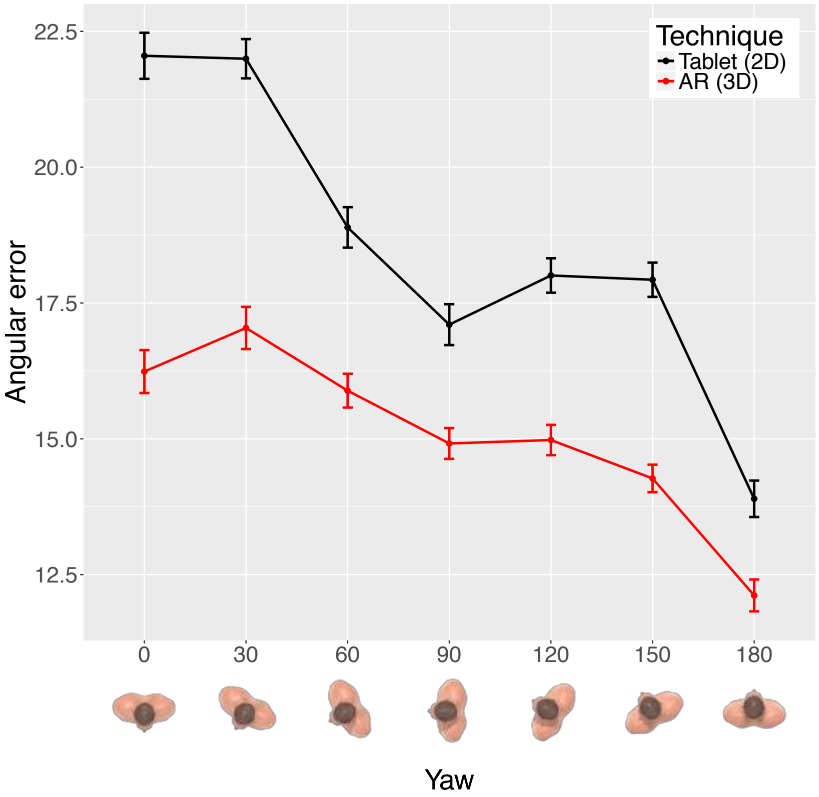


Fig. 5.9 Effect of *yaw* angles on angular *error* (lower is better).

0.038) as presented in Figure 5.10. The error bars in figures indicate the standard error (SE). Furthermore, most of the interaction effects had significant influences on *Error*, except for *Technique:Distance*, *Technique:Distance:Pitch* and *Technique:Distance:Pitch:Yaw* (Table 5.1).

In order to confirm the reliability of the ANOVA results above, a Friedman test [122] was also conducted for a double check of the data and same results were got that all main factors had significant effects on *Error*. In addition to this, pairwise t tests were performed to check if there was significant difference between each levels of every factors. The results of pairwise t tests on *Distance*, *Yaw* and *Pitch* are listed in Table 5.2, Table 5.3 and Table 5.4 separately.

In a post-experiment questionnaire, participants were asked to rate the avatar display techniques on a 5-point Likert scale (1 - I do not like it at all; 5 - I like it very much.). Results suggest that participants preferred the AR technique (μ_{AR} = 3.64, SD = 0.84) over the tablet (μ_{Tablet} = 2.29, SD = 0.91). Moreover, a Wilcoxon signed rank test was adopted to have further analysis on the data. The results also present a significant difference between AR preferred rating and tablet preferred rating ($p < 0.05$). In addition, when asked to comment

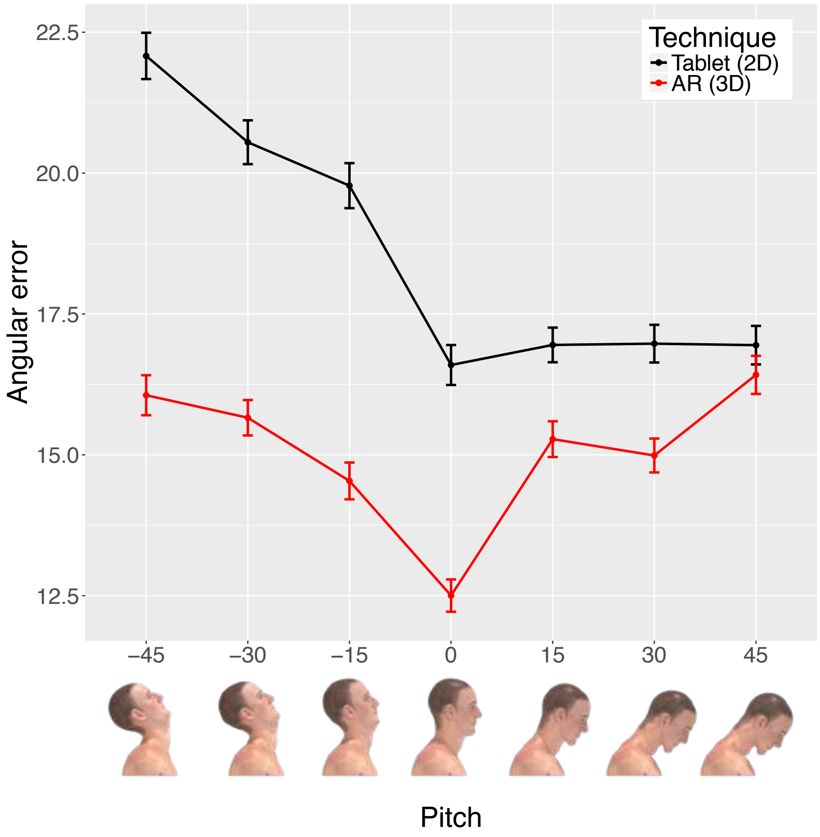


Fig. 5.10 Effect of *pitch* angles on angular *error* (lower is better).

about their choice, most of the participants (11/14) commented on the realistic and natural avatar display in the AR mode. e.g.: “It was easier to follow person’s gaze in AR"; “It seemed to be more natural in AR, and in some orientations it was really hard to estimate the direction in tablet mode"; “In the AR mode, the orientation of the avatar was easier to perceive".

5.1.7 Discussion

Distance and Technique

The results confirm the hypotheses to some extent. Figure 5.7 suggests that the mean error showed a rising trend in general as the avatar distance increases. In addition, for every tested distance, the mean error for AR technique was obviously lower than the tablet display, which is also illustrated clearly in Figure 5.8. An avatar displayed stereoscopically in the AR technique has depth and volume. On the tablet technique the avatar is “flattened” and therefore the volume of the avatar is lost. Furthermore, in the AR mode participants could get added parallax effects from slightly translating their heads horizontally.

Table 5.1 The results of analysis of variance (ANOVA).

Main effects	F	p	$p < .05$	η^2
Technique	26.473	1.883e-04	$*$	0.060
Distance	4.773	6.285e-03	$*$	0.013
Pitch	4.996	2.235e-04	$*$	0.038
Yaw	11.888	1.975e-09	$*$	0.080

Interaction effects	F	p	$p < .05$	η^2
Technique:Distance	2.139	1.108e-01		0.001
Technique:Pitch	17.027	1.814e-12	$*$	0.019
Distance:Pitch	6.426	7.607e-13	$*$	0.019
Technique:Yaw	4.919	2.585e-04	$*$	0.009
Distance:Yaw	17.062	3.372e-33	$*$	0.074
Pitch:Yaw	3.694	3.767e-11	$*$	0.037
Technique:Distance:Pitch	1.513	8.624e-02		0.003
Technique:Distance:Yaw	1.819	2.407e-02	$*$	0.004
Technique:Pitch:Yaw	2.105	2.786e-04	$*$	0.010
Distance:Pitch:Yaw	2.186	2.182e-10	$*$	0.028
Technique:Distance:Pitch:Yaw	1.150	1.477e-01		0.013

Table 5.2 The results of paired t-test on distances.

	0.2m	*0.45m*	*1.2m*
0.45m	0.35	-	-
1.2m	1.8e-07	1.9e-05	-
3.6m	9.1e-12	3.9e-09	0.15

However, the two display techniques showed different trends in specific proxemic distances. Both display techniques had the smallest angular error in the minimum distance of 0.2m (AR: 14.220°, tablet: 17.601°). In the intimate space (0.2m-0.45m), the angular error for the AR technique had a slight increase with distance, while the angular error for the tablet technique increased faster within this distance interval. i.e. in the intimate space, the difference of angular error between the two display techniques became larger (Tablet: +0.4° $>$ AR: +0.06°, difference = +0.366°). In the personal space, the angular error for both techniques started to increase more rapidly. In the social space, the angular error for the AR display continued to increase rapidly, while the angular error for the tablet display showed a modest decrease.

In summary, the AR display technique maintained a steady error throughout the intimate space and then degraded in personal and social space. Conversely, the tablet display technique had the largest error at the border between personal and social space (1.2m). Extrapolating

Table 5.3 The results of paired t-test on yaw angles.

	0°	*30°*	*60°*	*90°*	*120°*	*150°*
30°	0.70964	-	-	-	-	-
60°	4.7e-05	6.3e-08	-	-	-	-
90°	3.2e-15	< 2e-16	0.00037	-	-	-
120°	7.1e-12	< 2e-16	0.02987	0.51969	-	-
150°	1.1e-15	< 2e-16	0.00037	0.77149	0.56457	-
180°	< 2e-16	< 2e-16	< 2e-16	< 2e-16	< 2e-16	< 2e-16

Table 5.4 The results of paired t-test on pitch angles.

	-45°	*-30°*	*-15°*	*0°*	*15°*	*30°*
-30°	0.06413	-	-	-	-	-
-15°	1.0e-05	0.06413	-	-	-	-
0°	< 2e-16	< 2e-16	3.8e-12	-	-	-
15°	4.7e-15	6.9e-08	0.01847	1.3e-05	-	-
30°	3.9e-16	9.3e-09	0.00617	0.00011	0.67316	-
45°	2.2e-09	0.00051	0.37229	3.2e-09	0.24660	0.13400

beyond 3.6m, it is reasonable to predict that the difference of angular error between AR and tablet display technique will dissipate because of the gradual loss of cues as both the tablet and the avatar become smaller (Figure 5.7).

Yaw

Analysis of the yaw angles (Figure 5.9) suggests that yaw angles had a different effect and distribution on the error of the two display techniques. Overall, the error exhibited a decreasing trend as the yaw angle increased (i.e. as the avatar torso turned to the right and eventually backwards) with a sizable drop from 150° to 180° for both display techniques. i.e. Participants could best localize the avatar gaze when the avatar was facing away from them (in this case participant and avatar line of sight is practically collinear). It is postulated that avatar geometry accounts only partially for this result. Perhaps due to this position being directly across from the participants, they could point to it from a rest pointing position (elbow resting on the ribs) without moving their forearm much and as such their accuracy does not degrade as much as when making large sweeping motions [163]. This is a limitation of this study and a solution to this is perhaps instead of indicating the estimated gaze point with the wand, to control a cursor that rolls on the sphere using some sort of trackball.

In order to verify the hypothesis of asymmetry around the yaw axis (H2), the yaw levels were reorganized into a new factor called *horizontal heading* with two levels based on the avatar's horizontal direction: *towards host*, when the avatar faced the participant (0°, 30° and 60°) and *away from host*, when avatar faced away from the participant (120°, 150° and 180°). Side trials were excluded (90°).

The reorganized factor *horizontal heading* was subjected to an ANOVA test. A significant difference was found between the avatar's horizontal heading: *towards host* and *away from host* ($F_{1,13} = 14.20, p < 0.05, \eta^2 = 0.108$). Furthermore, for both display techniques, when the avatar faces away from the host the error is much smaller than when the avatar faces the host (towards host: e_{AR} = 16.39°, e_{tablet} = 20.98°; away from host: e_{AR} = 13.79°, e_{tablet} = 16.61°). This confirms the towards-away from host asymmetry hypothesis. This report of gaze estimation asymmetry data when an avatar is facing towards or away from a VR user, to the best of our knowledge, is the first of its kind in the literature.

Another possible explanation for this asymmetry and for the best performance at 180° in the experiment is that participants perhaps localize better when the avatar head and the gaze point on the sphere all fit within the bounds of the participant's field of view (FOV). If they do, participants could be better at estimating the relative spatial attributes. Conversely, if participants need to shift their FOV (and their torso in many cases) to find the target and thus the avatar leaves their FOV, they could be performing worse. In an attempt to verify this, the effect of avatar gaze point Z coordinate on *error* was plotted (Figure 5.11). i.e. the Z coordinate of the point where the avatar's gaze half-line intersected the sphere. During the experiment, participants stood at the origin (0, 0, 0) of the virtual space facing the avatar. So essentially, when the gaze point Z coordinate was positive, both the avatar and the corresponding gaze point would be located in front of the participant; Conversely, when the Z coordinate was negative, the target gaze point would be located in some area behind the participant, outside their FOV, and therefore localization would be more challenging. Added to this, positions behind the participant demanded a pivot in place by the participants which further exacerbates localization error.

The data for the tablet and AR display technique were fitted using local regression [52] ($\alpha = 1$). The fitted curves support the new hypothesis to a certain extent, nevertheless more rigorous study is needed before firm conclusions can be made.

Pitch

Regarding pitch, Figure 5.10 shows that generally, the slope of the error for the AR display technique had a symmetric distribution and the best performing pitch angle for both techniques was 0°, i.e. when the avatar head was level. Surprisingly the tablet display technique exhibited greater asymmetry; beyond 0°, there was only a minor increase in error as the avatar looked downwards.

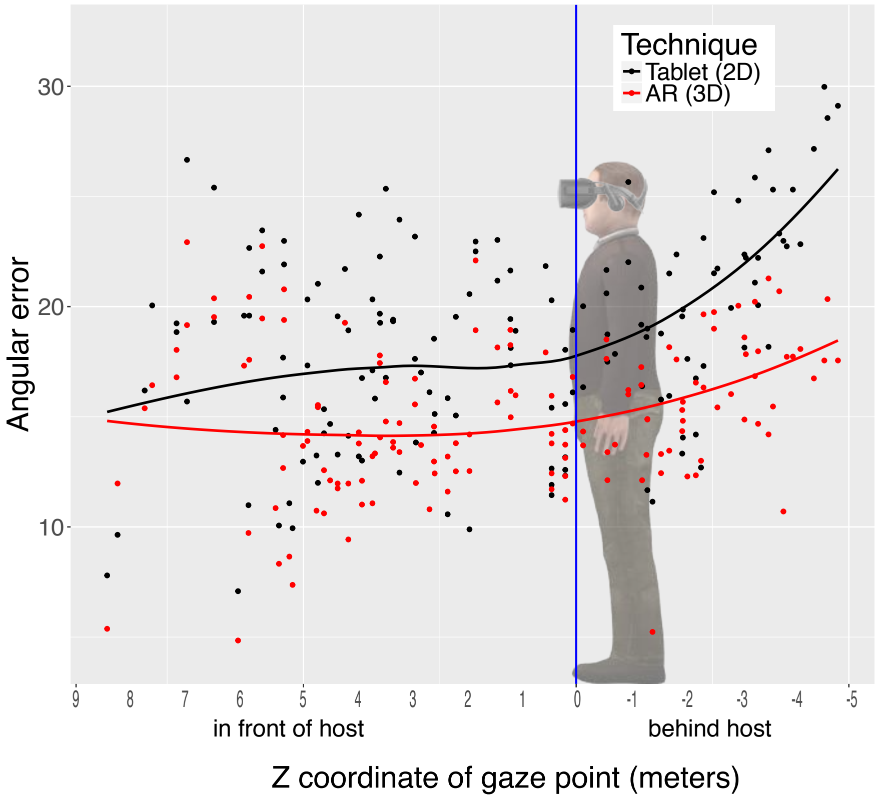


Fig. 5.11 Effect of gaze point Z coordinate on error. Curves are a product of fitting with local regression ($\alpha = 1$).

Considering the hypothesis on the asymmetry about the pitch axis, the data was reorganized into a new factor *vertical heading* with two levels: up (when avatar faced upwards i.e. -45°, -30° and -15°) and down (when the avatar faced downwards i.e. 15°, 30° and 45°).

Again, the reorganized data *vertical heading* was analyzed with ANOVA. A significant difference between the avatar's vertical heading: *up* and *down* was found with $F_{1,13} = 4.85, p < 0.05, \eta^2 = 0.070$. In addition, for the AR display technique, there was no obvious difference of angular error when the avatar faced upwards or downwards (for AR: e_{up} = 15.42°, e_{down} = 15.56°); However, for the tablet display technique, there was a sizable difference in angular error (for tablet: e_{up} = 20.80°, e_{down} = 16.96°). This is also visible in Figure 5.10, when the tablet technique error levels off beyond 0 degrees pitch angle. This finding means that H2 regarding pitch was only partially verified; it was only true for the tablet display technique. Hence, it seems that with the flattened tablet display technique, there was poor estimation of the gaze point when the avatar looked downwards compared to the AR technique.

A potential explanation for this is that for pitch angles beyond 0°, there is poor visibility of the jawline and its shadow on the neck as well as the fact that the eyes are partially occluded by the forehead. Given an AR avatar, participants were more sensitive to those minor differences.

