UNIVERSIDADE D
COIMBRA

Luis Agnelo de Almeida

TOWARDS TELEPRESENCE AND ROBOT TELEOPERATION

Tese no âmbito do Doutoramento em Engenharia Eletrotécnica e de Computadores, Ramo de Especialização em Computadores e Electrónica, orientada pelo Professor Doutor Paulo Jorge Carvalho Menezes e pelo Professor Doutor Jorge Manuel Miranda Dias, e apresentada ao Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Dezembro de 2022

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D
COIMBRA

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Towards Telepresence and Robot Teleoperation

Luis Agnelo de Almeida

Thesis within the scope of Doctoral Program in Electrical and Computer Engineering, Specialization in Computers and Electronics, supervised by Professor Paulo Jorge Carvalho Menezes and by Professor Jorge Manuel Miranda Dias, and presented to the Department of Electrotechnical and Computer Engineering of the Faculty of Science and Technology at the University of Coimbra

December 2022

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D
COIMBRA

DEPARTAMENTO DE ENGENHARIA ELETROTÉCNICA E DE COMPUTADORES

Telepresença e Teleoperação de Robôs

Luis Agnelo de Almeida

Tese no âmbito do Doutoramento em Engenharia Eletrotécnica e de Computadores, Ramo de Especialização em Computadores e Electrónica, orientada pelo Professor Doutor Paulo Jorge Carvalho Menezes e pelo Professor Doutor Jorge Manuel Miranda Dias, e apresentada ao Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Dezembro 2022

The work presented in this thesis was carried out within the Immersive Systems and Sensory Stimulation (IS3L) and Artificial Perception for Intelligent Systems and Robotics (AP4ISR), groups of the Institute of Systems and Robotics of the University of Coimbra (ISR-UC), and within the Polytechnic Institute of Tomar, Portugal.

Acknowledgements

First, I would like to express my deep gratitude to my supervisors, Professor Paulo Jorge Carvalho Menezes and Professor Jorge Manuel Miranda Dias, for their guidance, encouragement, and patience with me. I would also like to thank the support of the Institute of Systems and Robotics (ISR) at the University of Coimbra (UC), and the Polytechnic Institute of Tomar (IPT) for granting me the conditions to develop the work in this thesis. Thanks to the many co-authors, colleagues, and friends at ISR and IPT with whom I had the pleasure of working and discussing ideas. A special thanks to Bruno Patrão for the laboratory work and Luis Oliveira for his encouragement and continuous support. Over these years, conversations over coffee or work with colleagues in the laboratory or at work have always been enriching, namely among many from ISR: Bruno Ferreira, David Portugal, Diego Faria, Francisco Vasconcelos, Gustavo Assunção, Hadi Aliakbarpour, João Filipe Ferreira, João Quintas, Jorge Lobo, José Prado, Juan Carlos Garcia, Kamrad Roudposhti, Luís C. Vitória dos Santos, Luís Santos, Maria Rita Nogueira, Micael Couceiro, Pedro Martins, Pedro Trindade, Ricardo Martins, Rui P. Rocha, and from IPT: Ana Lopes, Ana Vieira, António Manso, Carlos Queiroz, Casimiro Batista, Cecilia Batista, Gabriel Pires, João Patricio, José Casimiro Pereira, José Ramos, Luis Merca, Manuel Barros, Nuno Madeira, Paulo Santos, Pedro Correia, Pedro Neves, Renato Panda.

Most of all, I would like to thank my father, mother, brother, and sister-in-law for their never-ending support and patience, and to my nephews João and Maria Inês for the joy they convey.

I acknowledge that a PROTEC research fellowship has partially supported this research under Grant SFRH/PROTEC/67912/2010.

Abstract

Interpersonal communication and physical interaction with the immediate environment depend upon our awareness of space and motion. The physical distance to a person or object affects our behaviour, influencing how we establish eye contact to express ourselves or how we move. This suggests that virtual communication technologies should provide the sensation of the proximity of other people or our presence in a remote environment. Space perception and its use (proxemics) are supposed to replicate real situations. How a person perceives himself, the world and interacts with it, defines his presence experience. Our space and motion perception depends not only on visual perception, but also on the vestibular system, proprioception and cognitive processes. Several sensory input systems, such as vision, auditory, proprioception, kinesthetic, vestibular, olfactory and thermal, contribute with information to continuously update our representation of the spatial structure of the environment. All this information enables us to efficiently navigate in an environment and establish the relationship between our body and the world. As immersion aims at providing stimuli that illude the sensory system, it is possible to create a consistency between outside sensory feedback and inside sensory information (proprioceptive, vestibular), and the brain's cognitive models. This research explores means to induce the sense of telepresence in human-centered communications and in remote robot teleoperations. Since robots help humans experience and perform actions in distant places, results from studies on human factors are used to provide recommendations for the construction of immersive teleoperation systems. Moreover, a testbed has been developed to study perceptual issues affecting task performance while users manipulate the environment through traditional or immersive interfaces. In teleoperation research, we focus on designing and evaluating new immersive interacting mechanisms for teleoperating a remote robot. We explore the notion of telepresence and physical embodiment to create the tele-embodiment concept. Thus, contributing to virtually transferring the operator to the remote robot, enhancing robot operation, minimising the cognitive workload and improving task performances. Additionally, in human-centered mediated communications, real face-to-face meeting benefits, like eye-to-eye contact establishment, gesture reconnaissance, body language or facial expressions, are not supported by commodity conferencing technologies such as Zoom, Teams or Skype. To transmit these social cues and enhance copresence, this research proposes a low-cost framework to support three-dimensional conferencing through augmented reality (AR) based on telepresence. The contribution is an incremental online 3D model reconstruction solution useful for real-time interaction in mixed reality workspaces, augmented reality environments or human-computer interactions. The approach explores virtual view synthesis through body motion estimation and hybrid sensors composed of a video and depth camera. A wearable glove-based method was also developed for natural task execution in virtual interaction scenarios. Telepresence robots are becoming popular in social interactions involving health care, elderly assistance, guidance, or office meetings. In robot-mediated interactions, there are two types of human psychological experiences to consider: (1) telepresence, in which a user develops a sense of being present in a remote physical location, near the remote interlocutor, and (2) co-presence, in which a

user perceives the other person as being present locally with him or her. Moreover, this work presents a literature review on developments supporting robotic social interactions, contributing to improving the sense of presence and co-presence via robot mediation.

Keywords

presence; telepresence; immersion; immersive interfaces; teleoperation; telerobotic; social robotics; co-presence; copresence; social presence; cognitive robotics; human-centered computing; cognitive human robot interaction; HCI design and evaluation methods; computer vision; robotics; augmented reality; virtual reality; mixed reality; 3D reconstruction; computer graphics

Resumo

A comunicação interpessoal e a interação física com o ambiente imediato dependem de nossa consciência do espaço e do movimento. A distância física a uma pessoa ou objeto afeta o nosso comportamento, influenciando a forma como estabelecemos contacto visual para nos expressarmos ou como nos movimentamos. Isso sugere que as tecnologias de comunicação virtual devem proporcionar a sensação de proximidade de outras pessoas ou da nossa presença em um ambiente remoto. A percepção do espaço e seu uso (proxêmica) devem replicar situações reais. A percepção que a pessoa tem de si própra, e como interage com o mundo define sua experiência de presença. A nossa percepção de espaço e movimento depende não apenas da percepção visual, mas também do sistema vestibular, propriocepção e processos cognitivos. Existem vários sistemas de sensoriais, como visão, audição, propriocepção, kinestésico, vestibular, olfativo e térmico que contribuem com informações para atualizar continuamente nossa representação da estrutura espacial do ambiente.Todas essas informações permitem-nos navegar com eficiência em um ambiente e estabelecer a relação entre nosso corpo e o mundo. Dado que a imersão visa fornecer estímulos que iludem o sistema sensorial, é possível criar uma consistência entre o feedback sensorial externo e a informação sensorial interna (proprioceptiva, vestibular) e os modelos cognitivos do cérebro.Esta investigação explora meios para induzir a sensação de telepresença em comunicações centradas no ser humano e em teleoperações de robôs remotos. Uma vez que os robôs são úteis para permitir que humanos experimentem e executem ações em lugares distantes, resultados de estudos sobre fatores humanos são usados para fornecer um conjunto de recomendações para a construção de sistemas imersivos de teleoperação. Além disso, foi desenvolvido um setup de testes para estudar questões perceptivas que afetam o desempenho da tarefa enquanto os utilizadores manipulam o ambiente, seja por meio de interfaces tradicionais ou imersivas. Na pesquisa de teleoperação, focamos no design e avaliação de novos mecanismos de interação imersiva para teleoperar um robô remoto. Exploramos a noção de telepresença e incorporação física para criar o conceito de tele-incorporação. Assim, contribuindo para transferir virtualmente o operador para o robô remoto, melhorando a operação do robô, minimizando a carga de trabalho cognitiva e melhorando o desempenho das tarefas. Além disso, em comunicações mediadas centradas no ser humano, os benefícios reais de reuniões face a face, como estabelecimento de contato olho no olho, reconhecimento de gestos, linguagem corporal ou expressões faciais, não são suportados por tecnologias de conferência de commodities, como Zoom, Teams ou Skype . Para transmitir essas pistas sociais e melhorar a copresença, esta pesquisa propõe uma estrutura de baixo custo para suportar conferências tridimensionais por meio de realidade aumentada (AR) baseada em telepresença. A contribuição é uma solução de reconstrução de modelo 3D online incremental útil para interação em tempo real em espaços de trabalho de realidade mista, ambientes de realidade aumentada ou interações humano-computador. A abordagem explora a síntese de visão virtual através da estimativa de movimento corporal e sensores híbridos compostos por câmeras de vídeo e uma câmera de profundidade.Um método baseado em luvas vestíveis também foi desenvolvido

para execução natural de tarefas em cenários de interação virtual.Os robôs de telepresença estão se tornando populares em interações sociais envolvendo assistência médica, assistência a idosos, orientação ou reuniões de escritório. Nas interações mediadas por robôs, existem dois tipos de experiências psicológicas humanas a serem consideradas: (1) telepresença, na qual um usuário desenvolve a sensação de estar presente em um local físico remoto, próximo ao interlocutor remoto, e (2) copresença , em que um utilizador percebe a outra pessoa como estando presente localmente com ele ou ela. Além disso, este trabalho apresenta uma revisão de literatura sobre desenvolvimentos que suportam interações sociais robóticas, contribuindo para melhorar o sentimento de presença e copresença via mediação robótica.

Palavras-Chave

presença; telepresença; imersão; interfaces imersivos; teleoperação; telerobótica; robótica social; co-presença; copresença; presença social; robótica cognitiva; computação centrada no ser humano; interação cognitiva humano-robô; projecto e avaliação de interfaces humano-computador; visão computacional; robótica; realidade aumentada; realidade virtual; realidade mista; reconstrução 3D; computação gráfica

Contents

Acronyms

AR Augmented Reality.

DOF Degree of Freedom.

ECG Electrocardiography.

EEG Electroencephalography.

fMRI Functional Magnetic Resonance Imaging.

FoV Field of View.

GSR Galvanic Skin Response.

HMD Head Mounted Display.

HMI Human–Machine Interface.

HRI Human–Robot Interaction.

PTU Pan-and-Tilt unit.

ST Skin Temperature.

VE Virtual Enviroment.

VR Virtual Reality.

List of Figures

List of Tables

Chapter 1

Introduction

Human interactions depend to some extent on their self-conscientious and their space and motion perception. This research aims to explore these intersecting fields to induce the sense of telepresence in human-centered communications and in remote robot teleoperations. To induce sensations of *being there*, in the presence of other people or in a remote environment through a robot, the mediation technologies must mimic and simulate the sensory inputs encountered in nature. The result is a perceptual and cognitive experience for the user, which responses are supposed to be similar as if he/she were really present in the remote physical environment.

We contribute to providing a consistency between outside sensory feedback (vision, audio, haptic), inside sensory information (proprioceptive, vestibular) and cognitive models through technological means, eliciting telepresence.

1.1 Context and Motivation

Humans like to communicate with each other and socialize. During most of their ages, this is a straightforward task because of their global mobility and ability to meet people. Once they become old, with less strength and mobility, they become confined to their homes. This situation is causing home elderly loneliness and health problems that increase government health care costs [360][144]. In several countries, over half of 65+ persons are living alone. Phone and video conversations have enabled elderly individuals to stay in touch with family, friends and caregivers, overcoming social isolation. However, it does not replace social contact during daily shopping, public services, travelling, bar, coffee, garden walks, etc. Thus, such friendly environments should be recreated while simulating a face-to-face meeting. Therefore, means of communication that enable eye contact establishment, gestures reconnaissance, body language and facial expressions are required.

Telepresence robots are becoming popular in the context of social interactions. Typically, these systems enable people to look at a distant place via teleoperating a robot and interacting with another person at a remote location using built-in com-

munication devices. Research, such as the EU ExCITE project, has assessed the validity and robustness of mobile robotic telepresence (MRP), in supporting elderly people and encouraging the development of their social interactions (42-months' long evaluation) [384]. Immersive research on gaze-controlled telepresence robots has accessed its use for persons unable to use their hands because of a motor disability [532]. In short, applications include health care, elderly assistance, autism therapy, guidance or office meetings [11, 41, 221, 228, 368, 451, 496].