Applications

The findings from this experiment should serve as guidelines for telepresence system designers. E.g. in the apartment viewing scenario, since the environment is not dynamic, the guest could be supplied with a static image of the room; he or she could also jump back to previously captured positions using some VR teleportation technique [35]. The physical robot/avatar would then be free, considering the available surrounding space and the gaze estimation performance of each pose, to continuously re-position itself to cater for gaze perception by the host and his or her proxemic preferences. Conversely, in a museum there might be numerous people walking around. In this case, the live stream from the remote environment is very important for the teleoperator to understand the surroundings as well as interact and converse with other people. In such a situation the robot would most likely not be able to re-position itself continuously and the host would adjust his or her position to get an ideal perception of the avatar gaze direction.

Gaze estimation by the host is only *one* of the many factors that contribute to an effective *hosted telepresence* session. The importance for the host to be able to localize the guest's gaze direction is heavily dependent on the scenario. Some of the high performing avatar poses were facing away from the host; in these positions eye-contact, facial expressions and therefore emotions are unavailable to the host. It should potentially be left up to the host to choose how he/she wants the robotic platform and avatar to behave. If the host is interested in gaze localization then the host interface should provide some option to focus on that; in which case the robot/avatar should dynamically re-position itself to cater for this. If the host is interested in facial expressions/emotion then the tablet/avatar should always face the host and display the face of the guest.

Limitations

This study does not take into account minor errors introduced by pointing with a wand [208, 163]. As mentioned earlier an alternative pointing method based on a trackball could be devised to alleviate this problem. It should be also noted that eyelids and eyebrows were static on the avatar's face. In practice, however, gazing upwards or downwards modulates the opening of the eyelids and the position of the eyebrows. These cues could be added in future studies to enhance accuracy of the results. Also, fixing avatar height at 1.5m during the experiment was another limitation of this study, the avatar height should be able to change correspondingly according to participant height.

In addition, conducting user study in a simulated environment might have some limitations in terms of ecological validity. In real applications, the localization errors would be slightly lower than the results of this experiments, since the layout of the environment, the features of specific objects (like colors and materials) and voice communication between the guest and the host could also provide people with unique cues to help them localize the avatar gaze more accurately. The results of this experiments are conservatively measured in a virtual ideal environment without any other information except for visual cues from avatar gaze, hence, the localization errors from the experiments should be referenced as a *strict baseline* for short-time gaze localization of an avatar.

5.1.8 Conclusion

Two methods were presented for improving gaze direction estimation of a guest in a 360° *hosted telepresence*: constantly orienting the tablet to face the host and an AR avatar method. A formal user study (in simulated virtual environment) was conducted that shed some light into the performance of these two proposed display techniques. A summary of the findings in the experiments:

- Telepresence hosts were 19% better at estimating the gaze of a stereoscopically rendered avatar over a flattened avatar image.
- Certain combinations of yaw, pitch and distance were better for the host to estimate the gaze direction (e.g., yaw = 180° / pitch = 0° @1.2m with the AR display has the least localization error 4.84° among all the tested conditions). This is the first time such results have been reported in the literature.
- Yaw of the guest avatar head can affect localization accuracy by as much as 22.05° for a tablet display and 17.04° for an AR display.
- Pitch of the guest avatar head can affect localization accuracy by as much as 22.08° for a tablet display and 16.42° for an AR display.
- Host localization error exhibits an increasing trend with distance from the guest avatar, but not a linear one.

These results should provide reference for *hosted telepresence* system designers but also for any social VR systems where participants are required to estimate the gaze of an avatar.

Many open questions remain for further research. For example, The current study only focuses on a one-to-one application between teleoperator and host. But in some real scenarios, normally there are multiple hosts who need to localize teleoperator's gaze simultaneously. In this case, it is assumed that for every host, the localization error would be lower comparing with the results of single host study. Because the host would get reference from other hosts to help them fixing the localization error. However, this has to be verified in further study. Furthermore, if the robot reconstructs a 3D mesh of the room as it travels along, and the guest tele-ports to a previously visited spot, the robotic platform does not need to travel back to that point. The guest avatar then would be "detached" from the physical robot, creating a triangle of host-robot-guest, thus creating entirely new social dynamics in that space, raising further questions about telepresence.

5.2 Evaluation of Proxemics in Dynamic Interaction with an Avatar Robot

5.2.1 Motivation

When interacting with a robot, human observers wish to comprehend their current state, understand their purpose and predict their further actions [82, 93]. Earlier research [173, 121] suggests that if human could improve a robot's personality with some extent of social intelligence, the robot will be more predictable and easier to understand. However, this is very challenging to achieve considering the diverse shapes and different locomotion methods of robots. Humanoid robot designs [230, 153, 37, 201] might make robot's behavior easier to perceive and predict by introducing human-like gestures or facial expressions, but constraints on the mechanical build still exist while humanoid designs are not optimal for all kinds of tasks (like traveling in a narrow pipe).

Mixed reality (MR) technology provides a solution to this problem. A mixed reality avatar could maintain the original build of robots but present a metaphysical state of the robot by visualization such as gestures, animation or facial expression simultaneously. Furthermore, a mixed reality avatar could help people notice and recognize some tiny ground robots in advance in some complicated or crowded environments, so as to avoid distraction and discomfort when the robots suddenly appear nearby or invade their private space. In this section, the setup consisting of a robot and a mixed reality avatar is referred to as *avatar robot*.

The work in this section specifically focuses on proxemics in dynamic interaction with a mixed reality avatar robot, in which the avatar arm swing animation is chosen to reflect the current locomotion speed of the robot. Perceiving the moving speed of a wheeled robot correctly and then choosing suitable proxemic preferences in a dynamic situation like guidance or head-on encounters could reduce discomfort to nearby pedestrians. The application of mixed reality avatars together with mobile robots raises therefore the question: With a mixed reality avatar presented overlayed on top of the robot, which factors influence the proxemic preferences of humans in a dynamic interaction? The design space is explored by evaluating two elements in a series of user studies which manipulates robot locomotion speed and the avatar visibility.

5.2.2 Related Works

MR and AR in Human Robot Interaction

In a series of papers, Dragone and Holz [70, 71, 123] have raised the idea that displaying a humanoid avatar upon the robot platform to broadcast the current state of the robot could help people understand the robot's state more effectively. Those studies, however, include no user evaluation. Similarly, Young et al. [329] proposed a method using cartooning to express different states of the robot. However, there was no mention in that work on how to apply this approach to express spatial intent. Following up from Dragone et al.'s work, Katzakis et al. [146] used a mixed-reality avatar to signal abrupt direction changes of a robotic platform by using "Body" and "Path" cues. The experiments in this section complement that work, further exploring the augmented surrogate/avatar design space.

Other works explored how to apply external visualization and expression to depict the internal state of a robot. For example, Collett et al. [53] explored how to use AR-based visualization to help debugging various sensors of the robot. Hoenig et al. [120] suggested that mixed reality can reduce the gap between simulation and implementation by enabling the prototyping of algorithms on a combination of physical and virtual objects. Walker et al. [316] investigated how augmented reality could mediate human-robot interaction by communicating robot motion intent and found that the objective task efficiency could be improved with this method. Hedayati et al. [112] prototyped several aerial robot teleoperation interfaces using AR and reported improvement in interaction by liberating users from an attention-divided mode. Other research works [91, 294, 336] also explored how to improve path planning and navigation of robot with augmented reality for higher accuracy and reduced errors. Following up from these previous works, the study in this section is designed to investigate if avatar arm swing animation could indicate and present the moving speed of a mobile robot.

Proxemics in Human-Robot Interaction

Young et al. [328] implemented a dog-leash human-robot interface which enables a person to lead a robot simply by holding the leash. The authors evaluated the comfort distance between robot and human. Walters et al. [317] investigated human-robot and robot-human approaching distances and suggested that subjects' personality profiles influenced personal spatial zones in human-robot interaction experiments. Furthermore, they ran a study [318] which focused on long-term human-robot proxemics and found that the majority of human-robot proxemic adaptation occurred in the first two interaction sessions, the distance preferences remained relatively steady for the rest of time. In addition, Mead et al.'s work [190, 191] "Autonomous human-robot proxemics" proposed a socially aware navigation method based on interaction potential. Kim et al. [157] investigated how social distance can serve as a lens through which people can understand human-robot relationships. Similarly, Mumm et al. [206] explored how people physically and psychologically distance themselves from a robot. All of these works provided valuable references to the study in this section.

5.2.3 Hypotheses

The user experiment is designed in order to test the following hypotheses:

H1: *Participants will tend to choose a higher frequency of avatar arm swing when the robot travels faster speed.*

This is assumed due to the function of arm swing in human walking which is regarded as a passive movement of gait [196, 54]. In order to verify this hypothesis, *robot moving speed* was chosen as a manipulated factor in the first experiment.

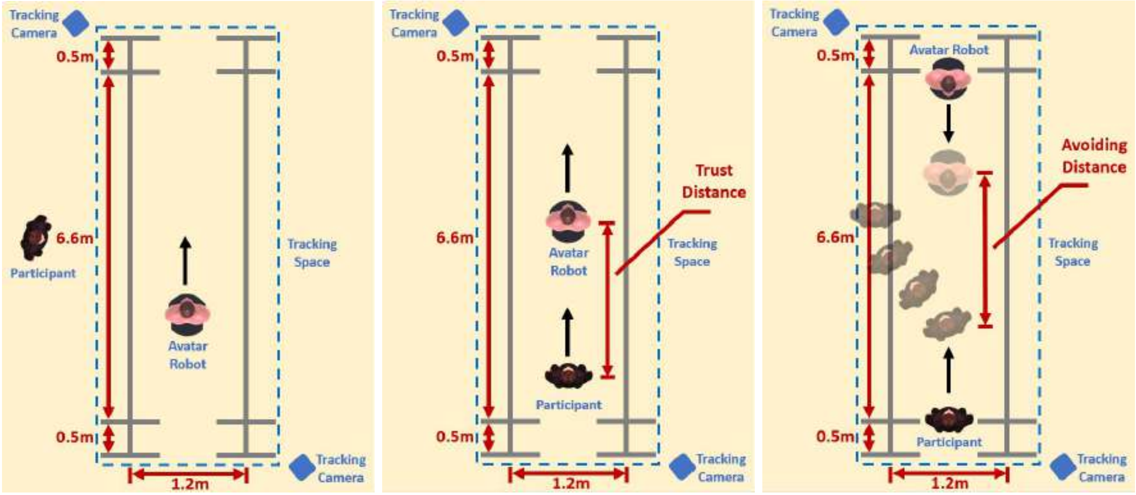
H2: *When following an avatar robot, with robot moving speed increase,* trust distance *between participants and robot as well as participant's walking variability will also increase.*

H3: *When participants are on a head-on collision trajectory with an avatar robot, a speed increase by the robot will result in participants having a closer* avoiding distance *from the robot.*

These two hypotheses are made considering the extent of trust between human and a robot in a dynamic situation. When following a robot that is moving rapidly, participants should theoretically maintain a longer trust distance between themselves and the robot to account for sudden direction or velocity changes of the robot. This is identical to how drivers maintain longer distances from a preceding vehicle when driving at high speeds on a highway. Similarly, when walking towards a faster moving robot, participants will need more time to perceive and understand the next potential movement of the robot so as to reduce the uncertainty, which as a result will cause a shorter avoiding distance between them.

H4: *Using avatar arm swing animation to present the moving speed of the robot will have an effect on the trust distance, walking variability and avoiding distance compared to the situation without avatar (robot only).*

It is assumed that using an arm swing animation on the avatar to depict the robot moving speed will help participants to perceive and predict a potential movement of robot better in dynamic interaction. In order to verify the Hypothesis 2-4, *robot moving speed* and *avatar visibility* were chosen as manipulated factors in the second experiment.



(a) (b) (c)

Fig. 5.12 Schematic layouts of experiments: (a) Avatar arm swing estimation: participants wearing HMD stood outside of tracking area, perceived and adjusted the most natural frequency of avatar arm swing with controllers during robot travelling in different speeds. (b) Walking following an avatar robot: distance between participant and avatar robot was defined as "*trust distance*". (c) Walking towards an avatar robot: distance between participant and avatar robot when participant changed his walking direction to avoid a potential collision was defined as "*avoiding distance*".

5.2.4 Experimental Setup

During the experiment, all participants were required to wear an HTC Vive Pro HMD with a resolution of 2880 × 1600 pixels (1440 × 1600 pixels per eye), which was working in AR

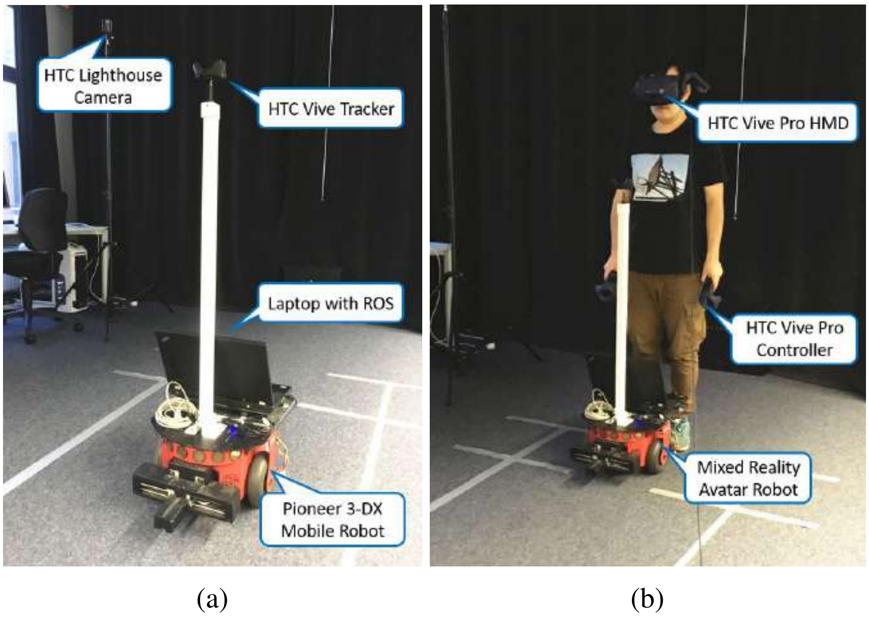


Fig. 5.13 The experimental and participant setup: (a) Mixed-reality avatar robot consisting of an HTC Vive tracker, a laptop running Robot Operating System (ROS) and a Pioneer 3-DX mobile robot. (b) Ready state of participant and robot in the task "walking following an avatar robot".

mode using the embedded front-facing cameras and the Vive SRWorks SDK toolkit. The diagonal field of view is approximately 110° and the refresh rate is 90Hz. An HTC Vive tracker was fixed on top of a pole for enhanced tracking (Figure 5.13). The mixed reality avatar had a height of 1.75m and the setup guarantees that the avatar would stay superimposed on the robot consistently throughout the experiment. Unity3D was used for rendering the mixed-reality avatar and communicating with the robot. A laptop running Robot Operating System (ROS) was connected with the robot to receive the commands from Unity and control the robot's movement. During the experiments, the lab environment was slightly dimmed and quiet. Participants used HTC Vive controllers as input devices to perform the specific tasks described below.

The avatar attached to the robot had a neutral facial expression (Figure 5.16). A few reasons motivated us to only show the avatar torso in the studies: First, avatar legs in mixed reality would have to be overlayed on the physical robot, which might have occluded the physical robot, confused participants and thus result in safety issues. Second, in actual scenarios like navigating through a crowd in a busy city, legs are often occluded by torsos of other bystanders. Furthermore, as a passive motion of gait, arm swing frequency provides information on step frequency as well. For these reasons, only the avatar torso was used.

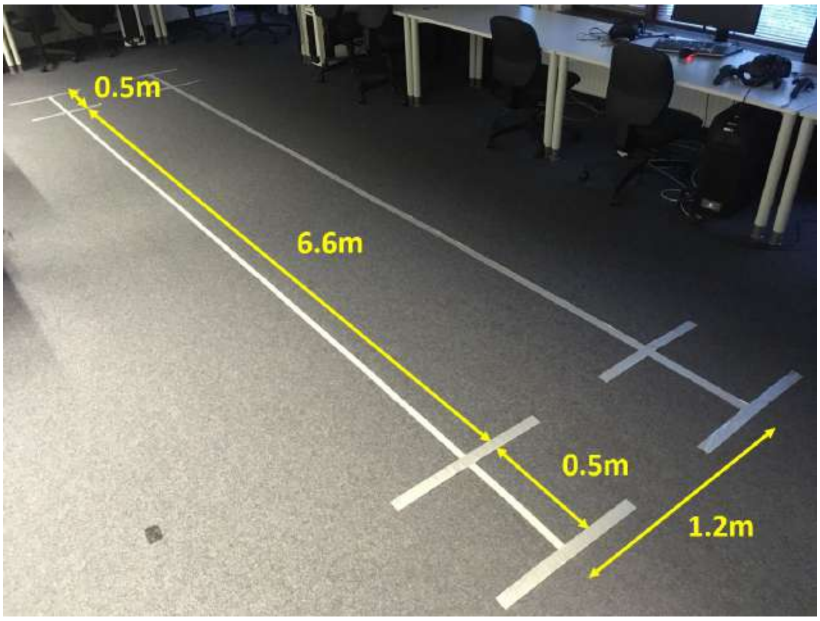


Fig. 5.14 The experimental environment.

5.2.5 Methods

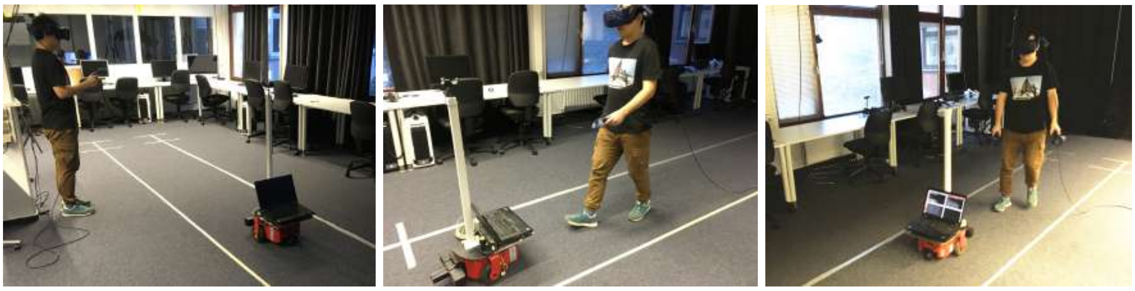
Experiment 1 (E1): Estimation of Avatar Arm Swing Frequency

The first experiment attempts to determine what is the most natural match between the avatar's arm swing frequency and the robot moving speed. In this part, participants were asked to decide the arm swing frequency of a mixed reality avatar based on the locomotion speed of a mobile robot from the perspective of a bystander. The schematic layout of this study could be seen in Figure 5.12a. During the experiment, participant mounted an HMD and held the controllers in their hands. They were required to stand outside the tracking space (Figure 5.15a). The distance from the standing point to the robot's moving trajectory was set as 1.5m in the social space based on Hall's proxemic zones [106]. When the study began, participants clicked the trigger button on the controllers to awake the avatar robot to move forward in a specific speed. Three speeds (0.8m/s, 1.0m/s and 1.2m/s) were chosen for the robot to match a person's walking based on previous research [118].

During the trial, participants were allowed to follow the robot movement by rotating their heads freely. The avatar was displayed 1 second after the robot's departure to avoid incongruities during the robot's acceleration to the respective speed. The default frequency of avatar arm swing was set as 1Hz (1 cycle for each arm per second). Furthermore, some gains were applied to enable participants to change avatar arm swing frequency. With this method, the real frequency of avatar arm swing could be decided as $f_r = g \cdot f_d$, where f_r was the real frequency of avatar arm swing, f_d was the default frequency set as 1Hz and g was the applied gain. The default gain when robot's departure was 1.0. When the gain was manipulated larger than 1.0, the avatar would have a faster arm swing animation which frequency was larger than 1Hz. Conversely, when the gain was smaller than 1.0, the avatar would swing the arm slower than 1Hz. With the robot moving, participants needed to observe the robot speed and adjust the avatar arm swing frequency to match it. Participants could accomplish this by pressing the left or right button on the touchpad of the controllers. The gain could be modulated in a range of 0.6 to 1.4 in steps of ±0.1 per click, and affect the avatar arm swing frequency in real time. In order to guarantee that for each speed participants had a fixed duration to observe and decide on the arm swing frequency, the avatar display time was set to 6 seconds for each trial. The gain applied to the avatar arm swing frequency at the end of the 6 second run was kept as the participant's final chosen gain for that trial. Each level in the speed factor was tested 4 times for each participant.

15 participants were invited from local department to take part in the first experiment (ages 21-40, mean age 26.87, SD = 6.323). All participants had normal or corrected to normal vision and most of them (13/15) had prior experience with a mixed reality headset. In summary, the experiment was a within subjects design with 3 robot speeds $\times$ 4 repetitions for a total of 12 trials. 15 participants $\times$ 12 trials per participant $=$ 180 total trials collected. All of the trials appeared in randomized order. The experiment lasted for around 15 minutes.

Experiment 2 (E2): Evaluation of Proxemics with Avatar Robot



(a) (b) (c)

Fig. 5.15 Experiment views from a third-person perspective: (a) Avatar arm swing frequency estimation. (b) Walking following an avatar robot. (c) Walking towards an avatar robot.

With the results from the first experiment, the avatar arm swing animation was used to present the corresponding robot moving speeds in the proxemics estimation experiment, in which the proxemics between a human subject and a mixed reality avatar robot was explored and evaluated in dynamic interactions. The experiment included two tasks: walking following and walking towards an avatar robot. The task order was shuffled based on the

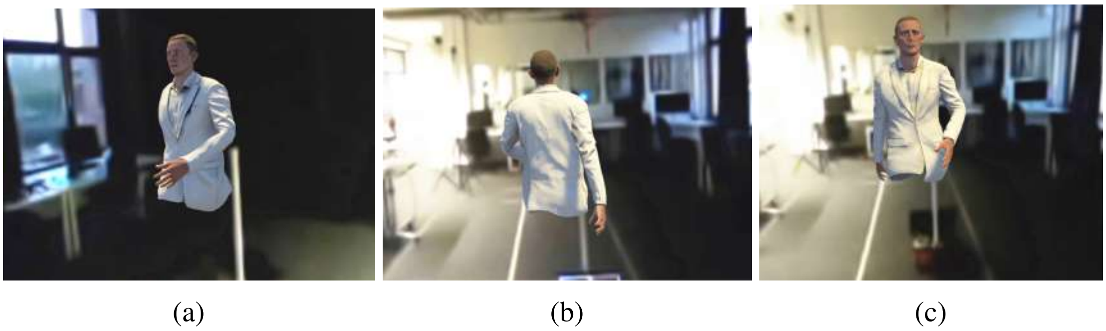

(a) (b) (c)

Fig. 5.16 View of participants during the experiment (images captured through HTC Vive Pro HMD): (a) Avatar arm swing frequency estimation. (b) Walking following an avatar robot. (c) Walking towards an avatar robot.

participants' ID: participants with odd ID started with *"following task"*, while the other participants conducted *"towards task"* firstly.

In the following task, the distance between a moving avatar robot and a following human subject was investigated in a simulated guidance scenario, which was defined as *"trust distance"* in Figure 5.12b. 3 robot moving speeds (0.8m/s, 1.0m/s and 1.2m/s) which has been tested in the first experiment and 2 avatar visibility (visible and invisible) were selected as factors in this task. Before the experiment, participants mounted HMD and holding controllers stood on one end of tracking space, a mixed-reality avatar robot was placed in front of the participants without space remained (Figure 5.13b). When the study began, participants clicked the trigger button on the controllers to awake the avatar robot to move forward in a specific speed. For visible situation, the avatar was displayed 1 second after the robot's departure to avoid incongruities during the robot's acceleration to the respective speed. The avatar arm swing frequency was set based on the results of the first experiment to present the state of robot. Participants were required to perceive the increasing distance between avatar robot and themselves. When the distance satisfied their requirements, they started to walk following the avatar robot (Figure 5.15b). Again, in order to guarantee that for each speed participants had a fixed duration to observe and decide the trust distance, the avatar display time was set to 6 seconds for each trial. After that, the avatar robot stopped automatically. For invisible situation, participants perceived and followed an robot only and adjusted the trust distance between each other in the process. For each trial, the trust distance in every frame was recorded with a frame rate of 75fps from participants' departure to the stop of avatar robot.

In the towards task, the distance where participants changed their directions to avoid a potential collision from a head-on approaching robot was investigated in a simulated encountering scenario, which was defined as *"avoiding distance"* in Figure 5.12c. Similarly, the same factors as the following task were selected: 3 robot moving speeds (0.8m/s, 1.0m/s and 1.2m/s) and 2 avatar visibility (visible and invisible). Before the experiment, participants mounted HMD and holding controllers stood on one end of the tracking space, while a mixed-reality avatar robot was located on the opposite end. When the study began, participants clicked the trigger button on the controllers to awake the avatar robot to move forward in a specific speed. At the same time, participants started walking towards the avatar robot (Figure 5.15c). Again, for visible situation, the avatar was displayed 1 second after the robot's departure and stayed for 6 seconds with corresponding arm swing frequency based on the moving speed. During the process, participants were required to perceive and evaluate the distance between the approaching avatar robot and themselves, and changed their directions in a suitable distance to avoid a potential collision. For invisible situation, participants walked towards an robot only and decided the avoiding distance between each other in the process. For each trial, positions of participants (HMD) and avatar robot (Vive tracker) were recorded in every frame with a frame rate of 75fps from the robot's departure to its stop.

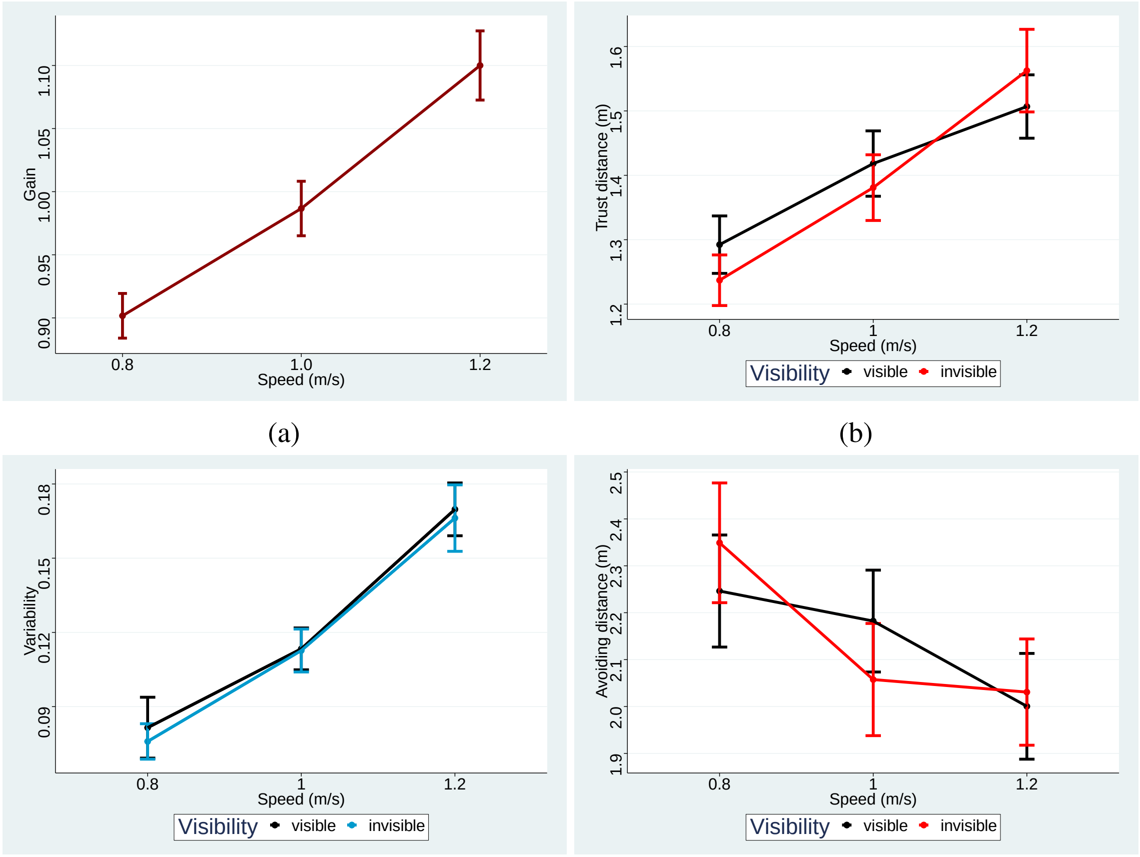
14 participants were invited from local department to take part in the second experiment (ages 20-37, mean age 26.36, SD = 4.765). All participants had normal or corrected to normal vision and most of them (10/14) had prior experience with a mixed reality headset before. In summary, the experiment was a within subjects design with 3 robot moving speeds × 2 avatar visibility × 2 repetitions for a total of 12 trials for each task. 14 participants × 12 trials × 2 tasks per participant = 336 total trials collected. For each task, all of the trials appeared in randomized order. The experiment lasted for around 40 minutes.

Before the experiment, participants were allowed to have some training trials to check if they understood the procedure. After each trial, the robot was manually positioned for the next trial. When the participant and the robot were ready, the operator would give a permission to the participants, then participants could click the trigger button again to begin the next trial. During the experiments, participants were allowed to have a break at any time.

5.2.6 Results

Before the analysis, a normality assumption check was performed for all factor levels using the Shapiro-Wilk test [262], and in a few cases the results did not show a strong indication of normal distribution. However, as shown in previous research [92, 110, 180], moderate deviations from normality can be tolerated by ANOVA.

For arm swing frequency estimation experiment, a plot for robot moving speed on preferred avatar arm swing frequency could be seen in Figure 5.17a. As the only factor tested in this experiment, *robot moving speed* had a significant effect on *preferred avatar arm swing frequency* ($F_{2,28} = 12.58, p < 0.01, \eta^2 = 0.280$). For the *walking following* task



(a)

(b)

(c)

(d)

Fig. 5.17 Results of experiments: (a) Effect of robot moving speeds on preferred avatar arm swing frequency. (b) Trust distances with different robot moving speeds and avatar visibility. (c) Walking variability with different robot moving speeds and avatar visibility. (d) Avoiding distances with different robot moving speeds and avatar visibility. The error bars in figures indicate the standard error (SE).

in proxemics experiment, a plot for robot moving speed and avatar visibility on trust distance could be found in Figure 5.17b. ANOVA results showed that *robot moving speed* had a significant effect on *trust distance* ($F_{2,26} = 44.41, p < 0.01, \eta^2 = 0.568$). However, there was no significant effect from *avatar visibility* on *trust distance*. There was also no significant interaction effect (robot moving speed:avatar visibility) on *trust distance*. In addition, a plot for robot moving speed and avatar visibility on walking variability could be found in Figure 5.17c. ANOVA results suggested that *robot moving speed* had a significant effect on *walking variability* ($F_{2,26} = 33.84, p < 0.01, \eta^2 = 0.542$). No significant effect from *avatar visibility* and no significant interaction effect (robot moving speed:avatar visibility) on *walking variability* was verified. For the *walking towards* task, a plot for robot moving speed and avatar visibility on avoiding distance could be found in Figure 5.17d. ANOVA results represented that *robot moving speed* had a significant effect on *avoiding distance* ($F_{2,26} = 5.084, p < 0.05, \eta^2 = 0.137$). No significant effect from *avatar visibility* and no significant interaction effect (robot moving speed:avatar visibility) on *avoiding distance* was found.

Furthermore, for the significant factor *robot moving speed*, a series of pairwise t tests were conducted in order to check if there was significant difference between each level of the factor. The results of pairwise t tests on *robot moving speed* was presented in Table 5.5.

Table 5.5 Results of pairwise t tests on *robot moving speed*.

Arm Swing Frequency			*Trust Distance*		
	0.8	1.0		0.8	1.0
1.0	0.0031	-	1.0	0.009	-
1.2	6.5e-08	0.0031	1.2	1.3e-06	0.014
Walking Variability			*Avoiding Distance*		
	0.8	1.0		0.8	1.0
1.0	0.00034	-	1.0	0.272	-
1.2	3.3e-12	1.5e-06	1.2	0.054	0.358

In a post-experiment questionnaire, participants were invited to comment on their experience with the avatar robot. Positive comments were mainly on the improvement of the perceptive process and potential use, e.g.: “I trusted the avatar and thus I did not have to look at the floor in order to guess where the robot is. The avatar helped me to avoid collisions and the interaction was more natural to me"; “I like how robots could be represented in future. Very useful for small robots near the ground". 2 participants gave negative feedback by complaining about the resolution of HMD and slight motion sickness due to occasional tracking time delay.

5.2.7 Discussion

Preferred Arm Swing Frequency

The results confirmed the hypotheses to some extent. Figure 5.17a suggested that the preferred avatar arm swing frequency showed a linear correlation with robot moving speed. In addition, standard deviation also increased slightly with the robot moving speed growing. Results suggest that in general human tend to match a faster robot moving speed with a higher avatar arm swing frequency. A possible explanation for this is that as a passive motion of walking gait, arm swing frequency could be also regarded equal to the frequency of legs. When walking in a faster speed, most people preferred to choose keeping their step length but increasing their step frequency as the solution, which conversely caused the results in the arm swing frequency estimation study. Considering the significant effect and the relevant results above, avatar arm swing frequency is an effective method to illustrate the current moving speed of a robot.