Additionally, augmented reality (AR) and particularly tele-immersion can provide the technology means that enable users to interact remotely and experience the benefits of a face-to-face meeting [46][274]. The tele-immersive technology combines virtual reality for rendering and display purposes, computer vision for image capturing and 3D reconstruction, and various networking techniques for transmitting data between remote sites in real-time with minimal delay [266][399]. Virtual meeting spaces allow the possibility of socialization, collaborative work on 3D data, 3DTV [95] [357], remote training and monitoring, and remote teaching of physical activities (e.g., rehabilitation, dance)[49]. Stimulus control and consistency that support repetitive actions and real-time performance feedback are some strengths of virtual reality environments already used on rehabilitation (e.g. phobia treatments, motor exercises, elderly fall preventions) [273][417][265].

1.2 Research Questions and Objectives

Neurophysiological and phenomenological studies [82][433] evidence the relation between space and motion perception and self-conscientious. Blanke [73] describes how integrating multisensory signals and brain processing contributes to self-conscientiousness. In his studies, several subjects received ambiguous multisensory information about the location and appearance of their own body and revealed that specific brain areas reflect the conscious experience of identifying with the body (self-identification (also known as body-ownership)), the experience of where am 'I' in space (self-location) and the experience of the position from where 'I' perceive the world (first-person perspective). Mel Slater found that when embodied in a virtual body, the perspective position and point of view provided to the user minimizes the importance of visual-tactile synchronisation [463]. Biocca [71] refers that intermodal integration (intersensory integration) may be a key psychological mechanism contributing to a sense of presence in virtual environments. Sensorimotor processes associated with multimodal integration may integrate perceptual cues and motor actions into a coherent experience and relatively consistent model of objects and spaces.

Current state-of-art does not address low-cost solutions for telepresence. Phone and internet chat/audio/video conferencing programs (ex: VOIP, NetMeeting, Skype, WhatsUp) have been used for socialization nevertheless, they cannot create a remote person's presence feeling. The dynamic nature of 3D meetings and remote communications put the challenge of optimal representation for graphics and vision while opening various research opportunities, including real-time performance imaging, multi-view video, virtual view synthesis, etc.

One of the research lines of this thesis proposes to study and develop a low-cost framework that supports three-dimensional conferencing through augmented reality (AR) based on telepresence to enhance copresence. The aim is to develop an incremental online 3D human reconstruction solution useful for real-time interaction on mixed reality workspaces validated through *telepresence* measures.

Given the importance of immersive interfaces in these solutions and to allow users to explore remote spaces, research evolved into the teleoperation of telepresence robots.

This research branch focuses on the study, design and evaluation of new immersive interfaces for the teleoperation a remote robot. We explore the notion of telepresence and physical embodiment to create the tele-embodiment concept. The objective is to induce in the operator the feeling of being present in the remote environment while maximizing task performance and minimizing the operator's physical and cognitive workload.

1.3 Contributions

This research explores means to induce the sense of telepresence in human-centered communications and remote robot teleoperations. Given that immersion aims at providing stimuli that illude the sensory system, the proposed solutions maintain the consistency between outside sensory feedback and inside sensory information (proprioceptive, vestibular), and the brain's cognitive models. The space and motion perception and the consequent interactions with the mediated world (virtual or real) should be as natural as the user was there. This work shows that people can experience and perform actions in remote places through a robotic agent having the illusion of being physically there. The sensation can be compelled through immersive interfaces; however, technological contingencies can affect human perception. Based on the human factors results of related works, we provide a set of recommendations for the design of immersive teleoperation systems aiming to improve the sense of telepresence for typical tasks (ex. Table 3.5). The mitigation of issues such as system latency, field of view, frame of reference, or frame rate contributes to enhancing the sense of telepresence. The presented evaluation methodology enables analyzing how perceptual issues affect task performance. By decoupling the flows of an immersive teleoperation system, we start to understand how vision and interaction fidelity affect spatial cognition. Task experiments with participants using traditional vs. immersive interfaces allowed quantifying the disturbance introduced by each component of the system.

The human role in the teleoperation control loop is fundamental because it is the operator who can decide, react, and adjust operations in the presence of noisy and incomplete data (especially in unstructured and unpredictable scenarios). This fact made human factor analysis an essential tool for designing new teleoperation interfaces aiming simultaneously for better performances and decreasing the number of failures caused by operator faults. One recommendation is that the interface systems should be developed so that the operator (surgeon, pilot, or other) receives the necessary information to perform the task without the need

to search for it in unusual places. Vital information should always be placed in a visible and salient way so that the user can perceive it immediately.

Regarding immersive teleoperation, research has demonstrated that it is possible to generate the remote physical embodiment feeling by letting the user perceive the robot's structure as his/her own body. To evolve from tele-operation to embodied operation, this research proposes a view transfer using an Head Mounted Display (HMD) (i.e. an egocentric controlled view in which the user will see what robot can see), and the use of natural commands (implicit commands instead explicit ones). A key development in this research is the cockpit concept in which the user feels inside the robot, perceiving and acting naturally.

By exploring computer graphics, spatial audio, computer vision and reconstruction techniques were demonstrated the potential of inducing sensations of being physical in the *presence* of other people. Namely, regarding human-centered mediated communications, this research proposes a low-cost framework to support three-dimensional conferencing through Augmented Reality (AR) based on telepresence. It aims to achieve the real face-to-face meeting benefits, in which important social cues such as eye-to-eye contact establishment, gesture reconnaissance, body language or facial expressions are transmitted, (presently not supported by commodity conferencing technologies such as Zoom, Teams or Skype). The contribution is a free viewpoint system framework that synthesizes views of an online 3D reconstructed model dependent on the observer's point of view. The approach explores virtual view synthesis through motion body estimation and hybrid sensors composed of video cameras and a low-cost depth camera based on structured light. The solution addresses the geometry reconstruction challenge from traditional video cameras array, that is, the lack of accuracy in low-texture or repeated pattern regions. We present a full 3D body reconstruction system that combines visual features and shape-based alignment. The modelling is based on meshes computed from dense depth maps to minimize processed data resulting in a global 3D mesh representation independent of the viewpoint. Research contributions include an incremental version of the Crust algorithm that efficiently adds new vertices to an already existing surface without having to recompute previously generated meshes and a topological incremental reconstruction approach based on confidence measures that avoid redundant data information computation. With this online reconstructed 3D model, it is possible to provide a synchronous point of view for an observer that moves in front of a display of a face-to-face meeting application, thus enhancing the presence sensation.

Additionally, a wearable glove-based method was also developed for natural task execution in virtual interaction scenarios.

Moreover, this work presents a literature review on developments supporting robotic social interactions, contributing to improving the sense of presence and co-presence via robot mediation. It aims to gather knowledge to help roboticists design improved user- and environment-adaptive systems and technical methods. Reviews have addressed user-adaptive systems and environment-adaptive systems for social robotics (in which the robot is generally an autonomous agent serving the bystander user). However, we further explore telepresence social robotics, emphasizing the relationship between the robot's operator and the by-

stander user.

1.4 Thesis Structure

Chapter 1 presents the research context, motivation, objectives and contributions towards telepresence and robot teleoperation.

Chapter 2 addresses concepts regarding immersion, presence, telepresence and copresence and related works on telepresence approaches.

Chapter 3 addresses interface transparency issues in teleoperation and presents a testbed framework for evaluating immersive interfaces, including a study case.

Chapter 4 presents a set of experimental works demonstrating that telepresence, embodiment, natural interaction, and immersive interfaces enhance robot teleoperation, minimizing cognitive workload and improving task performance.

Chapter 5 proposes a fast 3D model acquisition framework contributing to copresence in human-centered mediated communications, HRI and HMI.

Chapter 6 presents a literature review on developments supporting robotic social interactions, contributing to improving the sense of presence and copresence via robot mediation.

Chapter 7 presents the main contributions, conclusion, and future work towards telepresence and robot teleoperation.

1.5 Publications

The contributions of this thesis resulted in the following publications in international peer-reviewed conferences and journals:

Journal papers:

1. Luis Almeida, Paulo Menezes, Jorge Dias, "Telepresence Social Robotics towards Co-Presence: A Review", in Applied Sciences. 2022; 12(11):5557. https://doi.org/10.3390/app12115557

2. Luis Almeida, Paulo Menezes, Jorge Dias, "Interface Transparency Issues in Teleoperation", in Applied Sciences. 2020; 10(18):6232. https://doi.org/10.3390/app10186232

3. J. Garcia Sanchez, B. Patrão, L. Almeida, J. Perez, P. Menezes, J. Dias, P. Sanz, "Design and Evaluation of a Natural Interface for Remote Operation of Underwater Robots", in IEEE Computer Graphics and Applications, vol. 37, no. 1, pp. 34-43, Jan.-Feb. 2017. http://dx.doi.org/10.1109/MCG.2015.118

4. H. Aliakbarpour, L. Almeida, P. Menezes, J. Dias - "Multi-sensor 3D Volumetric Reconstruction Using CUDA". Journal of 3D Research, Volume 2, Number 4, ISSN: 2092-6731, Springer, December 2011. http://dx.doi.org/10.1007/3DRes.04(2011)6

Peer-reviewed books chapters

1. L. Almeida, P. Menezes, J. Dias, "Augmented reality framework for the socialization between elderly people", peer reviewed chapter in "Handbook of Research on ICTs for Healthcare and Social Services: Developments and Applications", Isabel Maria Miranda & Maria Manuela Cruz-Cunha (Eds.), IGI Global, 2013. http://dx.doi.org/10.4018/978-1-4666-3986-7.ch023

Conference papers:

1. Luis Almeida, Elio Lopes, Beril Yalçinkaya, Rodolfo Martins, Ana Lopes, Paulo Menezes, and Gabriel Pires, "Towards natural interaction in immersive reality with a cyber-glove", 2019 IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2019), 06-09 October Bari, Italy. http://dx.doi.org/10.1109/SMC.2019.8914239

2. Luis Almeida, Paulo Menezes, Jorge Dias, "Improving robot teleoperation experience via immersive interfaces", in 2017 4th Experiment International Conference (exp.at'17), Faro, Portugal, June 6-8, 2017. http://dx.doi.org/10.1109/EXPAT.2017.7984414

3. L. Almeida, P. Menezes, J. Dias, "Incremental Reconstruction Approach for Telepresence or AR Applications", in SciTecIN'15 - Sciences and Technologies of Interaction, EPCGI'2015: The 22nd Portuguese Conf. on Computer Graphics and Interaction, Coimbra, Portugal, November 12 – 13, 2015. https://doi.org/10.2312/pt.20151202

4. J. Quintas, L. Almeida, E. Sousa, P. Menezes, "A context-aware immersive interface for teleoperation of mobile robots", in SciTecIN'15 - Sciences and Technologies of Interaction, EPCGI'2015: The 22nd Portuguese Conf. on Computer Graphics and Interaction, Coimbra, Portugal, November 12 – 13, 2015. https://doi.org/10.2312/pt.20151213

5. L. Almeida, B. Patrão, P. Menezes, J. Dias, "Be the Robot: Human Embodiment in Tele-Operation Driving Tasks", Ro-Man 2014: The 23rd IEEE International Symposium on Robot and Human Interactive Coapunication, Edinburgh, UK, August 25-29, 2014. http://dx.doi.org/10.1109/ROMAN.2014.6926298

6. J. Quintas, L. Almeida, M. Brito, G. Quintela, P. Menezes, J. Dias, "Context-based understanding of interaction intentions", Ro-Man 2012: The 21st IEEE International Symposium on Robot and Human Interactive Coapunication, Paris, France, 9-13 September 2012. https://doi.org/10.1109/ROMAN.2012.6343803

7. Luis Almeida, Paulo Menezes and Jorge Dias, "On-Line 3D Body Modelling for Augmented Reality", GRAPP 2012 - International Conference on Computer Graphics Theory and Applications, Rome, Italy, February 24-26, 2012. http://dx.doi.org/10.5220/0003866304720479

8. L. Almeida, P. Menezes, L. D. Seneviratne, J. Dias, "Incremental 3D Body Reconstruction Framework for Robotic Telepresence Applications", in Robo2011: The 2nd IASTED International Conference on Robotics, Pittsburgh, USA, November 7 – 9, 2011. http://dx.doi.org/10.2316/P.2011.752-068

9. L. Almeida, F. Vasconcelos, J.P. Barreto, P. Menezes, and J. Dias, "On-line incremental 3D human body reconstruction for HMI or AR applications", CLAWAR2011 - The 14th International Conference on Climbing and Walking Robots And the Support Technologies for Mobile Machines, Paris, France, September 2011. https://doi.org/10.1142/8305

10. Luis Almeida, Paulo Menezes and Jorge Dias, "Stereo Vision Head Vergence Using GPU Cepstral Filtering", VISAPP 2011 - Proceedings of the Fifth International Conference on Computer Vision Theory and Applications, Vilamoura, Algarve, Portugal, March 5-7, 2011. http://dx.doi.org/10.5220/0003319406650670

11. Luis Almeida, Paulo Menezes and Jorge Dias, "Vergence Using GPU Cepstral Filtering", in Proceedings of the DoCEIS'11 - Doctoral Conference on Computing, Electrical and Industrial Systems. Lisbon, Portugal, February, 2011. Springer - ISBN 978-3-642-19169-5. http://dx.doi.org/10.1007/978-3-642-19170-1_35

Posters:

1. Luis Almeida, Paulo Menezes, Jorge Dias, "3D Modelling framework: an incremental approach". In Eurographics 2016 - Posters, the 37th Annual Conference of the European Association for Computer Graphics, Lisbon, Portugal, 9-13 May, 2016. http://dx.doi.org/10.2312/egp.20161047

Chapter 2

Background and Related Work

Virtual Reality (VR) creates a sensory and psychological experience for users as an alternative to reality. VR is an ever-growing set of tools, technologies and techniques that can be used to create the psychological sensation of being in an alternate space. Underpinning the methods used to create compelling virtual environments is the basic observation that information is fated for processing by a human sensory and perceptual system that has evolved to interact with regularities occurring in the physical world [177]. The more one can provide the system with sensory inputs that simulate and effectively mimic those encountered in nature, the more convincing the resulting perceptual and cognitive experience will be for the user. The ultimate goal of designers and users of VR environments is a computer-generated simulation that is indistinguishable to the user from its real-world equivalent [81]. The hardware and software used to create a VR system are designed to replicate the information available to the sensory/perceptual system in the physical world. System components create illusions for each of the senses, particularly for vision, hearing, and touch [91].

AR technology supplements (combines) the real world with virtual objects (computer-generated) that appear to coexist in the same space as the real world [46]. The virtual scene or ambience generated by the computer is designed to enhance the user's sensory perception of the world they see or interact with. Azuma defines an AR system as having the following properties: combines real and virtual objects in real environments; runs interactively in real-time; and register (aligns) real and virtual object with each other. AR is not restricted to the sense of sight or particular display technology; it can be applied to all senses, including hearing, touch and smell. Additionally, Milgram and Fumio Kishino defined Milgram's Reality-Virtuality Continuum in 1994 [343]. They describe a continuum spanning from an entirely real environment to a purely virtual one. In between are Augmented Reality (closer to the real environment) and Augmented Virtuality (closer to the virtual environment).

Displays and Input Devices

Augmented reality/Virtual reality low-cost technologies typically are based on the following components [91][112]:
Hardware:

The computational device: either a desktop, a PDA, a smartphone or a laptop pc equipped with an advanced image graphics card (GPU); Different peripheral devices: visual, aural or haptic devices (e.g. video cameras, microphones, force-feedback manipulator device); A non-immersive or immersive image display system: a screen, a high frame rate screen with stereoscopic shutter glasses, a polarised screen with polarised filter glasses or a head-mounted display (HMD);
A motion sensor (or tracking device), usually integrated with the HMD or a printed marker video camera tracking-based, provides information to the computer where the user is looking based on their head movement.
VR uses many input devices, some specifically designed for it, such as hand controllers and cyber gloves. Simpler and commonly found devices are baseless trackballs, joysticks, and even one-handed keyboards such as the Twiddler.
Spatial trackers commonly attached to the head and hand enable tracking the user's body. Having three Degree of Freedom (DOF) affords position in three-dimensional space, or orientation (yaw, pitch, roll.), although 6 DOF are ideal, affording both position and orientation of the tracked object. For example, these devices track head movements and ensure correct information for rendering images provided to the user (e.g. for HMDs). This information is also useful for providing the location/direction of the user for other purposes, such as for interaction (e.g. producing appropriate ambient sounds when entering different spaces) or for behavioural analysis.

Software:
According to the hardware and software used in AR/VR applications, various virtual settings can be distinguished:
Desktop/Monitor AR, based on subjective immersion: In these systems, users achieve a feeling of immersion through stereoscopic vision tools. They can interact with the virtual world using head or body movements, a mouse, a joystick, or other VR peripherals such as data gloves.
Fully immersive VR: Users can be fully immersed in a computer-generated environment. Systems create this illusion by providing immersive stimuli through output devices such as HMDs, glasses, force feedback robotic arms, and a head/body tracking system that coordinates the user's movements with the environment's feedback. The user is unable to see the real world around them and is surrounded by the virtual environment. CAVE: A CAVE is a small room that projects a computer-generated world on its walls, including both the front and sidewalls. This solution is suitable for collective VR experiences because it allows different people to share the same experience at the same time.
Telepresence systems: Users can influence and operate in the real world, even from a different location. They can observe the current situation through remote cameras and perform actions using a remote robotic agent, which may include arms.

2.1 Immersion, Presence, TelePresence, Social presence, Co-presence

Immersion, also known as *sensorimotor immersion*, refers to the extent and fidelity of physical stimulation affecting the human sensory systems and the system's responsiveness to the motor inputs [82]. The immersive level depends on the number and range of sensors and the motor channels connected to a remote agent in a real environment (e.g., a robot) or to a mediated virtual environment. Immersion is determined by the naturalness and coherence between actions (head, body, and gesture movements) and the expected sensory feedback [75, 456, 462, 463].

Presence is the psychological product of technological immersion, defined as *the perceptual illusion of non-mediation* [298] or simply referred to as the sense of *being there* in a mediated virtual environment [67, 82].

Sheridan [450] differentiates presence (virtual) from telepresence (experiential), in which *presence* describes the experience of being present within a virtual world, while *telepresence* refers to the sense of being in a mediated remote real environment [297, 429].

Social presence has been defined as the *sense of being together with another*, which includes primitive reactions to social cues and automatic creation of simulations or mental models of *"other minds"* [69]. Short et al. [453] defined *social presence* as "the degree of salience of the other person in the interaction and the consequent salience of the interpersonal relationship".

Co-presence is a different concept, introduced by Goffman [180] to describe the active state in which a person perceives their interlocutor, and the interlocutor also perceives him or her. *Copresence* refers to a "psychological connection to and with another person", in which "interactants feel they were able to perceive their interaction partner and that their interaction partner actively perceived them" [371]. With co-presence being a subjective concept, it involves different dimensions and interpretations depending on the social science discipline and application area (e.g., sociology or psychology) [378, 401, 538]. *Co-presence* has been used to refer to the *sense of being together* with others in a mediated environment, either remote real or virtual [68, 538]. As described in the definitions, concepts such as co-presence and social presence should not be confused as they are assessed differently [371].

2.1.1 The Utility of Presence and its Effect

To use the presence concept in a practical situation, understanding the results or consequences of presence is essential. The following subsection overviews theories and empirical studies on the usefulness of presence. Schuemie [440] identified in the literature the following consequences: subjective sensation, task performance, response and emotions and simulator sickness.

Subjective sensation: Most of the theories on presence refer to the experience of the subjective sensation of *being there* while immersed in a Virtual Enviroment (VE), and this sensation is, in fact, part of most definitions of presence. This subjective sensation can apply to the environment currently being experienced or to memories of past experiences. As Slater notes [460], a key result of presence is that a person remembers the VE as a place rather than a set of pictures.

Task performance: Research has demonstrated a positive association between the presence and human performance [56, 187, 306]. Mania and Chalmers [314] confirm that presence need not be related to task performance in an empirical study with three conditions: lectures were given on a specific topic in the real world, in a virtual classroom, and an auditory-only environment. In a between-subject design, 18 subjects were assigned to each condition. A preliminary analysis of the data was done by comparing means and standard deviations and applying the ANOVA method. Presence was found not to be correlated with the task performance of acquiring knowledge during the lecture. Kim [257] found a weak but significant correlation between presence and performance in terms of factual memory and average recognition speed for recognizing static pictures from a TV infomercial. The most interesting finding is evidence that there may be two dimensions to telepresence; one, they have labelled *arrival* and the other *departure*. The sense of arrival appears to be close to the sense of *being there* in the virtual environment. But the sense of *being there* that they call arrival may not be equivalent to or as powerful as the sense of departure, the sense of *not being here* in the physical environment.

Responses and emotions: Considered one of the most important consequences of presence is that a virtual experience can evoke the same reactions and emotions as a real experience. Hodges[212] in a between-subject experiment with ten subjects on a waitlist and ten subjects being treated for fear of heights in VR, showed that the subjects, who were all acrophobic, did show increased subjectively reported anxiety when confronted with height in the VE. These studies and other similar studies have demonstrated that VR treatment reduces acrophobia (Gandy, et al., 2010) [170]. North [370] found that people can show signs of fear of public speaking when confronted with a virtual audience. Other authors [166] have tested persons when confronted with visual cues suggesting motion and concluded that a person would tend to correct for the perceived motion by adjusting their body posture. Such effect is explored by portable console flight (or car) simulator games like Sony PSP or iPhone.

Simulator sickness: One problem associated with using VR is that it can cause nausea and dizziness, the phenomenon known as “simulator sickness”. Slater found a positive correlation between simulator sickness and presence.

2.1.2 Measuring Presence

Measures for presence are often based on the expected results of presence. A distinction can be made between:

- subjective measures, requiring introspection by the subjects

- objective measures, such as behavioural measures and physiological measures

Subjective measures: The most commonly used measures in presence research are based on subjective ratings through questionnaires. Witmer and Singer[518] developed a presence questionnaire (PQ) to measure presence in VEs. In addition, they developed an immersive tendencies questionnaire (ITQ) to measure differences in the tendencies of individuals to experience presence. Questions try to evaluate factors like:
Control factors, the amount of control the user had on events in the VE.
Sensory factors, the quality, number and consistency of displays
Distraction factors, are the degree of distraction by objects and events in the real world.
Realism factors, the degree of realism of the portrayed VE.

Users can rate their experience in the VE according to these factors on questions with a 7-point scale. The presence score is the sum of these ratings. Other authors define questions to evaluate user virtual environment (VE) experiences like:
Spatial Presence (SP), the relation between the VE as a space and the own body
Involvement (INV), the awareness devoted to the VE
Realness (REAL), the sense of reality attributed to the VE.
On ITC Sense Of Presence Inventory (ITC-SOP) Lessister [280] attempted to create standard questionnaires to analyzing factors like:
Physical Space, for example, “I had a sense of being in the scenes displayed”, “I felt I was visiting the places in the displayed environment”, “I felt that the characters and/or objects could almost touch me”.
Engagement, for example, “I felt involved (in the displayed environment)”, “I enjoyed myself”, “My experience was intense.”
Naturalness, for example, “The content seemed believable to me”, “I had a strong sense that the characters and objects were solid”, “The displayed environment seemed natural”.
Negative effects include “I felt dizzy”, “I felt disorientated”, “I felt nauseous.”
Lombard [298] analyzed user factors presence involved on a 3D IMAX movie:
Immersion relates to the sense of immersion, involvement and engagement in the mediated environment;
Parasocial Interaction relates to interacting with other people in real-time in the mediated environment;
Parasocial Relationships concern feelings of friendship, etc., toward people in the VE;
Physiological Response concerns, amongst others, simulator sickness;
Social Reality relates to how likely the events are to occur in reality;
Interpersonal Social Richness relates to how well the user can observe interpersonal communication cues; and
General Social Richness relates to items such as unemotional vs emotional, unresponsive vs responsive, and impersonal vs personal.

Other subjective measures include continuous measures. Instead of administering a questionnaire only after a virtual experience, some approaches suggest a

continuous measure of presence during the experience using, for example, a hand-operated slider that could be used to indicate the level of presence experienced at that moment

Objective measures: people's behaviour tends to be influenced by mediated stimuli as if they were unmediated when they experience a high level of presence. Examining *people's reaction* to mediated stimuli can lead to an objective measure of presence. Sheridan proposes measuring reflex responses, such as automatically catching a ball or avoiding a rapidly approaching object. Freeman [166] attempted to use *postural response* as a measure for presence but found no significant correlation between this measure and reported presence. In Barakova studies [55] human players were asked to act out several scenarios. From the sensor data analysis, it became apparent that the walking and hand movements are most informative about the emotional state of human subjects. Head gaze attention can also be analyzed either using an eye tracker or a head pose tracker. Bailenson demonstrated that conference interactants in the rendered head movement condition rated a higher level of co-presence[78]. Active multisensory perception using spatial maps has been the object of study by Ferreira & Dias 2010 [150] enabling the visuoauditory-driven gaze shift analyses through Bayesian framework algorithms. Gesture and body pose tracking can provide objective analysis measures [334].

Physiological signals are used as presence measures, although Sheridan [450] warns that " *Presence* is a subjective sensation, much like *mental workload* and *mental model*- it is a mental manifestation, not so amenable to objective physiological definition and measurement". Research experiments [213][458][458] attempt to measure presence using

- Functional Magnetic Resonance Imaging (fMRI)
- Electrocardiography (ECG) (Heart rate)
- Eye movements (Eye Scanpath)
- Skin Temperature (ST)
- Galvanic Skin Response (GSR)
- Electroencephalography (EEG)

2.1.3 Applications

Revolutionary advances in the underlying VR/AR enabling technologies (i.e., computation speed and power, graphics and image rendering technology, display systems, interface devices, immersive audio, haptics tools, tracking, intelligent agents, and authoring software) have supported development resulting in more powerful, low-cost PC-driven VR systems. Such technological advances and accessibility have provided the hardware platforms needed for the conduct of human research and treatment within more usable, useful and lower-cost VR systems. Alongside evolving technology, VR/AR applications have blossomed in a wide range of areas. VR/AR has proven useful for gaining basic scientific

knowledge, in medical diagnosis and treatment, commerce and entertainment (especially in the realm of desktop VR), training, and cultural heritage. For illustrative purposes, Bohil [81], Lange[273] and Rizzo [417] present a recent sampling of VR/AR applications.