Trust Distance and Walking Variability

Analysis of the robot moving speed and avatar visibility on the trust distance (Figure 5.17b) suggests that robot moving speed had a significant effect on the trust distance, however, avatar visibility and interaction effect had no significant influence on the trust distance. Overall, for both avatar visibility (visible and invisible), the trust distance exhibited an increasing trend as the robot moving speed increased, which suggests when following an avatar robot with a faster moving speed, participants tended to choose a larger trust distance between each other. This result has verified the second hypothesis. A potential explanation on this result was due to increasing difficulty in perceiving and predicting the subsequent movements or sudden direction changes of the robot. When it became more difficult and complicated to get accurate perception and prediction on a uncertain situation, people were used to keep more space for safety, which led to a larger trust distance between human and robot in the *walking following* case.

Furthermore, it was worth noticing that the growth of the trust distance for the visible avatar had an obvious deceleration within a robot speed range of 1.0m/s - 1.2m/s, while the trust distance for the invisible avatar showed a nearly linear correlation throughout the whole interval of selected robot speed. This result suggests that within some specific range of robot moving speed, the arm swing animation could help participants to perceive and predict a further motion of robot better and as a result influenced the trust distance to some extent. In addition to this, the standard deviation of trust distance when walking following a robot only in invisible situation became larger with increased robot speed, while when walking following a robot with a visible avatar, the standard deviation of the trust distance did not show an obvious change. This result suggests that participants could better adapt to the change in speed when following a robot with a visible avatar. However, more rigorous study is needed before more strong conclusions are made.

To evaluate the walking stability, participant's walking variability within each trial was analyzed. Variability exhibited a rising trend with the robot moving speed increase (Figure 5.17c). This result suggests that participants had a more stable walking gait when walking following an avatar robot with a lower moving speed. In addition to this, avatar visibility showed a very similar changing trend except for some slight difference in the speed of 0.8m/s and 1.2m/s. The effect of avatar visibility will be discussed in more details in Section 5.2.7 below.

Avoiding Distance

In general, avoiding distance exhibited a decreasing trend with the robot moving speed increasing in Figure 5.17d. One explanation of this phenomenon was that, when walking towards an approaching avatar robot with a faster moving speed, it is more challenging for participants to perceive and predict the subsequent motion of the robot than when the avatar robot in moving with a slower speed. In this case, participants usually needed a longer perceiving and reacting time to predict robot's potential trajectory, such that they could improve the prediction accuracy and make an effective avoiding behavior. Therefore, avoiding distance showed a decreasing trend with the robot moving speed increasing generally.

In addition, in different speed ranges, the effect of avatar visibility on avoiding distance was slightly different. For visible avatar, avoiding distance decreased slowly within the range of 0.8m/s - 1.0m/s but dropped fast within the range of 1.0m/s - 1.2m/s. While for invisible avatar, avoiding distance showed a rapid decrease within 0.8m/s - 1.0m/s but a slow decrease within 1.0m/s - 1.2m/s. One possible reason for this was due to the difference of participants' perceiving ability to the robot moving speed when facing a towards-moving robot with or without an avatar. This conversely suggests that attaching an avatar to the robot to present the current state changes an observer's perceptive ability. However, this analysis would require further verification in future research.

Avatar Visibility

For all the measurements (trust distance, walking variability and avoiding distance) tested above, avatar visibility did not show a significant effect. The potential reason of this result was due to the spatial layout of robot and mixed-reality avatar. In other word, when walking following or towards an avatar robot, what participants focused on and perceived was not only the robot or the avatar, but a cylinder space that consisted of robot and avatar together (Figure 5.18). The proxemics between participants and avatar robot should be decided by this cylinder space which could be also regarded as the real working range or an effective factor that would significantly influence the proxemics during a dynamic interaction between human and avatar robot. According to this analysis and conjecture, in the finished study, the mixed-reality avatar was attached on top of the robot and the size of them were almost the same, which did not have a significant change on the cylinder space comparing with the situation of robot with invisible avatar even though there was slight improvement on perception. Therefore, there was no significant effect of avatar visibility on trust distance, walking variability and avoiding distance found in the study.

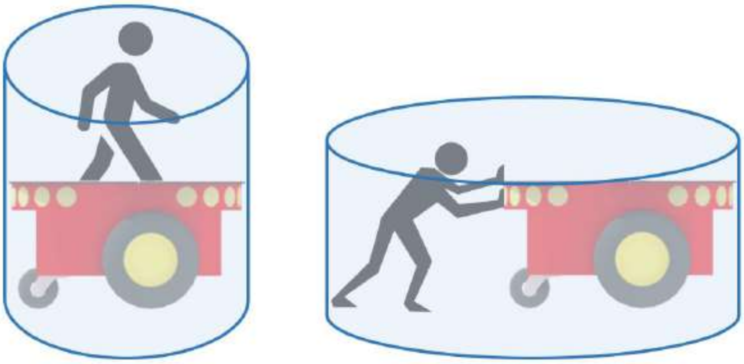


Fig. 5.18 Two cylinder spaces due to different spatial layouts of robot and mixed reality avatar.

Limitations

Normally, arm swing amplitude when walking in different speeds should be slightly different. In the first study, however, arm swing amplitude was fixed to the same level. Arm swing frequency was the only manipulated factor. However, given the limited speed range in the study, the influence of this limitation can be ignored. Another limitation comes from the limited effective tracking range of the tracking system. This tracking limitation was the reason why a longer tracking space was not built for participants to walk inside. A walking space beyond effective tracking range would cause problems like tracking offset or unstable avatar attachment. The limitation of limited testing distance was solved by giving participants enough training trials before the experiment, such that they could get familiar with the avatar robot well in advance and reach a stable level soon after the experiment was started.

5.2.8 Conclusion

A method for using arm swing frequency of a mixed-reality avatar attached with robot was presented to effectively communicate the moving speed. Using this arm swing method, a series of studies were performed to test and evaluate proxemic preferences including trust and avoiding distance between a human and a mixed reality avatar robot during dynamic interaction. The findings suggest that robot moving speed has a significant effect on the proxemics between a human and a mixed-reality avatar robot while avatar visibility did not show a significant influence on the trust and avoiding distance. The data was analyzed and some potential explanations were offered which would be valuable for the future research of mixed reality robotics. In the studies, visual information was designed as the only way to perceive robot motion without taking any other cognitive channels like audio and haptic assistance into consideration, thus the results in this section could be regarded as a conservative reference for some application scenarios, such as using robot for guidance in hospital or for ground cleaning in train station or building corridor.

There are still interesting questions that remain for future work, for example: how could people evaluate and decide proxemics in a more complicated scenario like multiple moving avatar robots? How will the avatar's body posture, facial expression or audio information influence the interactive process? What will be the results if the avatar is separated away from the robot and establish different spatial layouts of robot and avatar? These questions remain outside the scope of this thesis, but might stimulate for future research.

Chapter 6

Conclusion

6.1 Discussion

In this dissertation, we have introduced, evaluated, and discussed telepresence technologies, which enable humans to visit a RE and interact with objects and other people without the requirement for physical travel. Users would ideally feel a sense of presence in the RE during the teleoperation tasks. However, most of today's telepresence systems restrict this sensation as well as other relevant user experience severely, especially with respect to spatial presence, movement control, and natural interaction.

In order to address these challenges, this dissertation introduced a novel solution by integrating 360° cameras and VR HMDs into telepresence systems and exploiting the human user's natural walking as locomotion method for platform control. Based on this setup, several user studies were performed in order to explore the effect of global illumination and shadows of virtual objects on the localization tasks and user experience in 360° MR environments, understand the human perceptive ability to the locomotion manipulations in 360° video-based REs, and investigate the human cognition and social perception with a focus on gaze localization and proxemic preference during interaction with a MR avatar robot.

As explained in Chapter 1, 360° video-based VEs usually lack of depth cues for spatial presence, but can be augmented by virtual objects such as buildings, cars, or avatars to blend real content (from the immersive video) with computer-generated virtual objects. The combination of real and virtual objects in 360° video-based VEs forms new scenarios, which are referred to as 360° MR environments, and therefore, raises the challenge of consistent global illumination. In order to understand the effect of global illumination and shadows of virtual objects on the user's ability to localize objects as well as the overall user experience in 360° MR environments, a user study was performed. The results show the importance of global illumination and shadow effect of virtual objects, which could significantly reduce the localization error and improve user experience, especially in the aspect of experienced realism. Participants also give higher preference evaluations on the experience with global illumination and shadows of virtual objects.

In addition, redirected walking by means of gains for rotation and translation has been introduced into the 360° VR-based telepresence system to enable human users to explore and interact with a much larger RE than the LE. We evaluated the user's ability to detect the manipulations of translation and rotation in 360° video-based REs, in order to guarantee that such redirected manipulations could not be identified by human users to influence the naturalness of interaction. The studies provide detection thresholds of RDW in 360° video-based REs, within which a natural and smooth interaction between the user and RE could be guaranteed. The results show that human users are not able to reliably discriminate the difference between physical motions in the LE and the virtual motions perceived from the 360° video-based RE when virtual translations are down-scaled by 5.8% and up-scaled by 9.7%, and virtual rotations are about 12.3% less or 9.2% more than corresponding physical rotations. This suggests that manipulations in 360° video-based REs have similar influence on users as manipulations in fully computer-generated VEs, i.e., human users tend to perceive shorter distances, but larger rotation angles in the LE as natural. One point to be noticed is that, the ranges of detection thresholds for translation and rotation in 360° video-based REs are both smaller than the results in VEs. These differences indicate that users are better in identifying the difference between physical motions in the LE and corresponding virtual motions in 360° video-based REs than in fully computer-generated VEs. A better experience realism and a higher familiarity in 360° video-based REs could be potential explanations for this phenomenon, but this has to be verified in future research. However, a greater range of gains might also be accepted by users in actual applications without noticing that they are manipulated, because they need to focus on other tasks like object selections, collaborations, etc.

Furthermore, we investigated the human's perception on the social cues expressed by a MR avatar virtually attached on top of a robotic surrogate and its corresponding animations. Two independent user studies were conducted. The first study concentrated on the perception of the MR avatar's gaze direction in RE. Two potential technical solutions, i.e., (i) a stereoscopically rendered 3D avatar in AR, and (ii) a 2D avatar on a constantly host-oriented flattened screen, were proposed to visualize and indicate the teleoperator (guest) gaze for the host when interacting with a 360° video-based telepresence robot. Based on this, the effect of the distance to the robotic surrogate, display technique, rotations of the avatar's head around the pitch and yaw axes on the localization accuracy were evaluated. The results verify that humans perform better in these tasks when interacting with a stereoscopically rendered 3D avatar. In addition, participants preferred to interact with a 3D avatar because of its higher realism. These findings motivated the need for 3D avatar display in the further design of telepresence systems.

The results of the study also show that certain combinations of yaw, pitch and distance can form conditions, in which the host could estimate the avatar gaze direction more accurately. For example, the condition of yaw = 180°, pitch = 0°, distance = 1.2m with the AR display has the least localization error 4.84° among all the tested conditions. However, in this condition, the avatar pose was facing away from the host, which means that eye contact, facial expressions, and, therefore, facial emotions are unavailable to the host in this position. Gaze estimation by the host is only one of the many factors that contribute to an effective telepresence session. It should potentially be left up to the host to choose how he/she wants the robotic surrogate and avatar to "show and behave". In other words, the host should decide a trade-off between an accurate localization and a social interaction by himself/herself depending on different scenarios. If the host is interested in gaze localization, then the host interface should provide some option to focus on that; in which case the robot/avatar should dynamically reposition itself to cater for this. If the host is interested in facial expressions/emotion, then the tablet/avatar should always face the host and display the face of the guest.

The second study focuses on a MR avatar arm swing technique, which subtly communicates the velocity of a robotic surrogate it is attached to. Based on this, a series of studies were conducted to evaluate the effectiveness of this method and the proxemic preference when human users dynamically walk following or towards a robotic surrogate. The results suggest that the moving speed of the robot has a significant effect on the proxemic preference between human users and the MR avatar robot, whereas attaching an avatar on top of the robot did not have a significant effect on the proxemics compared to the baseline situation (robot only). However, participants reported that this method helped to improve perception and prediction on the robot state and gave positive comments regarding its potential applications, for example, with respect to noticing a tiny ground robot.

In our studies, visual information was designed as the only cue to perceive robot motion without taking any other cognitive channels like audio or haptic assistance into consideration, thus, the results in our studies should be regarded as a conservative reference for actual application scenarios, such as using robot for guidance in hospital or for floor cleaning in airport. In summary, this work offers novel guidelines for expressions of the robot state with MR technology to bystanders or collaboration partners.

6.2 Outlook

In this dissertation, essential research questions (see Section 1.4) related to 360° immersive telepresence, human perception in 360° video-based REs and human-robot interaction with a 360° video-based telepresence robot were explored and evaluated. Based on the findings, further research questions have evolved and could be addressed in future studies.

For instance, for the 360° VR-based telepresence system, the usage of a 360° camera and a mobile robotic platform introduces larger latency compared to typical VR environments. For this reason, the resolution of 360° video stream from the RE has to be compressed before being transmitted to the LE, in order to reduce the influence of latency to guarantee a fluent interaction. However, the compression of resolution will lead to an influence on the human perception in 360° video-based REs. Future work could focus on the balance between latency and compressed resolution, in order to provide human users with the best interactive and perceptual experience.

For the interaction with an avatar-robot surrogate, the current study in this dissertation only focuses on a one-to-one situation. However, in some actual scenarios, there would be multiple human users, who interact with multiple avatar-robot surrogates simultaneously. In such a many-to-many situation, perception and cognition of human users on the social cues expressed by avatar-robot surrogates might be different compared to the one-to-one situation. In the application of gaze localization, we assume that for every human user, the localization error would become lower compared with the results in a one-to-one situation. Since in this situation, every user could receive reference and hint from other users to help them fixing the localization error. In the application of proxemics perception, with multiple human users and robotic surrogates, the complexity of a dynamic environment would become higher, which could lead to a larger uncertainty of the human perception on robotic surrogates and more dynamic environments. In this case, the avatar-robot surrogate will not be the only target that human users concern about, the movement trajectories of other users within the same physical space should be taken into account as well. However, these have to be verified in further studies.

Another interesting question arises for situations in which the MR avatar and physical robot could be detached in some specific situations. For example, if the robotic surrogate is equipped with other sensors and could reconstruct a 3D mesh of the RE as it travels along, when the user teleports to a previously visited location, the robotic platform does not need to travel back to that point. In this case, the MR avatar would then be detached from the physical robot, creating a triangle field of "human-robot-avatar", thus introducing entirely new social dynamics in that space and raising further questions.

In summary, the research of human-robot interaction provides new perspectives for our future, in which humans will live and work together with autonomous systems, artificial intelligence, and machines. Therefore, it is important to focus not only on the technical development, but take humans in the focus of design and development, and include an ethical, social, and moral perspective as well in order to ensure efficient, usable, and acceptable HRI.

References

[1] Adalgeirsson, S. O. and Breazeal, C. (2010). Mebot: A robotic platform for socially embodied telepresence. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 15–22. IEEE.

[2] Adams Jr, R. B., Albohn, D. N., and Kveraga, K. (2017). Social vision: applying a social-functional approach to face and expression perception. *Current directions in psychological science*, 26(3):243–248.

[3] Adams Jr, R. B. and Kleck, R. E. (2003). Perceived gaze direction and the processing of facial displays of emotion. *Psychological science*, 14(6):644–647.

[4] Admoni, H. and Scassellati, B. (2017). Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63.

[5] Afzal, S., Chen, J., and Ramakrishnan, K. (2017). Characterization of 360-degree videos. In *Proceedings of the Workshop on Virtual Reality and Augmented Reality Network*, pages 1–6.

[6] Ahn, J.-g. and Kim, G. J. (2016). Remote collaboration using a tele-presence mobile projector robot tele-operated by a smartphone. In *2016 IEEE/SICE International Symposium on System Integration (SII)*, pages 236–241. IEEE.

[7] Akkil, D., James, J. M., Isokoski, P., and Kangas, J. (2016). Gazetorch: Enabling gaze awareness in collaborative physical tasks. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1151–1158. ACM.

[8] Alers, S., Bloembergen, D., Bügler, M., Hennes, D., and Tuyls, K. (2012). Mitro: an augmented mobile telepresence robot with assisted control. In *AAMAS*, pages 1475–1476.

[9] Alsina-Jurnet, I. and Gutiérrez-Maldonado, J. (2010). Influence of personality and individual abilities on the sense of presence experienced in anxiety triggering virtual environments. *International journal of human-computer studies*, 68(10):788–801.

[10] Andrew (2009). Film review: Avatar. https://flicksandbricks.wordpress.com/2017/08/11/film-review-avatar-2009.

[11] Ardouin, J., Lécuyer, A., Marchal, M., and Marchand, E. (2013). Navigating in virtual environments with 360 omnidirectional rendering. In *2013 IEEE Symposium on 3D User Interfaces (3DUI)*, pages 95–98. IEEE.

[12] Ardouin, J., Lécuyer, A., Marchal, M., and Marchand, E. (2014). Stereoscopic rendering of virtual environments with wide field-of-views up to 360. In *2014 IEEE Virtual Reality (VR)*, pages 3–8. IEEE.

[13] Asadi, H., Mohamed, S., Lim, C. P., and Nahavandi, S. (2016). A review on otolith models in human perception. *Behavioural brain research*, 309:67–76.

[14] Asimov, I. (2004). *I, robot*. Spectra.

[15] Aykut, T., Burgmair, C., Karimi, M., Xu, J., and Steinbach, E. (2018a). Delay compensation for actuated stereoscopic 360 degree telepresence systems with probabilistic head motion prediction. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2010–2018. IEEE.

[16] Aykut, T., Karimi, M., Burgmair, C., Finkenzeller, A., Bachhuber, C., and Steinbach, E. (2018b). Delay compensation for a telepresence system with 3d 360 degree vision based on deep head motion prediction and dynamic fov adaptation. *IEEE Robotics and Automation Letters*, 3(4):4343–4350.

[17] Aykut, T., Xu, J., and Steinbach, E. (2019). Realtime 3d 360-degree telepresence with deep-learning-based head-motion prediction. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(1):231–244.

[18] Azuma, R., Neely, H., Daily, M., and Leonard, J. (2006). Performance analysis of an outdoor augmented reality tracking system that relies upon a few mobile beacons. In *Mixed and Augmented Reality, 2006. ISMAR 2006. IEEE/ACM International Symposium on*, pages 101–104. IEEE.

[19] Ballantyne, G. H. (2002). Robotic surgery, telerobotic surgery, telepresence, and telementoring. *Surgical Endoscopy and Other Interventional Techniques*, 16(10):1389–1402.

[20] Ballantyne, G. H. and Moll, F. (2003). The da vinci telerobotic surgical system: the virtual operative field and telepresence surgery. *Surgical Clinics*, 83(6):1293–1304.

[21] Baños, R. M., Botella, C., Alcañiz, M., Liaño, V., Guerrero, B., and Rey, B. (2004). Immersion and emotion: their impact on the sense of presence. *Cyberpsychology & behavior*, 7(6):734–741.

[22] Banton, T., Stefanucci, J., Durgin, F., Fass, A., and Proffitt, D. (2005). The perception of walking speed in a virtual environment. *Presence: Teleoperators & Virtual Environments*, 14(4):394–406.

[23] Barraza, J. F. and Grzywacz, N. M. (2002). Measurement of angular velocity in the perception of rotation. *Vision research*, 42(21):2457–2462.

[24] Bazzano, F., Lamberti, F., Sanna, A., Paravati, G., and Gaspardone, M. (2017). Comparing usability of user interfaces for robotic telepresence. In *VISIGRAPP (2: HUCAPP)*, pages 46–54.

[25] Beck, S., Kunert, A., Kulik, A., and Froehlich, B. (2013). Immersive group-to-group telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 19(4):616–625.

[26] Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403(6767):309–312.

[27] Benson, A. (1970). Interactions between semicircular canals and gravireceptors. In *Recent advances in aerospace medicine*, pages 249–261. Springer.

[28] Bertel, T., Campbell, N. D., and Richardt, C. (2019). Megaparallax: Casual 360° panoramas with motion parallax. *IEEE transactions on visualization and computer graphics*, 25(5):1828–1835.

[29] Berthoz, A. (2000). *The brain's sense of movement*, volume 10. Harvard University Press.

[30] Bijlani, K., Rangan, P. V., Subramanian, S., Vijayan, V., and Jayahari, K. (2010). A-view: Adaptive bandwidth for telepresence and large user sets in live distance education. In *2010 2nd International Conference on Education Technology and Computer*, volume 2, pages V2–219. IEEE.

[31] Boissy, P., Corriveau, H., Michaud, F., Labonté, D., and Royer, M.-P. (2007). A qualitative study of in-home robotic telepresence for home care of community-living elderly subjects. *Journal of telemedicine and telecare*, 13(2):79–84.

[32] Bouguila, L., Ishii, M., and Sato, M. (2002a). Realizing a new step-in-place locomotion interface for virtual environment with large display system. In *Proceedings of the workshop on Virtual environments 2002*, pages 197–207. Eurographics Association.

[33] Bouguila, L., Ishii, M., and Sato, M. (2002b). Virtual locomotion system for human-scale virtual environments. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, pages 227–230. ACM.

[34] Bourke, P. (2016). Converting to/from cubemaps. http://paulbourke.net/miscellaneous/cubemaps.

[35] Bowman, D. A., Koller, D., and Hodges, L. F. (1997). Travel in immersive virtual environments: An evaluation of viewpoint motion control techniques. In *Virtual Reality Annual International Symposium, 1997., IEEE 1997*, pages 45–52. IEEE.

[36] Bowman, D. A. and McMahan, R. P. (2007). Virtual reality: how much immersion is enough? *Computer*, 40(7):36–43.

[37] Breazeal, C. (2003). Emotion and sociable humanoid robots. *International journal of human-computer studies*, 59(1-2):119–155.

[38] Brown, E. and Cairns, P. (2004). A grounded investigation of game immersion. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1297–1300.

[39] Bruder, G., Lubos, P., and Steinicke, F. (2015a). Cognitive resource demands of redirected walking. *IEEE transactions on visualization and computer graphics*, 21(4):539–544.

[40] Bruder, G., Sanz, F. A., Olivier, A.-H., and Lécuyer, A. (2015b). Distance estimation in large immersive projection systems, revisited. In *2015 IEEE Virtual Reality (VR)*, pages 27–32. IEEE.

[41] Bruder, G. and Steinicke, F. (2014). Threefolded motion perception during immersive walkthroughs. In *Proceedings of the 20th ACM symposium on virtual reality software and technology*, pages 177–185.

[42] Bruder, G., Steinicke, F., Wieland, P., and Lappe, M. (2012). Tuning self-motion perception in virtual reality with visual illusions. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1068–1078.

[43] Bryden, M. P. (1963). Ear preference in auditory perception. *Journal of experimental psychology*, 65(1):103.

[44] Buck, L. E., Young, M. K., and Bodenheimer, B. (2018). A comparison of distance estimation in hmd-based virtual environments with different hmd-based conditions. *ACM Transactions on Applied Perception (TAP)*, 15(3):1–15.

[45] Bulich, C., Klein, A., Watson, R., and Kitts, C. (2004). Characterization of delay-induced piloting instability for the triton undersea robot. In *2004 IEEE Aerospace Conference Proceedings (IEEE Cat. No. 04TH8720)*, volume 1. IEEE.

[46] Burridge, R. R. and Hambuchen, K. A. (2009). Using prediction to enhance remote robot supervision across time delay. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5628–5634. IEEE.

[47] Calleja, G. (2011). *In-game: From immersion to incorporation*. mit Press.

[48] Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P. W., Iversen, S. D., and David, A. S. (1997). Activation of auditory cortex during silent lipreading. *science*, 276(5312):593–596.

[49] Card, S. K. (2018). *The psychology of human-computer interaction*. Crc Press.

[50] Cherubini, A., Passama, R., Crosnier, A., Lasnier, A., and Fraisse, P. (2016). Collaborative manufacturing with physical human–robot interaction. *Robotics and Computer-Integrated Manufacturing*, 40:1–13.

[51] Chipps, J., Brysiewicz, P., and Mars, M. (2012). A systematic review of the effectiveness of videoconference-based tele-education for medical and nursing education. *Worldviews on Evidence-Based Nursing*, 9(2):78–87.

[52] Cleveland, W. S., Grosse, E., and Shyu, W. (1992). 1992local regression models. *Statistical models in S, edited by John M. Chambers and Trevor J. Hastie*, pages 309–376.

[53] Collett, T. H. J. and Macdonald, B. A. (2010). An augmented reality debugging system for mobile robot software engineers.

[54] Collins, S. H., Adamczyk, P. G., and Kuo, A. D. (2009). Dynamic arm swinging in human walking. *Proceedings of the Royal Society B: Biological Sciences*, 276(1673):3679–3688.

[55] Colombet, F., Paillot, D., Mérienne, F., and Kemeny, A. (2011). Visual scale factor for speed perception. *Journal of Computing and Information Science in Engineering*, 11(4).

[56] Coltin, B., Biswas, J., Pomerleau, D., and Veloso, M. (2011). Effective semi-autonomous telepresence. In *Robot Soccer World Cup*, pages 365–376. Springer.

[57] Coradeschi, S., Cesta, A., Cortellessa, G., Coraci, L., Gonzalez, J., Karlsson, L., Furfari, F., Loutfi, A., Orlandini, A., Palumbo, F., et al. (2013). Giraffplus: Combining social interaction and long term monitoring for promoting independent living. In *2013 6th international conference on Human System Interactions (HSI)*, pages 578–585. IEEE.

[58] Coradeschi, S., Loutfi, A., Kristoffersson, A., Cortellessa, G., and Eklundh, K. S. (2011). Social robotic telepresence. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 5–6. IEEE.

[59] Corley, A.-M. (2009). The reality of robot surrogates. https://spectrum.ieee.org/robotics/humanoids/the-reality-of-robot-surrogates.

[60] Creem-Regehr, S. H., Stefanucci, J. K., Thompson, W. B., Nash, N., and McCardell, M. (2015). Egocentric distance perception in the oculus rift (dk2). In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception*, pages 47–50.

[61] Creem-Regehr, S. H., Willemsen, P., Gooch, A. A., and Thompson, W. B. (2005). The influence of restricted viewing conditions on egocentric distance perception: Implications for real and virtual indoor environments. *Perception*, 34(2):191–204.

[62] Dael, N., Mortillaro, M., and Scherer, K. R. (2012). Emotion expression in body action and posture. *Emotion*, 12(5):1085.

[63] Dargahi, J. and Najarian, S. (2004). Human tactile perception as a standard for artificial tactile sensing—a review. *The international journal of medical robotics and computer assisted surgery*, 1(1):23–35.

[64] Day, B. L. and Fitzpatrick, R. C. (2005). The vestibular system. *Current biology*, 15(15):R583–R586.

[65] De Gelder, B. (2006). Towards the neurobiology of emotional body language. *Nature Reviews Neuroscience*, 7(3):242–249.

[66] De Greef, L., Morris, M., and Inkpen, K. (2016). Teletourist: Immersive telepresence tourism for mobility-restricted participants. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, pages 273–276. ACM.

[67] Derks, D., Bos, A. E., and Von Grumbkow, J. (2007). Emoticons and social interaction on the internet: the importance of social context. *Computers in human behavior*, 23(1):842–849.

[68] Dichgans, J. and Brandt, T. (1978). Visual-vestibular interaction: Effects on self-motion perception and postural control. In *Perception*, pages 755–804. Springer.

[69] Dodge, R. (1923). Thresholds of rotation. *Journal of Experimental Psychology*, 6(2):107.

[70] Dragone, M., Holz, T., and O'Hare, G. M. (2006). Mixing robotic realities. In *Proceedings of the 11th international conference on Intelligent user interfaces*, pages 261–263. ACM.

[71] Dragone, M., Holz, T., and O'Hare, G. M. (2007). Using mixed reality agents as social interfaces for robots. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*, pages 1161–1166. IEEE.

[72] Draper, J. V., Kaber, D. B., and Usher, J. M. (1998). Telepresence. *Human factors*, 40(3):354–375.

[73] Ebert, R. (2009). Your legs got nothin' to do some machine's doin' that for you. https://www.rogerebert.com/reviews/surrogates-2009.

[74] Edelmann, J., Gerjets, P., Mock, P., Schilling, A., and Strasser, W. (2012). Face2face—a system for multi-touch collaboration with telepresence. In *2012 IEEE International Conference on Emerging Signal Processing Applications*, pages 159–162. IEEE.

[75] Edwards, J. (2011). Telepresence: virtual reality in the real world [special reports]. *IEEE Signal Processing Magazine*, 28(6):9–142.

[76] Engel, D., Curio, C., Tcheang, L., Mohler, B., and Bülthoff, H. H. (2008). A psychophysically calibrated controller for navigating through large environments in a limited free-walking space. In *Proceedings of the 2008 ACM symposium on Virtual reality software and technology*, pages 157–164. ACM.

[77] Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E.-J., and Shadlen, M. N. (1994). fmri of human visual cortex. *Nature*.

[78] Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433.

[79] Escolano, C., Antelis, J. M., and Minguez, J. (2011). A telepresence mobile robot controlled with a noninvasive brain–computer interface. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(3):793–804.

[80] Fan, C.-L., Lee, J., Lo, W.-C., Huang, C.-Y., Chen, K.-T., and Hsu, C.-H. (2017). Fixation prediction for 360 video streaming in head-mounted virtual reality. In *Proceedings of the 27th Workshop on Network and Operating Systems Support for Digital Audio and Video*, pages 67–72.

[81] Fasola, J. and Mataric, M. J. (2012). Using socially assistive human–robot interaction to motivate physical exercise for older adults. *Proceedings of the IEEE*, 100(8):2512–2526.

[82] Fiore, S. M., Wiltshire, T. J., Lobato, E. J., Jentsch, F. G., Huang, W. H., and Axelrod, B. (2013). Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior. *Frontiers in psychology*, 4:859.

[83] Fowler, C. and Mayes, T. (1997). Applying telepresence to education. *BT Technology Journal*, 15(4):188–195.

[84] Frenz, H., Lappe, M., Kolesnik, M., and Bührmann, T. (2007). Estimation of travel distance from visual motion in virtual environments. *ACM Transactions on Applied Perception (TAP)*, 4(1):3.

[85] Frith, C. (2009). Role of facial expressions in social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3453–3458.

[86] Furness, W. B. T. A. (1995). *Virtual environments and advanced interface design*. Oxford University Press on Demand.

[87] Gaemperle, L., Seyid, K., Popovic, V., and Leblebici, Y. (2014). An immersive telepresence system using a real-time omnidirectional camera and a virtual reality head-mounted display. In *2014 IEEE International Symposium on Multimedia*, pages 175–178. IEEE.

[88] Gallace, A. and Spence, C. (2014). *In touch with the future: The sense of touch from cognitive neuroscience to virtual reality*. OUP Oxford.

[89] Garber-Barron, M. and Si, M. (2012). Using body movement and posture for emotion detection in non-acted scenarios. In *2012 IEEE International Conference on Fuzzy Systems*, pages 1–8. IEEE.

[90] Gemmell, J., Toyama, K., Zitnick, C. L., Kang, T., and Seitz, S. (2000). Gaze awareness for video-conferencing: A software approach. *IEEE MultiMedia*, 7(4):26–35.

[91] Giesler, B., Salb, T., Steinhaus, P., and Dillmann, R. (2004). Using augmented reality to interact with an autonomous mobile platform. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, volume 1, pages 1009–1014. IEEE.

[92] Glass, G. V., Peckham, P. D., and Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of educational research*, 42(3):237–288.

[93] Gockley, R., Forlizzi, J., and Simmons, R. (2007). Natural person-following behavior for social robots. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 17–24. ACM.

[94] Goldberg, J. M. and Fernandez, C. (1984). The vestibular system. *Handbook of physiology–the nervous system III. American Physiological Society, Bethesda, Md*, pages 916–977.

[95] Goza, S. M., Ambrose, R. O., Diftler, M. A., and Spain, I. M. (2004). Telepresence control of the nasa/darpa robonaut on a mobility platform. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 623–629.

[96] Grassini, S. and Laumann, K. (2020). Questionnaire measures and physiological correlates of presence: A systematic review. *Frontiers in psychology*, 11:349.

[97] Grau, O. (2003). *Virtual Art: from illusion to immersion*. MIT press.

[98] Grewe, P., Kohsik, A., Flentge, D., Dyck, E., Botsch, M., Winter, Y., Markowitsch, H. J., Bien, C. G., and Piefke, M. (2013). Learning real-life cognitive abilities in a novel 360-virtual reality supermarket: a neuropsychological study of healthy participants and patients with epilepsy. *Journal of neuroengineering and rehabilitation*, 10(1):42.

[99] Grice, P. M. and Kemp, C. C. (2019). In-home and remote use of robotic body surrogates by people with profound motor deficits. *PloS one*, 14(3).

[100] Grigg, P. (1994). Peripheral neural mechanisms in proprioception. *Journal of Sport Rehabilitation*, 3(1):2–17.

[101] Grill-Spector, K. and Malach, R. (2004). The human visual cortex. *Annu. Rev. Neurosci.*, 27:649–677.

[102] Gross, M., Gross, M., Würmlin, S., Naef, M., Lamboray, E., Spagno, C., Kunz, A., Koller-Meier, E., Svoboda, T., Van Gool, L., et al. (2003). blue-c: a spatially immersive display and 3d video portal for telepresence. In *ACM Transactions on Graphics (TOG)*, volume 22, pages 819–827. ACM.

[103] Grunwald, M. (2008). *Human haptic perception: Basics and applications*. Springer Science & Business Media.

[104] Guadarrama, S., Riano, L., Golland, D., Go, D., Jia, Y., Klein, D., Abbeel, P., Darrell, T., et al. (2013). Grounding spatial relations for human-robot interaction. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1640–1647. IEEE.