Training: Virtual environments are often ideal for meeting training needs. They provide cost-effective standardised interactive experiences as they are potentially reusable by a wide audience. They are safe learning experiences (i.e., mistakes only lead to virtual consequences, not costly or dangerous outcomes in the real world that make on-the-job training hazardous). They are compelling (users often report higher levels of engagement in completing a virtual task relative to more traditional methods such as listening to a lecture or reading a book). It is well-known that high levels of motivation and engagement lead to improved learning outcomes. Training can be done using either fully immersive or desktop virtual environments.

Communication Skills: Researchers at Case Western Reserve University have created a training simulator to enhance communication skills in dental students. This desktop VR application makes use of the massively multiplayer online world of Second Life. Like DIANA, this training simulation focuses on fostering improved doctor/patient communication. Students get much-needed practice in collecting patient history information, informing patients about treatment options, and describing dental techniques.

Medicine

Several medicine research areas have been exploring VR/AR applications (Harders, 2008)[195]. VR/AR allows researchers to see patient behaviours and body structures in new ways and enables new and effective therapeutic approaches. VR/AR offers the potential to create systematic human testing, training, and treatment environments with precise control. It enables immersive, dynamic 3D stimulus presentations, enabling sophisticated interaction, behavioural tracking, and performance recording. Rizzo describes several case studies and summarizes through a SWOT analysis [417] for VR rehabilitation and therapy the *Strengths*, *Opportunities*, *Weaknesses* and *Threats*: Strengths: Enhanced Ecological Validity; Stimulus Control and Consistency; Real-Time Performance Feedback Process; Cuing Stimuli to Support “Error-Free Learning”; Self-Guided Exploration and Independent Practice; Interface Modification Contingent on User’s Impairments; Complete Naturalistic Performance Record; Safe Testing and Training Environment; Gaming Factors to Enhance Motivation; Low-Cost Environments That Can be Duplicated and Distributed. Weakness: The Interface Challenge 1: Interaction Methods; The Interface Challenge 2: Wires and Displays; Immature Engineering Process; Platform Compatibility; Front-End Flexibility; Back-End Data Extraction, Management, Analysis and Visualization; Side Effects.

Online Virtual Worlds Therapies:

Online 3-D virtual worlds are computer-based simulated environments mainly modelled by their users that can create and manipulate elements and thus experiences telepresence to a certain degree [66][77][53]. Second Life, There, IMVU, Active World, Roblox, Fortnite or Meta’s social VR platform Horizon (Metaverse) are some

of the 3-D virtual worlds where millions of users interact with each other daily through their avatars, that is three-dimensional graphical representations of themselves. 3-D virtual worlds represent a good opportunity to create innovative online health services based on the following features: an extended sense of presence (3-D virtual worlds transform health guidelines and provisions into experience); an extended sense of community (social presence): online worlds use hybrid social interaction and dynamics of group sessions to provide each user with targeted - but also anonymous, if required - social support in both the physical and virtual world. Gorini and Riva [186] refer to a Second Life Psychotherapy case study.

Rehabilitation:

Many medical researchers have explored the use of VR in rehabilitating stroke victims. At the University of Haifa, researchers have found a way to assess different patterns of stroke-induced brain damage. Patients' hand motions are recorded as they respond to virtual flying objects (tennis balls). The researchers' computer models use this motion's data to diagnose patients with high accuracy (approximately equivalent to that of human physicians). They expect that these models will allow diagnosis and rehabilitation decisions that outperform any doctor. At Rutgers University (Boian, et al., 2002)[83] researchers have used a desktop VR system equipped with data gloves for stroke rehabilitation. The patient exercises his or her affected hand and arm by manipulating an on-screen hand to interact with a virtual butterfly, play a virtual piano, and perform other tasks. Due to the increased engagement that this task creates for the participant, the system leads to marked improvements.

Lange, Rizzo and their colleagues [273] present a recent overview of rehabilitation research in their article *The Potential of Virtual Reality and Gaming to Assist Successful Aging with Disability*. They state that virtual reality (VR), and gaming applications have the potential to address clinical challenges for a range of disabilities. VR-based games can potentially provide the ability to assess and augment cognitive and motor rehabilitation under a range of stimulus conditions that are not easily controllable and quantifiable in the real world. They discuss an approach for maximizing function and participation for elderly people with and into a disability by combining task-specific training with advances in VR and gaming technologies to enable positive behavioural modifications for independence in the home and community. There is potential for the use of VR and game applications for rehabilitating, maintaining, and enhancing those processes that are affected by ageing with and into disability, particularly the need to attain a balance in the interplay between sensorimotor function and cognitive demands and to reap the benefits of task-specific training and regular physical activity and exercise. They address the following processes: Virtual reality and gaming technology for sensorimotor and cognitive rehabilitation; VR rehabilitation for balance impairments; VR rehabilitation for a home exercise program for the shoulder; VR rehabilitation for dexterous manipulation with the fingertips; VR rehabilitation and stimulated active seating for pressure ulcer prevention;

Our research pursuing a technological test bed to measure the sense of presence quantitatively is inspired by a notable work published by researchers from Georgia Center Georgia Institute of Technology, Dept. of Psychology North Carolina

State University. The Experiences with an AR Evaluation Test Bed: Presence, Performance, and Physiological Measurement (Gandy, et al., 2010)[170]. They discuss an experiment carried out in an AR test bed called *the pit*. It is a VR acrophobia study. The experimental goals are to explore whether VR presence instruments are useful in AR (and modify them where required), compare additional measures to these well-researched techniques, and determine if findings from VR evaluations can be transferred to AR. Their test bed, *the pit*, presents to the user a virtual hole in the floor that appears to drop three stories, and they analyzed the effect of the illusion at different frame rates, measuring the induced anxiety using several methodologies (AR presence questionnaires (subjective measure) and three lead electrocardiogram (ECG) sensor placed on their chest as well as galvanic skin response (GSR) and skin temperature sensors mounted on their non-dominant hand (physiological measures)). Four factors were identified via factor analysis: Interaction & Immersion, Interference & Distraction, and Audio & Tactile Experience; Moving in the Environment. They concluded that high presence feelings were reflected on the questionnaire correctly. They found that physiological measures were challenging, and heart rate data was too noisy, but GSR seemed more promising as it could be analyzed over small segments of time in the experiment and responded to within seconds.

Concerning rehabilitation, the task is strongly multidisciplinary, as it integrates different areas of video games, physical rehabilitation and computer vision. Its importance has already attracted the video game industry. With EyeToy, the gamers use the PlayStation 2's EyeToy camera to interact with objects on their TV screen in a "virtual" workout (the game puts the player into a game onscreen, representing an augmented virtuality example). Nintendo Wii video games have been used for physical therapy for patients after injuries and strokes in rehabilitation hospitals like the Sister Kenny Rehabilitation Institute in Minneapolis, USA [225]. Several computerized systems for virtual rehabilitation are commercially available. In the cognitive domain, virtual environment-based therapies are provided by companies such as Virtually Better (Atlanta GA), Lumosity Labs (San Francisco, CA) and the Nintendo DS Brain Age series. Game consoles are also currently being used in motor rehabilitation, with the Wii being the most popular game console adopted clinically. Cognitive games for the Wii train language (vocabulary) skills (My Word Coach) or memory and logic (Big Brain Academy) [90]. The Wii game console is not appropriate for individuals challenged by arm gravity loading or with severe shoulder, elbow or finger spasticity. It has been demonstrated that competition in games makes them less boring and more motivating than traditional therapy. Microsoft recently launched Xbox Kinect controller-free games system as a response to Nintendo's extremely successful Wii, enabling user game interaction through gestures and body motion. Kinect Xbox 360 Fantastic Pets game (another augmented virtuality example) enables players to step inside the world and onto the screen where they can play and care for their pets. Using augmented reality games to motivate people to do physical therapy has huge potential.

2.2 Tele-presence Approaches

Measuring the moving three-dimensional contours of the inhabitants of a room and its other contents can be accomplished in a variety of ways [274](Lanier, 2001). As early as 1993, Henry Fuchs of the University of North Carolina at Chapel Hill had proposed one method, known as the *sea of cameras* approach [168], in which the viewpoints of many cameras are compared. In typical scenes in a human environment, there will tend to be visual features, such as a fold in a sweater, that is visible to more than one camera. By comparing the angle at which these features are seen by different cameras, algorithms can piece together a three-dimensional model of the scene. This technique was explored in non-real-time configurations, notably in Takeo Kanade's work [242], which later culminated in the *Virtualized Reality* demonstration at Carnegie Mellon University. That setup consisted of 51 inward-looking cameras mounted on a geodesic dome. Because it was not a real-time device, it could not be used for tele-immersion. Instead, videotape recorders captured events in the dome for later processing. Ruzena Bajcsy, head of the GRASP (General Robotics, Automation, Sensing and Perception) Laboratory at the University of Pennsylvania, was intrigued by the idea of real-time seas of cameras. Starting in 1994, she worked with colleagues at Chapel Hill and Carnegie Mellon on small-scale *puddles* of two or three cameras to gather real-world data for virtual-reality applications. Bajcsy and her colleague Kostas Daniilidis took on the assignment of creating the first real-time sea of cameras - one that was scalable and modular so that it could be adapted to a variety of rooms and uses. They worked closely with the Chapel Hill team, which was responsible for taking the *animated sculpture* data and using computer graphics techniques to turn it into a realistic scene for each user [121]. But a sea of cameras in itself isn't a complete solution. Suppose a sea of cameras is looking at a clean white wall. Because there are no surface features, the cameras have no information with which to build a sculptural model. A person can look at a white wall without being confused. Humans don't worry that a wall might be a passage to an infinitely deep white chasm because we don't rely on geometric cues alone–we also have a model of a room in our minds that can rein in errant mental interpretations. Unfortunately, to today's digital cameras, a person's forehead or T-shirt can present the same challenge as a white wall, and today's software isn't smart enough to undo the confusion that results. To overcome this problem, we are proposing hybrid solutions composed of depth and video cameras.

In recent years, there has been a significant effort focusing on immersive video conferencing and immersive television, challenging research areas and consumers product industries [266][399][501][357]. 3D cinema, 3D console games, 3D contends, 3D broadcast or 3DTV LCDs are common technologies nowadays. As a key component, the display technology can now convey a stereoscopic perception of 3-D depth to the viewer either using light active shutter glasses, passive polarized glasses or even without glasses, using flat-panel autostereoscopic solutions employing lenticular lenses or parallax barriers. F. Isgro, Emanuele Trucco, Peter Kauff and Oliver Schreer present a good survey paper titled "Three-Dimensional Image Processing in the Future of Immersive Media" [229], where they discuss the three-dimensional image processing challenges posed by present and future

immersive telecommunications, especially immersive video conferencing and television. European-funded projects like VIRTUE, 3DTV, 3D4YOU, 2020-3D-MEDIA, MOBILE 3DTV, 3D PHONE and 3D Presence or TEEVE demonstrate an interest in the area.

2.2.1 Realistically represent the user's appearance

Avatars are a common method to represent the user inside the virtual environment, but it is not realistic. A full body 3D reconstruction can realistically represent the user's appearance and full dynamics of movement, such as facial expressions, chest deformation during breathing, and movement of hair or clothing [266]. Real-time human body 3D reconstruction approaches can be divided into three categories:

1. silhouette-based reconstruction,
2. voxel-based methods with space sampling
3. image-based reconstruction with dense stereo depth-maps.

In silhouette-based reconstruction, the 3D information is obtained via visual hulls that are formed by intersecting generalized cones between a silhouette and the camera center [201][442]. In the voxel-based method, depth is determined by sampling a uniform grid of space using colour consistency. The vision-based reconstruction [266][295] creates dense stereo depth maps by correlating slightly displaced views of the same scene. Szeliski [443] presents a comparative survey for these methods. Recent approaches are using multi-view image and time-of-flight (ToF) sensor fusion for dense 3D reconstruction [95][258]. Pollefey [9] proposed techniques to scan outdoor scenes from video imageries by a batch optimization process. Scanning of static and small objects is reported by Van Gool [109] with the help of structured, encoded lighting to enhance the accuracy of 3D data acquisitions. Reddy [415] introduces compressed sensing (CS) for multi-view tracking and 3D-voxel reconstruction. They apply the CS theory on sparse background-subtracted silhouettes to address various multi-view estimation problems. They use random projections (compressed measurements) of the silhouette images for directly recovering object parameters in the scene coordinates. To keep the computational requirements of this recovery procedure reasonable, they tessellate the scene into a bunch of non-overlapping lines and perform estimation on each of these lines. Kurillo[266] introduces a stereo mapping using adaptive triangulation which allows a faster reconstruction. The algorithm produces partial 3D meshes, instead of a dense point. It reduces the dense stereo depth map calculation from working pixel-by-pixel to working region-by-region. Space carving [442] theory can be used to create three-dimensional models of objects from a set of images to be used as input to virtual reality systems.