[105] Hajjaj, S. S. H. and Sahari, K. S. M. (2014). Review of research in the area of agriculture mobile robots. In *The 8th International Conference on Robotic, Vision, Signal Processing & Power Applications*, pages 107–117. Springer.

[106] Hall, E. T., Birdwhistell, R. L., Bock, B., Bohannan, P., Diebold Jr, A. R., Durbin, M., Edmonson, M. S., Fischer, J., Hymes, D., Kimball, S. T., et al. (1968). Proxemics [and comments and replies]. *Current anthropology*, 9(2/3):83–108.

[107] Haller, M., Drab, S., and Hartmann, W. (2003). A real-time shadow approach for an augmented reality application using shadow volumes. In *Proceedings of the ACM symposium on Virtual reality software and technology*, pages 56–65.

[108] Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5):517–527.

[109] Hansen, J. P., Alapetite, A., Thomsen, M., Wang, Z., Minakata, K., and Zhang, G. (2018). Head and gaze control of a telepresence robot with an hmd. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–3.

[110] Harwell, M. R., Rubinstein, E. N., Hayes, W. S., and Olds, C. C. (1992). Summarizing monte carlo results in methodological research: The one-and two-factor fixed effects anova cases. *Journal of educational statistics*, 17(4):315–339.

[111] Hebert, P., Ma, J., Borders, J., Aydemir, A., Bajracharya, M., Hudson, N., Shankar, K., Karumanchi, S., Douillard, B., and Burdick, J. (2015). Supervised remote robot with guided autonomy and teleoperation (surrogate): a framework for whole-body manipulation. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 5509–5516. IEEE.

[112] Hedayati, H., Walker, M., and Szafir, D. (2018). Improving collocated robot teleoperation with augmented reality. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 78–86. ACM.

[113] Held, R. M. and Durlach, N. I. (1992). Telepresence. *Presence: Teleoperators & Virtual Environments*, 1(1):109–112.

[114] Hertel, J. and Steinicke, F. (2021). Augmented reality for maritime navigation assistance-egocentric depth perception in large distance outdoor environments. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 122–130. IEEE.

[115] Heshmat, Y., Jones, B., Xiong, X., Neustaedter, C., Tang, A., Riecke, B. E., and Yang, L. (2018). Geocaching with a beam: Shared outdoor activities through a telepresence robot with 360 degree viewing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 359. ACM.

[116] Hettinger, L. J. and Riccio, G. E. (1992). Visually induced motion sickness in virtual environments. *Presence: Teleoperators & Virtual Environments*, 1(3):306–310.

[117] Hietanen, J. K. (2002). Social attention orienting integrates visual information from head and body orientation. *Psychological Research*, 66(3):174–179.

[118] Hirasaki, E., Moore, S. T., Raphan, T., and Cohen, B. (1999). Effects of walking velocity on vertical head and body movements during locomotion. *Experimental brain research*, 127(2):117–130.

[119] Hodgson, E., Bachmann, E., and Waller, D. (2011). Redirected walking to explore virtual environments: Assessing the potential for spatial interference. *ACM Transactions on Applied Perception (TAP)*, 8(4):22.

[120] Hoenig, W., Milanes, C., Scaria, L., Phan, T., Bolas, M., and Ayanian, N. (2015). Mixed reality for robotics. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5382–5387. IEEE.

[121] Hoffman, G., Zuckerman, O., Hirschberger, G., Luria, M., and Shani Sherman, T. (2015). Design and evaluation of a peripheral robotic conversation companion. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 3–10. ACM.

[122] Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric statistical methods*, volume 751. John Wiley & Sons.

[123] Holz, T., Campbell, A. G., O'Hare, G. M., Stafford, J. W., Martin, A., and Dragone, M. (2011). Mira—mixed reality agents. *International journal of human-computer studies*, 69(4):251–268.

[124] Honig, S. and Oron-Gilad, T. (2018). Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology*, 9:861.

[125] Hosseini, A. and Lienkamp, M. (2016). Enhancing telepresence during the teleoperation of road vehicles using hmd-based mixed reality. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 1366–1373. IEEE.

[126] Hosseini, M. (2017). View-aware tile-based adaptations in 360 virtual reality video streaming. In *2017 IEEE Virtual Reality (VR)*, pages 423–424. IEEE.

[127] Hosseini, M. and Swaminathan, V. (2016). Adaptive 360 vr video streaming: Divide and conquer. In *2016 IEEE International Symposium on Multimedia (ISM)*, pages 107–110. IEEE.

[128] Howard, I. P., Rogers, B. J., et al. (1995). *Binocular vision and stereopsis*. Oxford University Press, USA.

[129] Howe, R. D. and Matsuoka, Y. (1999). Robotics for surgery. *Annual review of biomedical engineering*, 1(1):211–240.

[130] Huang, J., Chen, Z., Ceylan, D., and Jin, H. (2017). 6-dof vr videos with a single 360-camera. In *2017 IEEE Virtual Reality (VR)*, pages 37–44. IEEE.

[131] Ikei, Y., Yem, V., Tashiro, K., Fujie, T., Amemiya, T., and Kitazaki, M. (2019). Live stereoscopic 3d image with constant capture direction of 360° cameras for high-quality visual telepresence. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 431–439. IEEE.

[132] Ikemoto, S., Amor, H. B., Minato, T., Jung, B., and Ishiguro, H. (2012). Physical human-robot interaction: Mutual learning and adaptation. *IEEE robotics & automation magazine*, 19(4):24–35.

[133] Interrante, V., Ries, B., and Anderson, L. (2006). Distance perception in immersive virtual environments, revisited. In *Virtual Reality Conference, 2006*, pages 3–10. IEEE.

[134] Interrante, V., Ries, B., Lindquist, J., Kaeding, M., and Anderson, L. (2008). Elucidating factors that can facilitate veridical spatial perception in immersive virtual environments. *Presence: Teleoperators and Virtual Environments*, 17(2):176–198.

[135] IrisVR (2018). The importance of frame rates. https://help.irisvr.com/hc/en-us/articles/215884547-The-Importance-of-Frame-Rates.

[136] Iwata, H. (1999). The torus treadmill: Realizing locomotion in ves. *IEEE Computer Graphics and Applications*, 19(6):30–35.

[137] Iwata, H., Yano, H., Fukushima, H., and Noma, H. (2005). Circulafloor [locomotion interface]. *IEEE Computer Graphics and Applications*, 25(1):64–67.

[138] Iwata, H., Yano, H., and Tomioka, H. (2006). Powered shoes. In *ACM SIGGRAPH 2006 Emerging technologies*, page 28. ACM.

[139] Jerald, J. (2015). *The VR book: Human-centered design for virtual reality*. Morgan & Claypool.

[140] Jerald, J., Peck, T., Steinicke, F., and Whitton, M. (2008). Sensitivity to scene motion for phases of head yaws. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization*, pages 155–162. ACM.

[141] Johnson, S., Rae, I., Mutlu, B., and Takayama, L. (2015). Can you see me now? how field of view affects collaboration in robotic telepresence. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2397–2406.

[142] Jones, J. A., Swan, J. E., Singh, G., and Ellis, S. R. (2011). Peripheral visual information and its effect on distance judgments in virtual and augmented environments. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization*, pages 29–36.

[143] Kappers, A. M. and Koenderink, J. J. (1999). Haptic perception of spatial relations. *Perception*, 28(6):781–795.

[144] Kasahara, S., Nagai, S., and Rekimoto, J. (2017). Jackin head: Immersive visual telepresence system with omnidirectional wearable camera. *IEEE transactions on visualization and computer graphics*, 23(3):1222–1234.

[145] Kasahara, S. and Rekimoto, J. (2015). Jackin head: An immersive human-human telepresence system. In *SIGGRAPH Asia 2015 Emerging Technologies*, pages 1–3.

[146] Katzakis, N. and Steinicke, F. (2018). Excuse me! perception of abrupt direction changes using body cues and paths on mixed reality avatars. *arXiv preprint arXiv:1801.05085*.

[147] Katzakis, N., Tong, J., Ariza, O., Chen, L., Klinker, G., Röder, B., and Steinicke, F. (2017). Stylo and handifact: Modulating haptic perception through visualizations for posture training in augmented reality. In *Proceedings of the 5th Symposium on Spatial User Interaction*, SUI '17, pages 58–67, New York, NY, USA. ACM.

[148] Kavanagh, S., Luxton-Reilly, A., Wüensche, B., and Plimmer, B. (2016). Creating 360 educational video: A case study. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction*, pages 34–39.

[149] Kawaguchi, I., Kodama, Y., Kuzuoka, H., Otsuki, M., and Suzuki, Y. (2016). Effect of embodiment presentation by humanoid robot on social telepresence. In *Proceedings of the Fourth International Conference on Human Agent Interaction*, pages 253–256. ACM.

[150] Kawaguchi, I., Kuzuoka, H., and Suzuki, Y. (2015). Study on gaze direction perception of face image displayed on rotatable flat display. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1729–1737. ACM.

[151] Kelly, J. W., Cherep, L. A., Klesel, B., Siegel, Z. D., and George, S. (2018). Comparison of two methods for improving distance perception in virtual reality. *ACM Transactions on Applied Perception (TAP)*, 15(2):1–11.

[152] Khurshid, J. and Bing-Rong, H. (2004). Military robots-a glimpse from today and tomorrow. In *ICARCV 2004 8th Control, Automation, Robotics and Vision Conference, 2004.*, volume 1, pages 771–777. IEEE.

[153] Kim, J.-Y., Park, I.-W., Lee, J., Kim, M.-S., Cho, B.-K., and Oh, J.-H. (2005). System design and dynamic walking of humanoid robot khr-2. In *Proceedings of the 2005 IEEE international conference on robotics and automation*, pages 1431–1436. IEEE.

[154] Kim, K., Nagendran, A., Bailenson, J. N., Raij, A., Bruder, G., Lee, M., Schubert, R., Yan, X., and Welch, G. F. (2017). A large-scale study of surrogate physicality and gesturing on human–surrogate interactions in a public space. *Frontiers in Robotics and AI*, 4:32.

[155] Kim, K. and Welch, G. (2015). Maintaining and enhancing human-surrogate presence in augmented reality. In *2015 IEEE International Symposium on Mixed and Augmented Reality Workshops*, pages 15–19. IEEE.

[156] Kim, S. and Dey, A. K. (2009). Simulated augmented reality windshield display as a cognitive mapping aid for elder driver navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 133–142. ACM.

[157] Kim, Y. and Mutlu, B. (2014). How social distance shapes human–robot interaction. *International Journal of Human-Computer Studies*, 72(12):783–795.

[158] Kinsbourne, M. (1974). Direction of gaze and distribution of cerebral thought processes. *Neuropsychologia*, 12(2):279–281.

[159] Knapp, A. (1938). An introduction to clinical perimetry. *Archives of Ophthalmology*, 20(6):1116–1117.

[160] Knapp, J. M. and Loomis, J. M. (2004). Limited field of view of head-mounted displays is not the cause of distance underestimation in virtual environments. *Presence: Teleoperators & Virtual Environments*, 13(5):572–577.

[161] Kober, S. E. and Neuper, C. (2013). Personality and presence in virtual reality: Does their relationship depend on the used presence measure? *International Journal of Human-Computer Interaction*, 29(1):13–25.

[162] Kohli, L., Burns, E., Miller, D., and Fuchs, H. (2005). Combining passive haptics with redirected walking. In *Proceedings of the 2005 international conference on Augmented tele-existence*, pages 253–254. ACM.

[163] Kopper, R., Bowman, D. A., Silva, M. G., and McMahan, R. P. (2010). A human motor behavior model for distal pointing tasks. *International journal of human-computer studies*, 68(10):603–615.

[164] Kostavelis, I., Nalpantidis, L., Boukas, E., Rodrigalvarez, M. A., Stamoulias, I., Lentaris, G., Diamantopoulos, D., Siozios, K., Soudris, D., and Gasteratos, A. (2014). Spartan: Developing a vision system for future autonomous space exploration robots. *Journal of Field Robotics*, 31(1):107–140.

[165] Kristoffersson, A., Coradeschi, S., and Loutfi, A. (2013). A review of mobile robotic telepresence. *Advances in Human-Computer Interaction*, 2013.

[166] Krupke, D., Einig, L., Langbehn, E., Zhang, J., and Steinicke, F. (2016). Immersive remote grasping: realtime gripper control by a heterogenous robot control system. In *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*, pages 337–338. ACM.

[167] Krupke, D., Steinicke, F., Lubos, P., Jonetzko, Y., Görner, M., and Zhang, J. (2018). Comparison of multimodal heading and pointing gestures for co-located mixed reality human-robot interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–9. IEEE.

[168] Kubo, M., Wagenaar, R. C., Saltzman, E., and Holt, K. G. (2004). Biomechanical mechanism for transitions in phase and frequency of arm and leg swing during walking. *Biological cybernetics*, 91(2):91–98.

[169] Langbehn, E., Lubos, P., Bruder, G., and Steinicke, F. (2017). Bending the curve: Sensitivity to bending of curved paths and application in room-scale vr. *IEEE transactions on visualization and computer graphics*, 23(4):1389–1398.

[170] LaViola Jr, J. J. (2000). A discussion of cybersickness in virtual environments. *ACM SIGCHI Bulletin*, 32(1):47–56.

[171] Lederman, S. J. and Klatzky, R. L. (2009). Haptic perception: A tutorial. *Attention, Perception, & Psychophysics*, 71(7):1439–1459.

[172] Lee, G. A., Teo, T., Kim, S., and Billinghurst, M. (2017). Mixed reality collaboration through sharing a live panorama. In *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications*, page 14. ACM.

[173] Lee, K. M., Peng, W., Jin, S.-A., and Yan, C. (2006). Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of communication*, 56(4):754–772.

[174] Leeb, R., Tonin, L., Rohm, M., Desideri, L., Carlson, T., and Millan, J. d. R. (2015). Towards independence: a bci telepresence robot for people with severe motor disabilities. *Proceedings of the IEEE*, 103(6):969–982.

[175] Lei, M., Clemente, I. M., and Hu, Y. (2019). Student in the shell: the robotic body and student engagement. *Computers & Education*, 130:59–80.

[176] Leithinger, D., Follmer, S., Olwal, A., and Ishii, H. (2014). Physical telepresence: shape capture and display for embodied, computer-mediated remote collaboration. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 461–470. ACM.

[177] Li, B., Zhang, R., Nordman, A., and Kuhl, S. A. (2015). The effects of minification and display field of view on distance judgments in real and hmd-based environments. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception*, pages 55–58.

[178] Liao, W.-S., Hsieh, T.-J., Liang, W.-Y., Chang, Y.-L., Chang, C.-H., and Chen, W.-Y. (2011). Real-time spherical panorama image stitching using opencl. In *2011 International Conference on Computer Graphics and Virtual Reality*, pages 113–119. Citeseer.

[179] Liu, C., Ishi, C. T., Ishiguro, H., and Hagita, N. (2013). Generation of nodding, head tilting and gazing for human–robot speech interaction. *International Journal of Humanoid Robotics*, 10(01):1350009.

[180] Lix, L. M., Keselman, J. C., and Keselman, H. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance f test. *Review of educational research*, 66(4):579–619.

[181] Loomis, J. M. and Knapp, J. M. (2003). Visual perception of egocentric distance in real and virtual environments. *Virtual and adaptive environments*, 11:21–46.

[182] Luévano, E., de Lara, E. L., and Castro, J. E. (2015). Use of telepresence and holographic projection mobile device for college degree level. *Procedia Computer Science*, 75:339–347.

[183] Mackey, B. A., Bremner, P. A., and Giuliani, M. (2020). The effect of virtual reality control of a robotic surrogate on presence and social presence in comparison to telecommunications software. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 349–351.

[184] Marschner, L., Pannasch, S., Schulz, J., and Graupner, S.-T. (2015). Social communication with virtual agents: The effects of body and gaze direction on attention and emotional responding in human observers. *International Journal of Psychophysiology*, 97(2):85–92.

[185] Matsumoto, K., Ban, Y., Narumi, T., Yanase, Y., Tanikawa, T., and Hirose, M. (2016). Unlimited corridor: redirected walking techniques using visuo haptic interaction. In *ACM SIGGRAPH 2016 Emerging Technologies*, page 20. ACM.

[186] Mayer, H., Nagy, I., Knoll, A., Braun, E. U., Bauernschmitt, R., and Lange, R. (2007). Haptic feedback in a telepresence system for endoscopic heart surgery. *Presence: Teleoperators and Virtual Environments*, 16(5):459–470.

[187] McCall, C. (2015). Mapping social interactions: the science of proxemics. In *Social Behavior from Rodents to Humans*, pages 295–308. Springer.

[188] McGinity, M., Shaw, J., Kuchelmeister, V., Hardjono, A., and Favero, D. D. (2007). Avie: a versatile multi-user stereo 360 interactive vr theatre. In *Proceedings of the 2007 workshop on Emerging displays technologies: images and beyond: the future of displays and interacton*, pages 2–es.