2.2.2 Virtual view synthesis and modelling

Virtual view synthesis and modeling are the potential graphic tools to create the eye-to-eye contact illusion in telepresence communications. The approach involves surface reconstruction while a basic task for object detection, manipulation and environment modeling. Generally, the object's surface is reconstructed by merging measurements from different views. This approach requires depth data and sensor pose data. When both, pose and depth, are unknown, structure from motion is a solution. Corresponding features in consecutive images are used to estimate the ego-motion of the sensor. Based on this ego-motion information the depth without absolute scale is estimated.

Since recent depth cameras also provide RGB data, 2D image processing algorithms are usable. Point feature mapping in RGB images can be improved by the associated depth data obtaining a 3D feature tracking. Most common methods for matching 2D image features are based on the KLT (Kanade-Lucas-Tomasi) [452][304][494], SIFT (Scale-Invariant Feature Transform) [302] or SURF (Speeded Up Robust Features) [57] approaches. If only the depth information but no pose is given, i.e. by using a stereo camera or a laser scanner system without inertial sensors, the Iterative Closest Point (ICP) algorithm can be used to register point clouds acquired from different perspectives [64][295]. Finally, if pose and depth are known, the registration procedure is dispensable and the data can simply be merged. In any case, the quality of surface reconstruction depends on the precision of sensor pose estimation and depth measurement. Calculating changes in the 3D pose based on these methods have been performed by several works, e.g. [204][347][9][328][334].

2.3 Conclusions

This chapter introduced background information regarding VR and AR technologies and key concepts for this thesis regarding immersion, presence, telepresence, and copresence. It addresses the utility of presence, its effects, how to measure it, and application areas. Additionally, it overviews related works on telepresence approaches focusing on realistically representing the user's appearance and virtual view synthesis and modeling methods in virtual or mixed environments.

Chapter 3

Interface Transparency Issues in Teleoperation

Transferring skills and expertise to remote places, without being present, is a new challenge for our digitally interconnected society. People can experience and perform actions in distant places through a robotic agent wearing immersive interfaces to feel physically there. However, technological contingencies can affect human perception, compromising skill-based performances. Considering the results from studies on human factors, a set of recommendations for the construction of immersive teleoperation systems is provided, followed by an example of the evaluation methodology. We developed a testbed to study perceptual issues that affect task performance while users manipulated the environment either through traditional or immersive interfaces. The analysis of its effect on perception, navigation, and manipulation relies on performances measures and subjective answers. The goal is to mitigate the effect of factors such as system latency, field of view, frame of reference, or frame rate to achieve the sense of telepresence. By decoupling the flows of an immersive teleoperation system, we aim to understand how vision and interaction fidelity affects spatial cognition. Results show that misalignments between the frame of reference for vision and motor-action or the use of tools affecting the sense of body position or movement have a high effect on mental workload and spatial cognition.

3.1 Introduction

A telepresence robot presents a solution for doctors and health care workers to consult, handle, or monitor people in remote places or in contaminated areas, avoiding self-exposure. For instance, in the present coronavirus pandemic (COVID-19) or in recent epidemics like Ebola virus disease (EVD), a physician could teleoperate a robot and, through it, look around, move, or communicate safely in a contained environment. The robot's capabilities of perception, manipulation, and mobility allow performing some disaster-response tasks [126, 354, 356]. Telerobotics is already present in areas like surgery (e.g., the Da Vinci Robot) [4], remote inspection, space exploration [52, 311], underwater maintenance, nuclear

disposal, hazardous environment cleaning, and search and rescue. However, most of these robotic interventions in critical tasks still rely on the human's control capabilities. Teleoperated robots quite often include semi-autonomous functionalities to assist operators. To this end, cognitive human-robot interaction architectures are being used to minimize the control workload, improve the task performance, and increase safety [496][99]. Thus, the design of such cognitive robotic systems can integrate the knowledge of human perceptual factors to predict the intended actions and needs of operators.

Human's actions depend on the perception of the environment, while the decisions rely on correct the recognition and interpretation of sensory stimuli [142]. For this, several sensory modalities contribute information, providing cues for the perception of space and motion. These perceptual cues are continuously acquired and matched with our mental models. While some cues are consistently matched, others do not fit. Consequently, our brain tries to solve these conflicts to avoid compromising consequent actions, e.g., we expect that a known object seems bigger when it is near us and smaller if it is further away. We also do not expect to "see" a radio's loudspeaker somewhere and "listen" to the respective sound coming from a different direction. Another example of unsolved conflict can occur while on a train looking for a parallel train, where we cannot distinguish which train is moving, whether it it is ours or the other one. Surprisingly, we deal well with other conflicting situations: when we are combing our hair in front of a mirror, the hand that really "touches" us is not the one that we are "seeing": it is the reflection of the hand; this results from a learning process because children do not know how to comb in front of a mirror. In short, real-world representations are built on sensory information and cognition to understand them, using thoughts and experience (bottom-up and top-down processes complement each other).

Currently, a person can act in a remote environment through a teleoperation system; however, the use of any type of mediation requires training. This interaction with the system can be simple, if the control interface remains intuitive and natural. Ideally, if the operator feels as though he/she is in a remote location, he/she will act as naturally as though he/she were physically there. The illusion of telepresence can be induced through a proper action-perception loop supported by technology [449]. A person experiencing sensory stimuli similar to those in the remote environment and acting in line with them can actually sense "being there" [297]. Such a feeling depends on the capability of the system to accurately display remote environment properties and provide enough information about the remote agent and the responsiveness of the system to motor inputs, i.e., immerse the person in media [82]. Sheridan [450] suggested a distinction between (virtual) presence and (experiential) telepresence: presence refers to the experience of being present in a virtual world and telepresence to the sense of being in a mediated remote real environment. The term telepresence was initially introduced by Minsky in the teleoperation context. It refers to the phenomenon that a human operator develops a sense of being physically present at a remote location through interaction with the system's human interface [345]. Telerobotics appears as a subclass of teleoperation where the human operator supervises and/or controls a remote semi-automatic robotic system. The teleoperation of robots involves two major activities: remote perception and remote manipulation. In [27], to improve

teleoperation, we proposed the use of immersive technologies to allow operators to perceive the remote environment as though being there, where the control of the robot was as simple as controlling their own bodies. Thus, operators could sense the robot's body as their own (embodiment [459]), simplifying the navigation control. The use of head-mounted displays (HMDs) to naturally control the point of view of the remote robot's camera and display control instruments was further explored in [174] to allow operators to feel that they were controlling the robot from inside. The sense of presence was improved using a "virtual cockpit" while the operators' faults and mental workload were minimized. Nevertheless, we realised that some factors compromised the sense of telepresence and degraded task performance in these works and other authors' related works [322]. It is not simple to dissociate the flows in a teleoperation system (visual feedback, haptic, control, etc.) because impairing one of them would make the robot uncontrollable. Therefore, the present research aims to contribute to human perceptual factors with an impact on the teleoperator's behaviour. Some factors degrade human performances, such as low video stream bandwidth, frame rates, time lags, frame of reference, lack of proprioception, two-dimensional views, attention switches, or motion effects [99, 407, 459]. Some related works [234] analyzed the global effect of these factors on physical workload and on task performance without discriminating their weights in the system. Other authors focused on the influence of multimodality concurrency on factors [504], the effect of a specific factor such as the field of view [261], or concentrate on virtual environments [315]. The present study aims to decouple the flows of an immersive teleoperation system to understand how the vision and interaction fidelity levels affect spatial cognition.

The immersive experience can be enhanced through the replication of real visual feedback, thus supporting the operator's traditional hand-eye coordination [438, 484]. Based on Slater's findings [459], we also aim at consistency between actions (movements) and multimodal feedback (e.g., visual and haptic). Pursuing this goal, we expect to provide enough spatial and motion cues of the remote environment to allow operators to perceive and behave naturally. Therefore, we contribute to the minimization of cognitive workload and performance improvement.

We propose simplifying the operation control by exploring and combining telerobotics with telepresence. We argue that if with the use of media devices, the operators experiment with the sensation of being in the remote environment, then the task performance becomes as natural as being there. This research proposes an immersive interface approach for teleoperation and evaluates it against traditional remote control systems. The introduction of telepresence systems influences and enlarges the range of applications that can benefit from its use. Nevertheless, given some current technological limitations and taking into account the results from studies on human factors, a set of recommendations for the design of these systems is presented. The evaluation of any interactive system is a crucial step in the development process. This evaluation can be divided into two parts: user task performance and usability and user acceptance. Although the ultimate goal is to maximize the performance metrics, given the very central role of the user, such performance is very dependent on not only how the user is able to execute the task, but how the system influences the perception of that task and if that may influence its execution positively. Simplification, effort reduction, and intuitiveness are some

keywords that positively influence the user and therefore his/her performance. To this end, some guidelines on which aspects are important to evaluate in a teleoperation system are also presented.

The work discussed in this chapter was published in the Applied Sciences journal paper [24].

Structure of the chapter: Section 3.2 provides insights about skills and performance in human activity. Section 3.3 describes the human role in teleoperation, namely the control and feedback flows with the human in the loop operating a remote device. Section 3.4 analyzes the technological constraints and human perceptual factors with relevance for immersive teleoperation design, providing a set of recommendations for the construction of these systems. Section 3.5 presents an evaluation methodology for immersive interfaces in teleoperation. As an example, an immersive interface is proposed, and the influence of perception on task execution is analyzed. Section 3.6 presents the quantitative and qualitative results and a discussion of the proposed immersive interface. Section 3.7 presents the findings and conclusions.

3.2 Skills and Performance

Human activities have been distinguished by Rasmussen into three types: knowledge-, rule-, and skill-based [413]. In fact, complex activities may involve knowledge, rules, and skills with some level of alternation or simultaneous execution among them. Most of the skills, in particular those that involve motor coordination, are acquired via training, and during this process, the human brain establishes the relationships between low level sensory signals and muscle actions,and higher level goals, rules, and knowledge. The skill acquisition may be as difficult as it is to distinguish the relevant sensory information from signal noise and how complex the relationship between that information and the proper actions to be executed is [142, 483].

Picking the example of writing some text, if a person perfectly masters the type-writing technique, the focus is on the search for which words to express the idea. However, on the other side, for someone who is not used to keyboards, the focus will be on searching for the keys to compose the words. In the latter case, as the foreground activity is the search for the keys, there will be an increased difficulty in producing the text directly from ideas. In this case, probably, it would be better to hand write it first to avoid distraction and the consequent ruining of the establishment of the intended line of thought. Either way, the performance will be lower, either in terms of ideas about the text or the time to produce it.

Skills are related to actions that we may delegate to more reactive levels of our brain, where actions are produced as answers to sensory signals without any conscious intervention [413]. There may exist one intention that establishes some behavior of reference that serves as a guide to the production of actions in response to sensory signals. Therefore, it is common to propose closed-loop-based models to study and explain these automatic behaviors [72]. There

are however alarm mechanisms that are activated when excessive errors or mismatches are detected in these signals. When these happen, higher complexity layers are called to intervene to select a change in behavior, induce a correction, or a simple parameter adjustment that will bring the actuation back to “normality”. In the typewriting example, this may happen when a finger strikes two keys at once and the touch sense reports that to the brain. This normally triggers a reaction of confirming the typed characters or words.

We may say that during skill acquisition, we start by the understanding of how to act on the controls or objects. In the second phase, the focus is on what we intend the object or system to do under our action while keeping some level of surveillance on the grasping of the object or command that enables us to detect slippage, abnormal resistance, or other aspects that may indicate the existence of some failure in the action or in the means of interaction. As the task execution becomes automatic and does not require any particular attention to it, the cognitive level is free to focus on more abstract levels [413], e.g., the text to be written. As another example, we may consider a taxi driver that is concentrated on choosing the best way to reach a given destination and apparently not paying any special attention to the other cars or the road limits. Nevertheless, as soon as the front car stop lights are lit, an obstacle appears on the road or the motor does not respond as usual to the gas pedal actuation, and that detail is brought immediately to the primal attention levels.

As skills are acquired for particular ways of performing activities, introducing changes in the way the activity is performed or sensed may result in performance reduction, at least until a re-adaptation is possible and completed. Again with the keyboard example, if it becomes less precise and typed keys produce zero or multiple characters, this requires the user to confirm the output constantly and check if the text is correct as expected. Similarly, a faulty screen where some parts of it show black bars instead of the content forces the user to make sure that all pressed keys are done in the correct order as some of the outputs cannot be checked visually a posteriori. Similar effects may be introduced by automated mechanisms that are intended to help the user, e.g., snap to grid in drawing programs or spelling correctors in text editors, which change fine drawing movements into unintended locations or foreign words into inappropriate and out of context ones.

As demonstrated by the above, the existence of imprecise or disturbing factors may reduce the performance of any task execution in particular, but is not limited to those that require motor-related skills.

3.3 The Human Role in Teleoperation

A teleoperated system can be seen as a control loop, as in Figure 3.1, where the operator plays the role of selecting the appropriate signals and acts on the controlled element to execute the desired task.

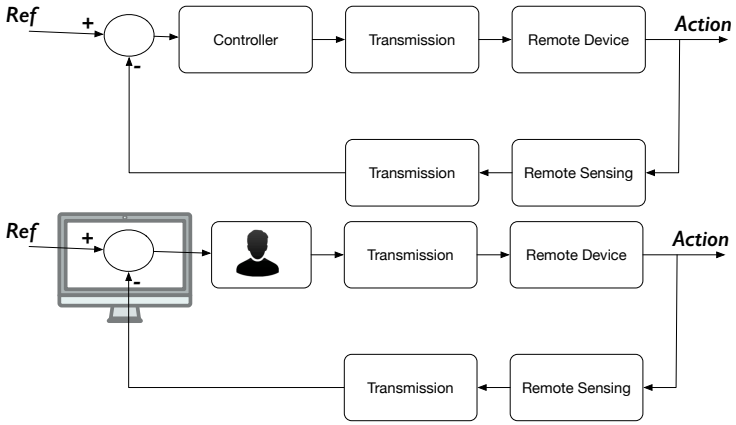


Figure 3.1: A remote control loop schema (top). A modified control loop where an operator visualizes the intended reference and remote signals to generate the control signals to operate the remote device (bottom).

The task can be one of following a trajectory, a pick-and-place operation, or a much more complicated one. As in any control loop, the objective is to have the system follow some reference signal, and that is accomplished by comparing that input signal with some measurable observation of the system output. As the operator plays the role of the controller, he/she needs to receive the information about the system variables that he/she intends to control [74, 350]. Frequently, the information received is very limited and corrupted by different types of noise. It is here that humans normally play still irreplaceable roles given their ability to make correct inferences from incomplete data. Humans can still extract the necessary information in the presence of the most variable conditions, where typically automated mechanisms can only be tuned to specific working conditions and fail if these conditions are not met. Humans have the ability to use temporal signals, to build mental image mosaics, and to use them to make and execute plans [142]. This comes with a price: it is very hard to follow a memorized plan and simultaneously keep track of multiple information sources. In fact, the required concentration level induces important fatigue, and as it rises, it reduces our attention and concentration capabilities. Humans have some difficulties in noticing changes in slowly varying stimuli even if their intensity is high [181]. Actually, the exposure to the same stimuli for some period of time induces a process of desensitization where a constant or repetitive stimuli loses importance and becomes progressively integrated as part of the “background”. On the other side, we are particularly good at learning automated skills; thus, operations like driving a car become simple reactive tasks where the conscious level is mostly left for the high-level plan, whereas lane keeping, velocity control, distance to other vehicles, etc., are performed mostly at the subconscious (reactive) level. It is well known that the reactive execution is simpler and less “energy consuming” than tasks that require reasoning. Similarly, the reactions to well distinguishable visual, auditory, or tactile stimuli generate much less fatigue than being focused on the detection of minor changes of the same stimuli. We are equipped with a set of complementary sensing mechanisms that together enable us to pay attention to most of the important events in our lives with minimum effort. For example, although the human eyes only produce high resolution information at

their fovea, to minimize the amount of normal processing workload, their peripheral vision can detect sudden changes and motion. These motion field-generated stimuli, as well as auditory stimuli are enough to direct the visual attention to the right spot in the surrounding space when needed. This means that although we frequently browse the surrounding areas around us, we rely more on particular event detection to attract and direct our attention, instead of using any type of constant and exhaustive scanning process. This has indeed two main advantages: it saves energy and allows performing focused tasks while still being able to detect nearby events, in particular those that may represent any type of danger and thus require immediate action.

Why are the above questions important when we are talking about teleoperation-based tasks? When we talk about executing tasks, we always define, implicitly or explicitly, success goals or success criteria. Once the goal is defined, the execution time, the amount of work done, or other task performance measures are used for evaluation and compare operator skills or the usability metrics of different systems. These analyses are very important as they are one of the few ways of evaluating interactive systems quantitatively and may complement other types of more subjective results obtained from user judgments or opinions. Nevertheless, these evaluations can only be done after the whole teleoperation system is built and set up. Furthermore, their analysis may reveal that the system is not adequate for the pre-established goals, but does not necessarily provide any guidance in discovering which are the elements that are responsible for the failure. From this, it seems clear that some set of recommendations and guiding rules can be very handy to anticipate the possible problem sources and avoid them during the early design phases. Some of these recommendations can be obtained from known studies from areas like psychophysics, neuroscience, physiology, or ergonomics and establish limits for various operation's parameters. Subsequently, an analysis of engineering models of the system to be designed may provide predictions for relevant parameters that may then be verified if they are inside the acceptable ranges defined by the former recommendations. Other recommendations can come out of heuristic experience and still provide valuable guidance to anticipate and avoid any possible problems or performance issues. The following section presents a set of recommendations for the design of teleoperation systems.

3.3.1 Human Factors, Tasks, and Telepresence

Human beings have a remarkable ability to adapt to unexpected constraints and still carry on with their actions [181]. Skilled people can perform tasks, including teleoperation, even when there are momentary failures in certain feedback channels, such as visual feedback. This means that the person has memorized all the operations and can predict the output of his/her actions under certain ranges of disturbances in the visual feedback. In the absence of the typical feedback modality, other senses are used to get the necessary feedback; memories are triggered, and adjustments can be made (e.g., haptic feedback, touch). Even in cases where the sequence of actions is to be memorized and the person is expected to be able to execute the sequence of actions without any kind of external stimuli, he or she uses various sensorial cues (e.g., proprioception and time

notion) to adjust the performance and correct any motor action deviations. With teleoperation, a user aims to transfer his/her abilities to a remote agent to perform tasks like navigation, perception, and manipulation. If during this process, he/she may have the impression of being at the remote site, these actions will benefit from first person perception and cognitive mechanisms, and therefore improve the achieved performance. This sense of being at the remote site may be defined as telepresence [450]. Damasio [120] and Metzinger [336] mentioned that there is a close link between self-experience, selfhood, and the first-person perspective. Metzinger also referred to "The Consciously Experienced First-Person Perspective" to support more complex forms of phenomenal content, such a conscious representation of the *relation* between the person and varying objects in his/her environments.

3.3.2 Issues Affecting Telepresence

A user can perceive a remote environment and navigate or manipulate objects through a teleoperated robotic system; however, several issues can affect such tasks, degrading the sense of telepresence [99, 407]:

Field of View (FoV): Observing the remote environment through a camera's video stream reduces the peripheral vision of the user, and it can negatively affect the spatial perception, compromising manipulation and navigation abilities.

Knowledge of the robot's orientation: Users need to know the position and orientation of the robotic agent in the remote place, as well as the robot's topology (e.g., arm position, body size, and pitch and roll angles).

Camera's view point and *frame of reference*: The placement of the cameras may affect visual perception by providing unnatural views to the user (i.e., compromising pose and position estimation). The egocentric (first-person) perspective comes up as the view for someone present in a space, enhancing, therefore, the sense of telepresence. However, sometimes, an exocentric camera view (third person) may present advantages in the execution of a specific task.

Depth perception: Viewing the remote scenario through a monocular camera can limit the acquisition of significant depth information. The projection of 3D depth information onto a 2D display surface foreshortens or compresses depth cues (e.g., distance underestimation).

Video image quality: Factors like low image resolution, reduced frame rate, or reduced number of colors can make user's remote spatial awareness, target localization, and consequently, response time difficult.

Latency: The time lag verified between the operator's input control action and the observed system response determines his/her control behavior. The aim is a continuous and smooth control operation; however, when the latency increases, the operator adopts the "move and wait" control strategy.

Motion effect: Performing manual tasks on top of a moving platform can be quite challenging and ultimately can induce the operator's motion sickness. Vibrations

and disturbances of the visual feedback can make the operator's input controls challenging.

3.4 Technological and Human Perceptual Factors: Effect on Skill-Based Behavior and Immersive Teleoperation Design

As mentioned in Section 3.2, the activities of human operators rely on three types of performance, namely skill-, rule-, and knowledge-based behavior. The higher the involvement of rule- and knowledge-based layers, the higher is the cognitive workload requested of those operators. Perceptual disturbances can lead to a higher intervention of more rational behaviors, contrasting the reactive mode of skill-based behavior [413]. Thus, it is important to keep activities at the skill-based behavior level where perceptual signs are essential to lower the workload. The present section discusses the technological and human perceptual factors with relevance for immersive teleoperation design to minimize workload.

Given that human vision provides over 70% of all the sensory information used in the interaction with the world, we start by presenting the ideal display specifications in Table 3.1, considering the human eye limits. The comparison of these references with those of current visual mediation systems allows the identification of critical factors for teleoperation.

Table 3.1: Ideal display specifications to match the human eye limits.

Display Propriety	**Range Value**	**Ref.**
Latency	$<$7–15 ms	[31, 313]
Angular Pixel Density	$>$60–200 pixels/degree	[94, 108]
Field of View	210° (H) $\times$ 135° (V)	[428]
Frame Rate	$>$1800 Hz	[114]
Color	10 Millions	[239]
Dynamic Range	1:10^9	[84]

Reliable perception is determinant for HRI task performance, the visual feedback being a major source of information. Actually, users perform better in simple or visually less complex environments (e.g., structured spaces, interactions with few objects and at similar depths, few concurrent tasks) [100, 122]. Related works aiming to minimize workload through visual features' enhancement show positive results [192, 390, 524]. As the task becomes visually more demanding, the involvement of other sensory channels, such as tactile, haptic, and auditory, can contribute to maintaining performance [159]. However, these studies also demonstrate that technological mediation of the visual sensory channel can limit visual features and consequently degrade the user's performance and increase

the user's workload. We present some findings concerning visual interface design that aim to mitigate the mentioned problems. Table 3.5, at the end of this paper, presents a summary of the specs for mediation teleoperation technology based on human capability requirements and based on related works' specs findings for a given a task.

3.4.1 The Human Eye and the Real Scene Resolution

Human visual acuity, that is the power to resolve fine details, relies on optics and neural factors. A person with "normal" visual acuity, i.e., 6/6 vision in meters (or 20/20 in feet), can discriminate the letter E in the Snellen chart at a distance of six meters. Given that the size of these optotypes is subtended in a visual angle of five arc minutes and the eye can resolve the gap between the five horizontal lines of the E letter, i.e., 1/5 of the arc, the human visual acuity is one arc minute.

At a distance of six meters (or 20 feet), the eye can detect an interval lower than 1.75 mm between two contours. The discrimination of a line through optics suggests at least three photo-detectors (stimulated, not stimulated, and stimulated), so for each arc minute, there is a photo-detector. Thus, as one arc minute is equal to 1/60 degrees, there are at least 60 photo-receptors/degree in an angle subtended at the fovea [108] (an analogy with a camera could be 60 pixels/degree).

The retina of the eye is composed of *cone* cells that are sensitive to color and *rod* cells that are sensitive to low light (about seven million cones and 120 million rod cells). It includes the fovea, which is a small pit region of the retina (1.5 mm wide) with a high density of cone cells (exclusively) and where the visual acuity is higher. Fixating on an object implies the movement of the eyes that makes the image fall into the fovea. The result is the sharp central vision essential for human tasks like reading or driving.

Resolution acuity enables identifying two very close lines or contours, and there are people whose vision exceeds 6/6. Human grating resolution ranges from one to 0.3 arc minutes [94, 491]. Related also, *detection acuity* refers to the ability to detect small elements in a well contrasted scenario, e.g., black points on a white background. Studies show that the human eye can detect a single thin dark line against a white background uniformly illuminated and subtended in just 0.5 arc seconds [260, 475] or 0.0083 arc minutes. It is approximately 2% of the diameter of the fovea's cones, which is a mechanism of the visual cortex and not exclusively based on the structure of the retina. Detection acuity seems to result from a spatial temporal averaging process involving the retina's peripheral region (even where rods and cones present a decreasing density) [276]. Considering that the eye's fovea has at least 60 photo-receptors/degree, an ideal display would have, at least, a resolution of 60 pixels/degree to stimulate the cells of the fovea. Of course, in the fovea's periphery, there is a decrease of photo-receptors; however, no one knows at what region of the display the user will look. Therefore, hypothetically, the ideal display would provide a resolution of 12,600 $\times$ 8100 pixels to approximate human's vision (considering both the eye's FOV, 60 pixels/degree $\times$ (210° (H) $\times$ 135° (V)), [428]).

Assuming that an immersive remote visualization system works like a *video see-through HMD*, whose cameras follow the exact user's head orientation, it is possible to analyze such a system without image transmission issues. Let us consider a setup composed of an HMD attached to cameras that stream video from the real world in front. *Video see-through HMD* may acquire images of the real scene using miniature digital cameras and present them through the HMD's LCD or OLED displays. Then, the viewing optics of the HMD adapt these images to the human eye, using an eyepiece lens. Ideally, a *video see-through HMD* should present images with resolutions similar to the real ones, but in practice, that may not be the case. Therefore, the perceived resolution of the real scenario is limited by the resolution specifications of either of these three system components: the video cameras, the HMD display, or the HMD's viewing optics.

3.4.2 Resolution

Generally, higher display resolution leads to better teleoperation performances. Resolution impact is not as noticeable as other limiting factors like latency, but it is significant. Evaluating teleoperation driving tasks at different display resolutions (1600 $\times$ 1200, 800 $\times$ 600, and 320 $\times$ 200 pixels per screen), Ross et al. [423] observed a 23% reduction in the rover's average speed and an increase of 69 times in the average time that the operator stopped the rover for planning or other reasons while comparing the highest and the lowest resolutions.