[189] McMahan, A. (2013). Immersion, engagement, and presence: A method for analyzing 3-d video games. In *The video game theory reader*, pages 89–108. Routledge.

[190] Mead, R. and Matarić, M. J. (2016). Perceptual models of human-robot proxemics. In *Experimental Robotics*, pages 261–276. Springer.

[191] Mead, R. and Matarić, M. J. (2017). Autonomous human–robot proxemics: socially aware navigation based on interaction potential. *Autonomous Robots*, 41(5):1189–1201.

[192] Mergner, T. and Becker, W. (1990). Perception of horizontal self-rotation: Multisensory and cognitive aspects. *Perception and control of self-motion*, pages 219–263.

[193] Mergner, T., Schweigart, G., Müller, M., Hlavacka, F., and Becker, W. (2000). Visual contributions to human self-motion perception during horizontal body rotation. *Archives italiennes de biologie*, 138(2):139–166.

[194] Messing, R. and Durgin, F. (2004). Space perception and cues to distance in virtual reality. In *Proceedings of the 1st Symposium on Applied perception in graphics and visualization*, pages 176–176.

[195] Messing, R. and Durgin, F. H. (2005). Distance perception and the visual horizon in head-mounted displays. *ACM Transactions on Applied Perception (TAP)*, 2(3):234–250.

[196] Meyns, P., Bruijn, S. M., and Duysens, J. (2013). The how and why of arm swing during human walking. *Gait & posture*, 38(4):555–562.

[197] Michaud, F., Boissy, P., Labonte, D., Corriveau, H., Grant, A., Lauria, M., Cloutier, R., Roux, M.-A., Iannuzzi, D., and Royer, M.-P. (2007). Telepresence robot for home care assistance. In *AAAI spring symposium: multidisciplinary collaboration for socially assistive robotics*, pages 50–55. California, USA.

[198] Milgram, P. and Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems*, 77(12):1321–1329.

[199] Minsky, M. (1980). Telepresence.

[200] Misawa, K., Ishiguro, Y., and Rekimoto, J. (2013). Livemask: A telepresence surrogate system with a face-shaped screen for supporting nonverbal communication. *Information and Media Technologies*, 8(2):617–625.

[201] Miwa, H., Itoh, K., Matsumoto, M., Zecca, M., Takanobu, H., Rocella, S., Carrozza, M. C., Dario, P., and Takanishi, A. (2004). Effective emotional expressions with expression humanoid robot we-4rii: integration of humanoid robot hand rch-1. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 3, pages 2203–2208. IEEE.

[202] Mohler, B. J., Thompson, W. B., Creem-Regehr, S. H., Pick, H. L., and Warren, W. H. (2007). Visual flow influences gait transition speed and preferred walking speed. *Experimental brain research*, 181(2):221–228.

[203] Moubayed, S. A., Skantze, G., and Beskow, J. (2013). The furhat back-projected humanoid head–lip reading, gaze and multi-party interaction. *International Journal of Humanoid Robotics*, 10(01):1350005.

[204] Mubin, O., Stevens, C. J., Shahid, S., Al Mahmud, A., and Dong, J.-J. (2013). A review of the applicability of robots in education. *Journal of Technology in Education and Learning*, 1(209-0015):13.

[205] Müller, J., Rädle, R., and Reiterer, H. (2016). Virtual objects as spatial cues in collaborative mixed reality environments: How they shape communication behavior and user task load. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1245–1249.

[206] Mumm, J. and Mutlu, B. (2011). Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 331–338. ACM.

[207] Murai, R., Sakai, T., Kawano, H., Matsukawa, Y., Kitano, Y., Honda, Y., and Campbell, K. C. (2012). A novel visible light communication system for enhanced control of autonomous delivery robots in a hospital. In *2012 IEEE/SICE International Symposium on System Integration (SII)*, pages 510–516. IEEE.

[208] Myers, B. A., Bhatnagar, R., Nichols, J., Peck, C. H., Kong, D., Miller, R., and Long, A. C. (2002). Interacting at a distance: measuring the performance of laser pointers and other devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 33–40. ACM.

[209] Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., and Winkler, I. (2001). 'primitive intelligence'in the auditory cortex. *Trends in neurosciences*, 24(5):283–288.

[210] Nagendran, A., Steed, A., Kelly, B., and Pan, Y. (2015). Symmetric telepresence using robotic humanoid surrogates. *Computer Animation and Virtual Worlds*, 26(3-4):271–280.

[211] Nakanishi, H., Murakami, Y., Nogami, D., and Ishiguro, H. (2008). Minimum movement matters: impact of robot-mounted cameras on social telepresence. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 303–312.

[212] Neth, C. T., Souman, J. L., Engel, D., Kloos, U., Bulthoff, H. H., and Mohler, B. J. (2012). Velocity-dependent dynamic curvature gain for redirected walking. *IEEE transactions on visualization and computer graphics*, 18(7):1041–1052.

[213] Newhart, V. A. and Olson, J. S. (2017). My student is a robot: How schools manage telepresence experiences for students. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 342–347.

[214] Nguyen, D. and Canny, J. (2005). Multiview: spatially faithful group video conferencing. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 799–808. ACM.

[215] Nguyen, T. N., Nguyen, H. T., et al. (2015). Real-time transmission of panoramic images for a telepresence wheelchair. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3565–3568. IEEE.

[216] Nilsson, N. C., Serafin, S., and Nordahl, R. (2014). Establishing the range of perceptually natural visual walking speeds for virtual walking-in-place locomotion. *IEEE transactions on visualization and computer graphics*, 20(4):569–578.

[217] Nitzsche, N., Hanebeck, U. D., and Schmidt, G. (2004). Motion compression for telepresent walking in large target environments. *Presence: Teleoperators & Virtual Environments*, 13(1):44–60.

[218] Noh, Z. and Sunar, M. S. (2009). A review of shadow techniques in augmented reality. In *2009 Second International Conference on Machine Vision*, pages 320–324. IEEE.

[219] Nooij, S. A., Nesti, A., Bülthoff, H. H., and Pretto, P. (2016). Perception of rotation, path, and heading in circular trajectories. *Experimental brain research*, 234(8):2323–2337.

[220] Oculus (2018). Guidelines for vr performance optimization. https://developer.oculus.com/documentation/native/pc/dg-performance-guidelines.

[221] Oh, Y., Parasuraman, R., McGraw, T., and Min, B.-C. (2018). 360 vr based robot teleoperation interface for virtual tour. In *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for Human-Robot Interactions (VAM-HRI)*, pages 78–82. ACM/IEEE.

[222] Okura, F., Kanbara, M., and Yokoya, N. (2012). Fly-through heijo palace site: historical tourism system using augmented telepresence. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1283–1284. ACM.

[223] Onix-Systems (2018). How to use 360° equirectangular panoramas for greater realism in games. https://medium.com/@onix_systems/how-to-use-360-equirectangular-panoramas-for-greater-realism-in-games-55fadb0547da.

[224] Orts-Escolano, S., Rhemann, C., Fanello, S., Chang, W., Kowdle, A., Degtyarev, Y., Kim, D., Davidson, P. L., Khamis, S., Dou, M., et al. (2016). Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 741–754. ACM.

[225] Otsuki, M., Kawano, T., Maruyama, K., Kuzuoka, H., and Suzuki, Y. (2017). Thirdeye: Simple add-on display to represent remote participant's gaze direction in video communication. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5307–5312. ACM.

[226] Oyekoya, O., Steptoe, W., and Steed, A. (2012). Sphereavatar: a situated display to represent a remote collaborator. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2551–2560. ACM.

[227] Paludan, A., Elbaek, J., Mortensen, M., Zobbe, M., Nilsson, N. C., Nordahl, R., Reng, L., and Serafin, S. (2016). Disguising rotational gain for redirected walking in virtual reality: Effect of visual density. In *Virtual Reality (VR), 2016 IEEE*, pages 259–260. IEEE.

[228] Pan, Y. and Steed, A. (2014). A gaze-preserving situated multiview telepresence system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2173–2176. ACM.

[229] Pan, Y. and Steed, A. (2016). Effects of 3d perspective on head gaze estimation with a multiview autostereoscopic display. *International Journal of Human-Computer Studies*, 86:138–148.

[230] Park, I.-W., Kim, J.-Y., Lee, J., and Oh, J.-H. (2005). Mechanical design of humanoid robot platform khr-3 (kaist humanoid robot 3: Hubo). In *5th IEEE-RAS International Conference on Humanoid Robots, 2005.*, pages 321–326. IEEE.

[231] Peck, T. C., Fuchs, H., and Whitton, M. C. (2009). Evaluation of reorientation techniques and distractors for walking in large virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, 15(3):383–394.

[232] Peck, T. C., Fuchs, H., and Whitton, M. C. (2011). An evaluation of navigational ability comparing redirected free exploration with distractors to walking-in-place and joystick locomotio interfaces. In *Virtual Reality Conference (VR), 2011 IEEE*, pages 55–62. IEEE.

[233] Pedersen, L., Kortenkamp, D., Wettergreen, D., Nourbakhsh, I., and Korsmeyer, D. (2003). A survey of space robotics.

[234] Pejsa, T., Kantor, J., Benko, H., Ofek, E., and Wilson, A. (2016). Room2room: Enabling life-size telepresence in a projected augmented reality environment. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pages 1716–1725.

[235] Pessoa, S., Moura, G., Lima, J., Teichrieb, V., and Kelner, J. (2010). Photorealistic rendering for augmented reality: A global illumination and brdf solution. In *2010 IEEE Virtual Reality Conference (VR)*, pages 3–10. IEEE.

[236] Petrangeli, S., Swaminathan, V., Hosseini, M., and De Turck, F. (2017). An http/2-based adaptive streaming framework for 360 virtual reality videos. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 306–314.

[237] Phillips, L., Ries, B., Kaeding, M., and Interrante, V. (2010). Avatar self-embodiment enhances distance perception accuracy in non-photorealistic immersive virtual environments. In *2010 IEEE Virtual Reality Conference (VR)*, pages 115–1148. IEEE.

[238] Pierwola, M. (2015). How 360 video works - explanation 01. http://worldin360.com/how-360-video-works/how-explanation-01.

[239] Pin, F. G. and Killough, S. M. (1994). A new family of omnidirectional and holonomic wheeled platforms for mobile robots. *IEEE transactions on robotics and automation*, 10(4):480–489.

[240] Piryankova, I. V., De La Rosa, S., Kloos, U., Bülthoff, H. H., and Mohler, B. J. (2013). Egocentric distance perception in large screen immersive displays. *Displays*, 34(2):153–164.

[241] Piumsomboon, T., Day, A., Ens, B., Lee, Y., Lee, G., and Billinghurst, M. (2017). Exploring enhancements for remote mixed reality collaboration. In *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications*, page 16. ACM.

[242] Piumsomboon, T., Lee, G. A., Hart, J. D., Ens, B., Lindeman, R. W., Thomas, B. H., and Billinghurst, M. (2018). Mini-me: An adaptive avatar for mixed reality remote collaboration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 46. ACM.

[243] Plumert, J. M., Kearney, J. K., Cremer, J. F., and Recker, K. (2005). Distance perception in real and virtual environments. *ACM Transactions on Applied Perception (TAP)*, 2(3):216–233.

[244] Pretto, P., Bresciani, J.-P., Rainer, G., and Bülthoff, H. H. (2012). Foggy perception slows us down. *Elife*, 1:e00031.

[245] Pretto, P., Ogier, M., Bülthoff, H. H., and Bresciani, J.-P. (2009). Influence of the size of the field of view on motion perception. *Computers & Graphics*, 33(2):139–146.

[246] Qian, F., Ji, L., Han, B., and Gopalakrishnan, V. (2016). Optimizing 360 video delivery over cellular networks. In *Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges*, pages 1–6.

[247] Rabbitt, R. D., Damiano, E. R., and Grant, J. W. (2004). Biomechanics of the semicircular canals and otolith organs. In *The vestibular system*, pages 153–201. Springer.

[248] Ragan, E. D., Wilkes, C., Bowman, D. A., and Höllerer, T. (2009). Simulation of augmented reality systems in purely virtual environments. *2009 IEEE Virtual Reality Conference*, pages 287–288.

[249] Razzaque, S. (2005). *Redirected walking*. University of North Carolina at Chapel Hill.

[250] Razzaque, S., Kohn, Z., and Whitton, M. C. (2001). Redirected walking. In *Proceedings of EUROGRAPHICS*, volume 9, pages 105–106. Manchester, UK.

[251] Reader, S. (2019). A reading list inspired by the 20th anniversary of the matrix (and that sequel announcement). https://www.barnesandnoble.com/blog/sci-fi-fantasy/a-reading-list-inspired-by-the-20th-anniversary-of-the-matrix-and-that-sequel-announcement.

[252] Renner, P. and Pfeiffer, T. (2017). Attention guiding techniques using peripheral vision and eye tracking for feedback in augmented-reality-based assistance systems. In *3D User Interfaces (3DUI), 2017 IEEE Symposium on*, pages 186–194. IEEE.

[253] Rhee, T., Chalmers, A., Hicks, M., Kumagai, K., Allen, B., Loh, I., Petikam, L., and Anjyo, K. (2018). Mr360 interactive: playing with digital creatures in 360° videos. In *SIGGRAPH Asia 2018 Virtual & Augmented Reality*, pages 1–2.

[254] Rhee, T., Petikam, L., Allen, B., and Chalmers, A. (2017). Mr360: Mixed reality rendering for 360 panoramic videos. *IEEE transactions on visualization and computer graphics*, 23(4):1379–1388.

[255] Rich, C., Ponsler, B., Holroyd, A., and Sidner, C. L. (2010). Recognizing engagement in human-robot interaction. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 375–382. IEEE.

[256] Riecke, B. E., Bodenheimer, B., McNamara, T. P., Williams, B., Peng, P., and Feuereissen, D. (2010). Do we need to walk for effective virtual reality navigation? physical rotations alone may suffice. In *International Conference on Spatial Cognition*, pages 234–247. Springer.

[257] Ries, B., Interrante, V., Kaeding, M., and Anderson, L. (2008). The effect of self-embodiment on distance perception in immersive virtual environments. In *Proceedings of the 2008 ACM symposium on Virtual reality software and technology*, pages 167–170.

[258] Riva, G., Mantovani, F., Capideville, C. S., Preziosa, A., Morganti, F., Villani, D., Gaggioli, A., Botella, C., and Alcañiz, M. (2007). Affective interactions using virtual reality: the link between presence and emotions. *CyberPsychology & Behavior*, 10(1):45–56.

[259] Rosenthal, S., Biswas, J., and Veloso, M. (2010). An effective personal mobile robot agent through symbiotic human-robot interaction. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 915–922. International Foundation for Autonomous Agents and Multiagent Systems.

[260] Rößler, P., Beutler, F., Hanebeck, U. D., and Nitzsche, N. (2005). Motion compression applied to guidance of a mobile teleoperator. In *Proceedings of the 2005 IEEE International Conference on Intelligent Robots and Systems (IROS 2005)*, pages 2495–2500.

[261] Rousset, T., Bourdin, C., Goulon, C., Monnoyer, J., and Vercher, J.-L. (2015). Does virtual reality affect visual perception of egocentric distance? In *2015 IEEE Virtual Reality (VR)*, pages 277–278. IEEE.

[262] Royston, J. P. (1982). An extension of shapiro and wilk's w test for normality to large samples. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(2):115–124.

[263] Ryskeldiev, B., Cohen, M., and Herder, J. (2018). Streamspace: Pervasive mixed reality telepresence for remote collaboration on mobile devices. *Journal of Information Processing*, 26:177–185.

[264] Sadowski, W. and Stanney, K. (2002). Presence in virtual environments.

[265] Sauppé, A. and Mutlu, B. (2014). How social cues shape task coordination and communication. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 97–108.

[266] Schmidt, S., Nunez, O. J. A., and Steinicke, F. (2019). Blended agents: Manipulation of physical objects within mixed reality environments and beyond. In *Symposium on Spatial User Interaction*, pages 1–10.

[267] Schmidt, S., Steinicke, F., Huang, T., and Dey, A. (2017). A projection-based augmented reality setup for blended museum experiences. In *ICAT-EGVE (Posters and Demos)*, pages 5–6.

[268] Schubert, T., Friedmann, F., and Regenbrecht, H. (2001). Igroup presence questionnaire.

[269] Schuchardt, P. and Bowman, D. A. (2007). The benefits of immersion for spatial understanding of complex underground cave systems. In *Proceedings of the 2007 ACM symposium on Virtual reality software and technology*, pages 121–124.

[270] Schuemie, M. J., Van Der Straaten, P., Krijn, M., and Van Der Mast, C. A. (2001). Research on presence in virtual reality: A survey. *CyberPsychology & Behavior*, 4(2):183–201.

[271] Schwaiger, M., Thümmel, T., and Ulbrich, H. (2007). Cyberwalk: Implementation of a ball bearing platform for humans. *Human-Computer Interaction. Interaction Platforms and Techniques*, pages 926–935.