The path decision was negatively affected by low resolution conditions, as the teleoperator drove slower due to difficulties in distinguishing obstacles. A quality assessment suggested that high resolution improves operator confidence and contributes to the sense of realism and presence. An earlier study analyzed the impact of resolution in the periphery of an HMD in three conditions (64 $\times$ 48, 192 $\times$ 144, and 320 $\times$ 240 pixels) for a simple search and identification object task [511]. Watson et al. found that the lowest resolution was significantly worse than the two higher resolutions considering the criterion of accuracy and search time. Commodity game industry HMDs keep pushing resolution to higher levels of realism and fidelity (e.g., the HTC Vive display has 1080 $\times$ 1200 pixels per eye, while the new model, HTC Cosmos, presents a resolution of 1440 $\times$ 1700 pixels per eye). In [487], two HMDs with different display resolutions were evaluated for echography examinations in a pre-clinical and clinical study: the HM2-T2 (Sony Corp.) with a dual display with 1280 $\times$ 720 pixels per screen and the Wrap 1200 (Vuzix Corp.) with a resolution of 852 $\times$ 480 pixels for each eye. The study showed that the image quality and the diagnostic performance were significantly superior using the HMD with the highest resolution. Laparoscopic surgery using a high resolution HMD enabled superior image quality and faster task performances as opposed to a low resolution HMD model [233].

3.4.3 Frame Rate

The frame rate (FR) is the number of images displayed per time unit, indicating the image refresh rate of the system (frames per second (fps) or Hz). Studies

reveal that generally, a low frame rate degrades operators' performance. Massimo and Sheridan [324] analyzed the efficiency of moving a robot arm to a target via a camera view, on a placement style task (accuracy and speed peg in hole task), for three frame rate levels (3 fps, 5 fps, and 30 fps). They found that increases in the frame rate improved the teleoperation efficiency significantly, and for levels below 5 Hz, there was a significant performance deterioration. In Chen [98], the participant experienced a significant performance degradation in a simulated target acquisition task with a frame rate of 5 Hz. Ware and Balakrishnan [510] assessed the impact of different frame rates (60 Hz, 15 Hz, 10 Hz, 5 Hz, 3 Hz, 2 Hz, and 1 Hz) on a 3D target acquisition task (Fitts' law style task) and also concluded that lower frame rates, 5 Hz or below, decreased users' performance; however, they suggested that 10 Hz was enough for the tested task. They also noticed that the effect of frame rate and the lag were closely related. The impact of frame rate in the sense of presence was analyzed in Meehan et al. [331] for a virtual environment (VE) placement task at 10 Hz, 15 Hz, 20 Hz, and 30 Hz. They found that presence increased significantly from 15 Hz to 20 Hz and kept growing from 20 Hz to 30 Hz. They also reported that lower frame rates, besides impairing the sense of presence can also cause balance loss and a heart rate increase. Claypool [106] found that first-person-shooter video game users significantly improved their target acquisition task performances while the frame rate varied from 3 Hz to 60 Hz. In Ross's rover teleoperation driving task [423], several display rates were evaluated at 10 Hz, 15 Hz, 20 Hz, 25 Hz, and 30 Hz. Negative impacts were more noticeable at lower frame rates, namely on speed and motion perception. There was a drop of 37% in the average rover's speed when the FR condition changed from 20 Hz to 10 Hz. Participants felt comfortable teleoperating at 25 Hz and started experiencing an initial discomfort at 20 Hz due to flickering. They found it difficult to perceive the rover's speed at frame rates under 15 Hz. Chen and Thropp [101] conducted a survey involving 50 studies, and they found that generally higher FR and a small standard deviation of the frame rate benefit users' psychomotor performances ($\geq$17.5 Hz for placement tasks, $\geq$10 Hz for tracing tasks, $\geq$16 Hz for navigation and tracking targets). In summary, they concluded that 15 Hz is a reasonable threshold for most of the tasks, including perceptual and psychomotor. The sense of presence and that of immersion benefit from higher values of FR (60 Hz to 90 Hz), which enable improved realism [377, 433].

3.4.4 System Latency

Latency in teleoperation refers to the elapsed time between a user's action and the consequent system's observed response. Several components of the system can contribute to this overall time lag: user-input device lag, motor actuator lag, video acquisition and display lag, synchronization lag, communication time delay, etc. This occurs in the remote agent control flow and the visual and/or haptic feedback flow.

The operator's control actions, such as head and hand motions, are mapped into the robot's commands using pose-tracking devices and hand controllers. These commands are then transmitted through communication links and executed remotely by the robot's actuators. Considerable delays can occur from the command

generation to its execution, due to motor actuator lag and or the latency of the network. The execution of the commands is viewed through images of the remote scene acquired by cameras and delivered to the operator using the network link. This transmission adds more time delays because of the network latency and the time required to transfer the data images. The time to send the command itself is influenced by bandwidth limitation, and although it is comprised of a few small data packets, it can take a long time when the emitter is far from the receptor (e.g., 2.56 s in the two way light time latency between the Earth and Moon [281]). A head-mounted display (HMD) can also introduce some delays. Immersive technologies related to VR refer to motion-to-photon latency ($l_{m2photon}$) as the time delay between the movement of the operator's head (t_{mov}) and the change of the display screen reflecting the operator's movement (t_{disp_mov}) [313], $l_{m2photon} = t_{disp\_mov} - t_{mov}$.

Lag is a crucial element to allow the operator's brain to think that he/she is physically interacting in the remote place, experiencing the sense of telepresence. For that, immersive interfaces should provide stimuli that trick the sensory system, ensuring consistency between vision, internal sensory information (e.g., vestibular, proprioceptive), and cognitive models. The non-compliance of expectations can break the experience of presence, negatively affecting the task performance and causing motion sickness and nausea [140, 332]. For example, when a user stops rotating his/her head, he/she expects not to see moving images on the display, or when he/she starts head motion, he/she is supposed to see images moving immediately and not after a delay. Research [313] has found that the *just noticeable difference* (JND) for latency discrimination should remain below 15 ms. The response of the system to head motions should be as fast as the human vestibulo-ocular reflex, the response of which ranges from 7 ms to 15 ms [31]. During head motion, this reflex stabilizes the retinal image at a chosen fixated point by rotating the eyes based on vestibular and proprioceptive information compensating the motion. In fast movements of the eye, such as *saccade movements* (up to 900°/s), the retinal image may become blurred. To avoid processing unclear information, the brain has a mechanism, called a *saccadic mask* [500], that temporally suspends the visual processing so that the motion of the eye or the gap in visual perception is unnoticed. *Saccade movements* enable fovea rendering approaches, optimizing display resources, and additionally, the *saccadic mask* suggests some tolerance in current VR eye-tracking-to-photon latency [12, 479].

NASA researchers [411] studying HMDs to support synthetic enhanced vision systems in flight decks found that for extreme head motions, such as higher than 100 deg/s, the system latency requirement must be below 2.5 ms. Anyway, for demanding tasks requiring headset displays with a large field of view and high resolutions, the recommended system latency is less than 20 ms. For example, Oculus RIFT developers recommend not exceeding 20 ms for motion-to-photon latency in a VR application [376]. In recent developments, they enabled reducing its tracking subsystem latency to 2 ms, allowing more time for the overall system lag. The Oculus Rift CV1 [375] and HTC Vive [218] headsets have a refresh rate of 90 Hz, meaning that they can update their displays every 11 ms.

Human performance studies show that a person can detect latency from 10–20 ms [140]. Lags occur in teleoperation of mobile robots or robotic arms when infor-

mation is transmitted across a communication network (i.e., end-to-end latency). When the latency is higher than 1 s, the operator tends to change his/her control approach to “move and wait”, rather than continuously commanding, predicting, and trying to compensate the delay [272].

The time delay factor may differently affect the performance of a specific task. The negative effect extends to over-actuation for the variable delay, affecting robot-to-operator direction information flow more than the other way. In a telemanipulation task related to laparoscopy surgery, negative effects were observed in the system usability and the performance of experienced surgeons for a delay $\geq$105 ms [264]; in another task related to telesurgery (precision, cutting, stitching, knotting), a degradation in accuracy, precision, and performance was reported for a delay $\geq$300 ms [305]; in a driving simulation task, a performance degradation was observed for a delay $\geq$170 ms [164]; in social interaction with mobile telepresence robots, the recommendation for teleoperation commands' latency is under 125 ms [496]; in a real teledriving task (six wheel all-terrain rover), evaluating the average speed and the average time stopped for path decision, the negative effect appeared for latency $\geq$480 ms [423]; in a car teledriving task on city roads at 30 Km, while analyzing the tracking line, obstacle detection, and performance, problems for delays $\geq$550–600 ms were reported [179, 216].

The mitigation of the effects of latencies related to visual feedback in teleoperation systems may involve a predictive display. The goal is to provide immediate visual feedback to the operator, displaying a scene model animated by the commands. Meanwhile, and in parallel, this scene model is updated with measures acquired by the remote teleoperator sensors [43, 348].

3.4.5 Field of View

The field of view (FOV) refers to the size of the visual field observed. Manipulation studies typically compare narrow viewpoints with wide panoramic views, revealing that narrow FOVs result in difficult navigation [47]. For example, they can limit the acquisition of contextual information about the space around a rover, compromising the perception of distance, size, and direction, having difficult cognitive map formation, and being more demanding of operator memory and attention.

In [423], the FOV had a significant impact on rover teleoperation tasks. Different horizontal FOVs were tested, 40°, 60°, 120°, and 200°, and it was found that an horizontal field of view (HFOV) of 200° allowed average speeds 40% higher than with an HFOV of 40°. The average stopped time with an HFOV of 40° was two times greater than with an HFOV of 120°. Path decisions were compromised at lower FOVs, and operators reported that a wide field of view benefited the situational awareness while enabling higher speeds. In [397], the field of view was manipulated at two levels, 30° (narrow FOV) and 60° (wide FOV), while navigating a virtual UGV. Widening the FOV resulted in a superior performance benefit, leading to lower times to complete obstacle navigation tasks, decreasing the number of collisions and the number of turnarounds, and having higher piloting comfort. In [37], three different levels of FOV (48°, 112°, and 176°) wearing

an HMD while walking, avoiding obstacles, estimating distances, and recalling spatial characteristics were analyzed. He reported that users' walking was more efficient with a wider FOV, but did not find significant effects on distance estimation (spatial understanding), on user's balance, nor on recalling the characteristics of the environment (i.e., memory recall).

Wider to moderate fields of view tend to contribute positively to performance. Subjectively, FOV significantly improves teleoperators' situational awareness and perception of robot position and motion. However, moderation is advised as a wider FOV can increase motion sickness [47, 441, 467]. The causes of motion sickness can be related to the fact that a wider FOV increases ocular stimulation and motion in operators' peripheral vision.

Video cameras have been used in robot navigation to perceive the environment, although this process can suffer from the "keyhole" effect [99, 520]. This means that just part of the environment can be acquired and presented to the human operator. Usually, the operator overcomes this situation with extra effort, manipulating the cameras to survey the environment and gaining similar scene awareness to direct viewing.

Human eyes provide a horizontal field of view of 210° by 135° in the vertical [428]; however, stereo central vision relies on the overlap of the FOV of both eyes, and it is around 114°. Thus, as a result of $210^\circ/2 + 114^\circ/2$, an ideal HMD should provide a view of 162° horizontally for each eye [114].

Robotic remote operations include remote spatial and motion perception, navigation, and remote manipulation. A limited field of view (FOV) degrades the remote perception in several ways; however, it affects tasks differently. Tasks like navigation, self-location identification in space, and target detection are negatively affected due to the video feedback constraints [123]. Manipulation tasks involving action in a limited space can overcome the FOV limitation by gathering new points of view of the scene with more or less extra effort.

Operators that perform navigation tasks based on a fixed camera mounted on a mobile robot tend to perceive the environment through a stack of images for which the points of view correspond uniquely to the robot's path. This "cognitive tunneling" effect present in egocentric visualization approaches contrasts with exocentric systems in which the frame of reference (FOR) remains unchangeable, requiring less mental transformation. In navigation tasks, operators usually focus their attention on a destination point without worrying much about the surrounding environment. Either in navigation or manipulation tasks, a limited FOV can cut crucial distance cues and limit user depth perception [517]. However, Knapp [261] proved that an HMD's limited field of view has no effect on perceived distance, reported in virtual environments.

Navigation performance research has identified several problems related to restricted FOV: drivers with difficulties in judging vehicle speed, object perception, time to collision, obstacle location, and start procedures to curve. The peripheral vision contributes to speed perception, lane following, and lateral control avoidance. Wider FOVs are common solutions in teleoperation indirect navigation. This broadens the view of the scene either with wide angle cameras or using extra cameras

on-board to cover the surrounding space. Nevertheless, such FOV increments, particularly when based on wide angle cameras, might induce a faster perception of speed. This effect is related to scene compression and quite often leads the operators reducing their speed. Researchers have pointed out that the scene distortion and resolution decrements, related to scene compression, may increase operator cognitive workload, degrade object localization tasks, and cause motion sickness symptoms [467]. These authors conducted a direct viewing driving and an indirect video driving of a military car. Three internal tiled LCDs provided a 100° panoramic view returned by a camera array mounted in the front roof of the car. They tested three sets of camera lenses, providing 150° (near unity), 205° (wide), and 257° (extended) camera FOVs. They found that map planning performance and spatial rotation improve with a wider FOV. Wider FOVs had an effect on spatial cognitive functions similar to peripheral cues for direct viewing. The best performances were achieved using vision displays with the FOV close to direct vision, using systems that enabled electronic adjustment of the FOV.