[272] Shamsuddin, S., Yussof, H., Ismail, L., Hanapiah, F. A., Mohamed, S., Piah, H. A., and Zahari, N. I. (2012). Initial response of autistic children in human-robot interaction therapy with humanoid robot nao. In *2012 IEEE 8th International Colloquium on Signal Processing and its Applications*, pages 188–193. IEEE.

[273] Sheridan, T. B. (1989). Telerobotics. *Automatica*, 25(4):487–507.

[274] Sheridan, T. B. (2016). Human–robot interaction: status and challenges. *Human factors*, 58(4):525–532.

[275] Shiarlis, K., Messias, J., and Whiteson, S. (2017). Acquiring social interaction behaviours for telepresence robots via deep learning from demonstration. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 37–42. IEEE.

[276] Shin, D. (2018). Empathy and embodied experience in virtual environment: To what extent can virtual reality stimulate empathy and embodied experience? *Computers in Human Behavior*, 78:64–73.

[277] Simon-Thomas, E. R., Keltner, D. J., Sauter, D., Sinicropi-Yao, L., and Abramson, A. (2009). The voice conveys specific emotions: Evidence from vocal burst displays. *Emotion*, 9(6):838.

[278] Singhal, S. and Neustaedter, C. (2017). Bewithme: An immersive telepresence system for distance separated couples. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 307–310. ACM.

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

[280] Slater, M. and Usoh, M. (1993). Presence in immersive virtual environments. In *Proceedings of IEEE Virtual Reality Annual International Symposium*, pages 90–96. IEEE.

[281] Smith, P. and Zheng, Y. (2013). From ear to uncertainty: vestibular contributions to cognitive function. *Frontiers in integrative neuroscience*, 7:84.

[282] Sodhi, R. S., Jones, B. R., Forsyth, D., Bailey, B. P., and Maciocci, G. (2013). Bethere: 3d mobile collaboration with spatial input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 179–188.

[283] Standaert, W., Muylle, S., and Basu, A. (2013). Assessing the effectiveness of telepresence for business meetings. In *2013 46th Hawaii International Conference on System Sciences*, pages 549–558. IEEE.

[284] Stark, L. W., Privitera, C. M., Yang, H., Azzariti, M., Ho, Y. F., Blackmon, T. T., and Chernyak, D. A. (2001). Representation of human vision in the brain: How does human perception recognize images? *Journal of Electronic Imaging*, 10(1):123–152.

[285] Steed, A., Steptoe, W., Oyekoya, W., Pece, F., Weyrich, T., Kautz, J., Friedman, D., Peer, A., Solazzi, M., Tecchia, F., et al. (2012). Beaming: an asymmetric telepresence system. *IEEE computer graphics and applications*, 32(6):10–17.

[286] Steinicke, F., Bruder, G., Hinrichs, K., Lappe, M., Ries, B., and Interrante, V. (2009). Transitional environments enhance distance perception in immersive virtual reality systems. In *Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization*, pages 19–26.

[287] Steinicke, F., Bruder, G., Jerald, J., Frenz, H., and Lappe, M. (2008a). Analyses of human sensitivity to redirected walking. In *Proceedings of the 2008 ACM symposium on Virtual reality software and technology*, pages 149–156. ACM.

[288] Steinicke, F., Bruder, G., Jerald, J., Frenz, H., and Lappe, M. (2010). Estimation of detection thresholds for redirected walking techniques. *IEEE transactions on visualization and computer graphics*, 16(1):17–27.

[289] Steinicke, F., Bruder, G., Kohli, L., Jerald, J., and Hinrichs, K. (2008b). Taxonomy and implementation of redirection techniques for ubiquitous passive haptic feedback. In *Cyberworlds, 2008 International Conference on*, pages 217–223. IEEE.

[290] Steinicke, F., Bruder, G., Ropinski, T., and Hinrichs, K. (2008c). Moving towards generally applicable redirected walking. In *Proceedings of the Virtual Reality International Conference (VRIC)*, pages 15–24. IEEE Press.

[291] Steinicke, F., Hinrichs, K., and Ropinski, T. (2005). Virtual reflections and virtual shadows in mixed reality environments. In *IFIP Conference on Human-Computer Interaction*, pages 1018–1021. Springer.

[292] Steinicke, F., Visell, Y., Campos, J., and Lécuyer, A. (2013). *Human walking in virtual environments*. Springer.

[293] Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. *Journal of communication*, 42(4):73–93.

[294] Stilman, M., Michel, P., Chestnutt, J., Nishiwaki, K., Kagami, S., and Kuffner, J. (2005). Augmented reality for robot development and experimentation. *Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-05-55*, 2(3).

[295] Suma, E. A., Bruder, G., Steinicke, F., Krum, D. M., and Bolas, M. (2012). A taxonomy for deploying redirection techniques in immersive virtual environments. In *Virtual Reality Short Papers and Posters (VRW), 2012 IEEE*, pages 43–46. IEEE.

[296] Suma, E. A., Clark, S., Finkelstein, S. L., and Wartell, Z. (2010). Exploiting change blindness to expand walkable space in a virtual environment. In *Virtual Reality Conference (VR), 2010 IEEE*, pages 305–306. IEEE.

[297] Suma, E. A., Clark, S., Krum, D., Finkelstein, S., Bolas, M., and Warte, Z. (2011). Leveraging change blindness for redirection in virtual environments. In *Virtual Reality Conference (VR), 2011 IEEE*, pages 159–166. IEEE.

[298] Tachi, S. (2015). Telexistence. In *Virtual Realities*, pages 229–259. Springer.

[299] Tachi, S. (2016). Telexistence: Enabling humans to be virtually ubiquitous. *IEEE computer graphics and applications*, 36(1):8–14.

[300] Tanaka, F., Takahashi, T., Matsuzoe, S., Tazawa, N., and Morita, M. (2013). Child-operated telepresence robot: A field trial connecting classrooms between australia and japan. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5896–5901. IEEE.

[301] Tanaka, F., Takahashi, T., Matsuzoe, S., Tazawa, N., and Morita, M. (2014). Telepresence robot helps children in communicating with teachers who speak a different language. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 399–406.

[302] Tanaka, N. and Takagi, H. (2004). Virtual reality environment design of managing both presence and virtual reality sickness. *Journal of physiological anthropology and applied human science*, 23(6):313–317.

[303] Tecchia, F., Alem, L., and Huang, W. (2012). 3d helping hands: a gesture based mr system for remote collaboration. In *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*, pages 323–328.

[304] Tee, K. P., Yan, R., Chua, Y., Huang, Z., and Liemhetcharat, S. (2014). Gesture-based attention direction for a telepresence robot: Design and experimental study. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4090–4095. IEEE.

[305] Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582):520–522.

[306] Tidoni, E., Abu-Alqumsan, M., Leonardis, D., Kapeller, C., Fusco, G., Guger, C., Hintermüller, C., Peer, A., Frisoli, A., Tecchia, F., et al. (2016). Local and remote cooperation with virtual and robotic agents: a p300 bci study in healthy and people living with spinal cord injury. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(9):1622–1632.

[307] Toris, R., Kammerl, J., Lu, D. V., Lee, J., Jenkins, O. C., Osentoski, S., Wills, M., and Chernova, S. (2015). Robot web tools: Efficient messaging for cloud robotics. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4530–4537. IEEE.

[308] Tuli, T. B., Terefe, T. O., and Rashid, M. M. U. (2020). Telepresence mobile robots design and control for social interaction. *International Journal of Social Robotics*, pages 1–10.

[309] Tunstel, E., Maimone, M., Trebi-Ollennu, A., Yen, J., Petras, R., and Willson, R. (2005). Mars exploration rover mobility and robotic arm operational performance. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, volume 2, pages 1807–1814. IEEE.

[310] Vasic, M. and Billard, A. (2013). Safety issues in human-robot interactions. In *2013 ieee international conference on robotics and automation*, pages 197–204. IEEE.

[311] Vasylevska, K., Kaufmann, H., Bolas, M., and Suma, E. A. (2013). Flexible spaces: A virtual step outside of reality. In *Virtual Reality (VR), 2013 IEEE*, pages 109–110. IEEE.

[312] Vasylevska, K., Podkosova, I., and Kaufmann, H. (2015). Walking in virtual reality: Flexible spaces and other techniques. In *The Visual Language of Technique*, pages 81–97. Springer.

[313] Vespa, P. M., Miller, C., Hu, X., Nenov, V., Buxey, F., and Martin, N. A. (2007). Intensive care unit robotic telepresence facilitates rapid physician response to unstable patients and decreased cost in neurointensive care. *Surgical neurology*, 67(4):331–337.

[314] Villani, D., Repetto, C., Cipresso, P., and Riva, G. (2012). May i experience more presence in doing the same thing in virtual reality than in reality? an answer from a simulated job interview. *Interacting with Computers*, 24(4):265–272.

[315] Vogl, W., Ma, B. K.-L., and Sitti, M. (2006). Augmented reality user interface for an atomic force microscope-based nanorobotic system. *IEEE transactions on nanotechnology*, 5(4):397–406.

[316] Walker, M., Hedayati, H., Lee, J., and Szafir, D. (2018). Communicating robot motion intent with augmented reality. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 316–324. ACM.

[317] Walters, M. L., Dautenhahn, K., Te Boekhorst, R., Koay, K. L., Kaouri, C., Woods, S., Nehaniv, C., Lee, D., and Werry, I. (2005). The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pages 347–352. IEEE.

[318] Walters, M. L., Oskoei, M. A., Syrdal, D. S., and Dautenhahn, K. (2011). A long-term human-robot proxemic study. In *2011 RO-MAN*, pages 137–142. IEEE.

[319] Wang, X., Love, P. E., Kim, M. J., and Wang, W. (2014). Mutual awareness in collaborative design: An augmented reality integrated telepresence system. *Computers in Industry*, 65(2):314–324.

[320] Whitney, J. P., Chen, T., Mars, J., and Hodgins, J. K. (2016). A hybrid hydrostatic transmission and human-safe haptic telepresence robot. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 690–695. IEEE.

[321] Wikipedia (2020a). Human–robot interaction. https://en.wikipedia.org/wiki/Human-robot_interaction.

[322] Wikipedia (2020b). Panorama. https://en.wikipedia.org/wiki/Panorama.

[323] Wilcox, R., Nikolaidis, S., and Shah, J. (2013). Optimization of temporal dynamics for adaptive human-robot interaction in assembly manufacturing. *Robotics*, 8:441.

[324] Willemsen, P. and Gooch, A. A. (2002). Perceived egocentric distances in real, image-based, and traditional virtual environments. In *Proceedings IEEE Virtual Reality 2002*, pages 275–276. IEEE.

[325] Willemsen, P., Gooch, A. A., Thompson, W. B., and Creem-Regehr, S. H. (2008). Effects of stereo viewing conditions on distance perception in virtual environments. *Presence: Teleoperators and Virtual Environments*, 17(1):91–101.

[326] Xu, B., Ellis, J., and Erickson, T. (2017). Attention from afar: Simulating the gazes of remote participants in hybrid meetings. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, pages 101–113. ACM.

[327] Youdas, J. W., Garrett, T. R., Suman, V. J., Bogard, C. L., Hallman, H. O., and Carey, J. R. (1992). Normal range of motion of the cervical spine: an initial goniometric study. *Physical therapy*, 72(11):770–780.

[328] Young, J. E., Kamiyama, Y., Reichenbach, J., Igarashi, T., and Sharlin, E. (2011). How to walk a robot: A dog-leash human-robot interface. In *2011 RO-MAN*, pages 376–382. IEEE.

[329] Young, J. E., Xin, M., and Sharlin, E. (2007). Robot expressionism through cartooning. In *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 309–316. IEEE.

[330] Zacharias, G. and Young, L. (1981). Influence of combined visual and vestibular cues on human perception and control of horizontal rotation. *Experimental brain research*, 41(2):159–171.

[331] Zatorre, R. J. and Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral cortex*, 11(10):946–953.

[332] Zhang, G., Hansen, J. P., Minakata, K., Alapetite, A., and Wang, Z. (2019). Eye-gaze-controlled telepresence robots for people with motor disabilities. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 574–575. IEEE.

[333] Zhang, J., Langbehn, E., Krupke, D., Katzakis, N., and Steinicke, F. (2018a). A 360 video-based robot platform for telepresent redirected walking. In *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for Human-Robot Interactions (VAM-HRI)*, pages 58–62. ACM/IEEE.

[334] Zhang, J., Langbehn, E., Krupke, D., Katzakis, N., and Steinicke, F. (2018b). Detection thresholds for rotation and translation gains in 360 video-based telepresence systems. *IEEE transactions on visualization and computer graphics*, 24(4):1671–1680.

[335] Zillner, J., Rhemann, C., Izadi, S., and Haller, M. (2014). 3d-board: a whole-body remote collaborative whiteboard. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 471–479.

[336] Zollmann, S., Hoppe, C., Langlotz, T., and Reitmayr, G. (2014). Flyar: Augmented reality supported micro aerial vehicle navigation. *IEEE transactions on visualization and computer graphics*, 20(4):560–568.

Appendix A

Questionnaires

Experience Questionnaire

- Have you used a virtual reality headset (Oculus Rift, HTC Vive etc.) before?

  □ Yes
  □ No

- How much experience do you have with VR headset?

  No Experience □ 0 □ 1 □ 2 □ 3 □ 4 Much Experience

- Do you have experience with 3D computer games?

  No Experience □ 0 □ 1 □ 2 □ 3 □ 4 Much Experience

- How many hours do you play per week?

  ____

- Do you have experience with 3D stereoscopic display (cinema, games etc.)?

  No Experience □ 0 □ 1 □ 2 □ 3 □ 4 Much Experience

Igroup Presence Questionnaire (IPQ)

- In the computer generated world I had a sense of "being there".

  Not at all □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 Very much

- Somehow I felt that the virtual world surrounded me.

  Fully disagree □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 Fully agree

- I felt like I was just perceiving pictures.

  Fully disagree □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 Fully agree

- I did not feel present in the virtual space.

  Did not feel □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 Felt present

- I had a sense of acting in the virtual space, rather than operating something from outside.

  Fully disagree □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 Fully agree

- I felt present in the virtual space.

  Fully disagree □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 Fully agree

- How aware were you of the real world surrounding while navigating in the virtual world (i.e. sounds, room temperature, other people etc.)?

  □ extremely aware
  □ moderately aware
  □ not aware at all

- I was not aware of my real environment.

  Fully disagree □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 Fully agree

- I still paid attention to the real environment.

  Fully disagree □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 Fully agree

- I was completely captivated by the virtual world.

  Fully disagree □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 Fully agree

- How real did the virtual world seem to you?

  Completely real □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 Not real at all

- How much did your experience in the virtual environment seem consistent with your real world experience?

  □ not consistent
  □ moderately consistent
  □ very consistent

- How real did the virtual world seem to you?

  About as real as an imagined world □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 Indistinguishable from the real world

- The virtual world seemed more realistic than the real world.

  Fully disagree □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 Fully agree

Simulator Sickness Questionnaire (SSQ)

- General discomfort

  None □ 0 □ 1 □ 2 □ 3 Severe

- Fatigue

  None □ 0 □ 1 □ 2 □ 3 Severe

- Headache

  None □ 0 □ 1 □ 2 □ 3 Severe

- Eyestrain

  None □ 0 □ 1 □ 2 □ 3 Severe

- Difficulty focusing

  None □ 0 □ 1 □ 2 □ 3 Severe

- Increased salivation

  None □ 0 □ 1 □ 2 □ 3 Severe

- Sweating

  None □ 0 □ 1 □ 2 □ 3 Severe

- Nausea

  None □ 0 □ 1 □ 2 □ 3 Severe

- Difficulty concentrating

  None □ 0 □ 1 □ 2 □ 3 Severe

- Fullness of head

  None □ 0 □ 1 □ 2 □ 3 Severe

- Blurred vision

  None □ 0 □ 1 □ 2 □ 3 Severe

- Dizzy (eyes open)

  None □ 0 □ 1 □ 2 □ 3 Severe

- Dizzy (eyes closed)

  None □ 0 □ 1 □ 2 □ 3 Severe

- Vertigo

  None □ 0 □ 1 □ 2 □ 3 Severe

- Stomach awareness

  None □ 0 □ 1 □ 2 □ 3 Severe

- Burping

None □ 0 □ 1 □ 2 □ 3 Severe

Demographic Questionnaire

- Height

  ____

- Age

  ____

- Gender

  □ male
  □ female
  □ other

- Profession / field of study

  ____

- Vision correction

  □ none
  □ glasses
  □ contact lenses

- Do you suffer from a displacement of equilibrium or similar?

  □ Yes
  □ No

- Do you have a known eye disorder?

  □ color blindness
  □ night blindness
  □ dyschromatopsia (red-green color weakness)
  □ strong eye dominance
  □ other ___________________________________

- Additional comments

  ____

Concentration Questionnaire

- How sure are you that you always chose the correct answer?

  Very low □ 0 □ 1 □ 2 □ 3 □ 4 Very High

- Did you have a cognitive strategy to answer the questions?

  ____

- How would you subjectively describe your level of attention during the experiment?

  Very low □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 Very High