3.4.6 Depth Perception

The majority of human interactions depend on space and motion perceptions. This ability enables navigation in an environment and handling objects. *Depth perception* is a filtering process that enables us to perceive the world's tridimensionality. Humans can identify information in images and correlate it with depth in the scene, using both psychological and physiological cues [294]. For example, if an object X partially covers another one Y, then object X is inferred as the nearest to us (i.e., occlusion). These different depth cues can be classified into three groups according to the sensory information: *oculomotor cues*, *monocular cues*, and *binocular cues*.

Oculomotor cues for depth/distance are based on:

- *vergence*, which results from the sense of inward movement that occurs when the two eyeballs rotate to fixate near or distant objects, and
- *accommodation*, the change in the eye lens's shape required to focus on the object at different distances. In fact, accommodation is a monocular cue, based on the kinesthetic sensation of stretching the eye's lens and sensing the tension of the eye's muscles [181].

Monocular cues provide distance/depth information by viewing the scene with one eye only. Besides accommodation, they include:

- *pictorial cues*, which identify depth information in a single two-dimensional image, and
- *movement-based cues*, which extract depth from the perceived movement.

Pictorial cues include (a) *occlusion*, (b) *relative height*, (c) *relative size*, (d) *perspective convergence*, (e) *familiar size*, (f) *atmospheric perspective*, (h) *texture gradient*,

and (i) *shadows*. The mentioned cues consider a stationary observer; however, new depth cues arise when a person rotates his/her head or starts walking. *Motion-produced cues* include (1) *motion parallax* and (2) *accretion and deletion*. *Motion parallax* results from the fact that when we move, nearby objects seem to move faster than the more distant ones. As we move and due to perspective projection, the displacement of the projection of the objects in the retinal image depends on the distance. *Accretion* and *deletion* result from sideways movements of the observer, leading some object parts to become hidden, and others visible. These cues, based on both occlusion and motion parallax, arise when superimposing surfaces seem to move in relation to one another. Two surfaces, at different depths, can be detected using these cues.

Binocular cues use information from both eyes to provide important distance/depth information. *Binocular disparity* (*binocular parallax*), i.e., the difference of corresponding image points in both eyes, combined with the geometry of the eyes (convergence) enable localizing a tridimensional point (triangulation). *Stereopsis* is the feeling of depth that results from information provided by binocular disparity.

In teleoperation, by changing the point of view of remote monocular or supporting stereo vision allows operators to get important cues for depth perception, crucial for navigation and telemanipulation. In telesurgery or laparoscopic tasks, the observation through an indirect method affects the surgeon's depth perception and diminishes his/her eye-hand coordination ability. Depth perception problems result from accommodation and convergence inconsistencies, the lack of shadows in endoscopic video images, and the lack of movement parallax and stereovision. Eye-hand coordination problems are due to the distance location of the monitor and because the video images of the doctor's hand movements appear mirrored, rotated, and amplified [87]. To overcome these contingencies, surgeons must have intensive, long, and specialized training. Technological solutions [80][42] include flexible and motorized endoscopes that compensate misorientation and provide movement parallax through probe positioning controlled by the surgeon's head movements.

3.4.7 Frame of Reference

Many teleoperation visuomotor tasks such as telemanipulation or telesurgery can be highly demanding for the operator because the frames of reference for vision and action are misaligned [127, 323, 504]. The operator's control actions are harder when there is no alignment between his/her point of view, the input device, and the coordinate frame of the robot. Any mental transformation, such as rotation or translation, imposed by the input/output interface to control the robotic system can increase mental workload and consequently decrease task performance. Thus, the interface should be intuitive enough, so when the operator acts on the input device, the robotic agent should move in the expected direction coherently with the input gesture. Moreover, the operator should observe or sense the robotic agent moving in the expected direction. Researchers have shown that the angular misalignment between the visualized axis of rotation and the controller's hand axis rotation can cause control response delays and decrease

accuracy in telemanipulation tasks [307, 493]. Human path following performances degrade for non-orthogonal angles relating control and visual frames [128]. The eye-hand coordination tasks should be as natural as if the operator were physically in the remote place. In teleoperation systems, a one-to-one correspondence between control and display devices is desirable, so for a given control movement, the consequent change in the image display should appear to be in the same relative place [449]. Humans can easily adapt to a lateral displacement between the expected point of view of a movement and the observed movement through a 3D perspective display, if less than 15° [390]. For higher angles $>15^\circ$, the adaption decreases.

One way to achieve the alignment between the operator's point of view, the input/output device coordinates, and the robot's coordinate frame is to design the teleoperator agent anthropomorphically, or ensure that the extent of the control action is easily perceived and mapped through the feedback of the agent (e.g., ensure that a steering wheel rotation is proportional to the robot's linear displacement in the diagonal). The egocentric visual feedback suggested in the present work aims to fulfill one of these natural and one-to-one correspondences (visual, haptic, proprioceptive).

Egocentric or First-Person View

Why provide different views of the remote site? Depth cues provided by human's binocular system are important for the perception of the environment. In the absence of pure stereo vision, different points of view can contribute to the visual perception. Additionally, mental models are matched against reality through egocentric information arriving from the different sensory modalities [336]. For example, an observer expects to see the scene's elements changing when he or she moves. He/she expects that parts of the scene initially occluded become visible and others become occluded. Moreover, it is expectable that the views of a nearby object or part of it present higher changes in images than those more distant. Thus, all these reality cues need to be replicated by the egocentric visualization system because the human brain uses them to estimate distances.

In traditional teleoperation systems, the operator controls the remote camera's orientation manually. He/she has to control two or three degrees of freedom of the camera and, additionally, several degrees of freedom of the robot, which increases the operator's efforts. To address these problems, we suggest a virtual cockpit for the operator where he/she experiments with the sensation of controlling the robot inside of it, i.e., telepresence [23, 174]. Operators can browse the vehicle's surroundings as if they were aboard it and simultaneously realize that the controls and instrumentation panel at his/her disposal present coherent feedback. This is implemented using a head-mounted display (HMD) whose orientation is used to control a pan and tilt unit (PTU) that supports the remote camera. With this solution, the user's head movements are implicitly transposed into camera movements. The approach enables the operator to look around, just by naturally rotating his/her head, viewing what was supposed to be seen in the remote place with that movement. By superimposing real video feeds with virtual elements,

synchronous with the operator's point of view, we contribute to the visual perception of being immersed in the remote place.

3.4.8 Working Memory

Spatial memory is part of our cognitive system. Those tasks, involving higher demands from human short- and long-term memory, typically result in higher mental workloads. *Working memory* refers to the mind's capability to maintain and manipulate important information to perform complex tasks such as reasoning, comprehension, and learning. Badlley's model [48] evolution proposes a four component model: a central executive system along with three short-term storage subsystems, the visuo-spatial sketchpad (processes visual-spatial information), the episodic buffer (with limited capacity), and the phonological loop (processes verbal-acoustic information). For example, route planning or thinking about time involve the visuo-spatial sketchpad subsystem and the central executive system. Maintaining information in the working memory involves a rehearsal process [194, 248]; however, working memory overload or significant distractions can interrupt this process. For example, operators that rely on working memory to label objects or to recall the spatial positions and orientation of the objects tend to experience higher cognitive workloads. Moreover, working memory data are necessary for mental transformations involving visual and proprioceptive frames of reference (rotations and translations). The human cognitive system uses the *working memory* to retain new information temporally for processing or to store in long-term memory. It is shown that working memory has a limited capacity and holds the information for a short time (a few seconds). The manipulation of this information enables reasoning and decisions and shapes event behavior. Miller [344] introduced the "chunks" concept, referring to any set of information to be associated with long-term memory. Moreover, he defended that human working memory can hold just 7 ± 2 "chunks" or "items" of information. More recent studies point to a number of "chunks" dependent on the type of information. Managing multiple and separate frames of reference, controlling multiple actuators, distributing attention over multiple instruments, memorizing the spatial position of multiple targets due the limitation of the field of view, register numbers with several digits, or changing sequential procedures frequently due to external unpredictable events are some examples that can compromise tasks. As a conclusion, teleoperation systems should avoid making a person recall more than 7 ± 2 "chunks" to evolve through a task.

3.5 Immersive Interface Testbed: An Evaluation Example

Traditional interfaces for teleoperated robots integrate multiple displays and controls, requiring specialized training by the operators. Such additional effort is not easy for a non-specialized operator, making them skeptical about technological advances. As noted in the previous sections, common commands, or perceptual

actions, not properly executed or reflected by the system, can become distracting factors. Additionally, letting operators experience the sensation of being in the remote environment (telepresence), with minimum disturbances of the mediation technology, makes the task performance more natural. Simpler operation control is achieved through the combination of telerobotics and telepresence. We propose an immersive interface testbed that enables the manipulation of perceptual factors and an analysis of their impact on immersion, presence, task performance, and workload. Ensuring perceptual factors that keep the operator's activities at the skill-based behavior level, with less intervention from higher decision levels, saves energy and enables focused tasks [413]. The approach suggests a familiar and traditional workspace where the operator can perform manipulations while disposing of a wide egocentric field of view and precise control of the actuator. The solution includes: providing a *first-person view* (*FPV*) of the workspace scene; enriching the operator's spatial perception, by virtually disposing of the mediated sources of information in conforming with natural layouts; providing consistency between action-movement and the human's sensory system (visual, haptic); transforming some explicit commands into implicit ones; providing a natural point of view of the remote site. These approaches contribute to reducing physical and cognitive workload while improving task performance. The following text describes the evaluation methodology for the proposed teleoperation immersive interface compared with the traditional approaches.

3.5.1 Participants

The experiments were performed at the Polytechnic Institute of Tomar and the Institute of Systems and Robotics, Coimbra, Portugal, with 25 participants, two females and 23 males. The participant group included students and researchers in fields such as engineering and computer science, with an overall average age of 30.3 years and a standard deviation of 8.65 years. All participants reported normal or corrected to normal vision. None of them had prior knowledge of the experience or technologies involved. Participation was voluntary, and ethical research principles were observed.

3.5.2 Experiment Design

To evaluate the effect of immersive technologies in teleoperation spatial perception, we designed several evaluation procedures where participants were invited to accomplish several hand-eye coordination tasks while their performances were analyzed. We compared immersive visual feedback against traditional visual feedback based on a traditional monitor and a remote fixed camera (Figure 3.2 exemplifies the setups used for the pick-and-place task). The proposed immerse visual feedback uses a head-mounted display (HMD) for visualization purposes while, the pose of the user's head is used to control the remote camera orientation, gathering different points of view.

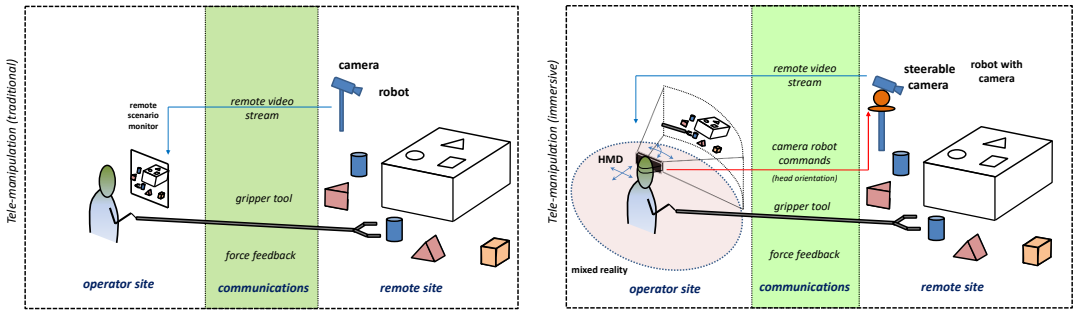


Figure 3.2: Traditional teleoperation setup (remote visualization through a fixed monitor and camera) vs. immersive interface (point of view transfer; the head-mounted display (HMD) controls the remote camera).

The developed setups manipulate the type of visualization of the remote environment and the control of the remote camera orientation for several tasks (see Table 4.1).

Table 3.2: The two different test setup combinations for semi-teleoperation tasks. Fixed view, first-person view (FPV).

Test	**Display**	**Camera Orientation**	**Stick/Arm Control**
1	Traditional 1 Monitor	Fixed	stick/gripper haptic feedback
2	Immersive via HMD (FPV)	head orientation from HMD IMU	stick/gripper haptic feedback

Each participant performed a given task using both setups. The evaluation consisted of analyzing a set of related performances (quantitative measures), recorded during the experiment, and the answers to a short questionnaire (subjective measures) given after each trial. Statistical significance was assessed using repeated measures (within subjects) ANOVA analysis.

However, given that the immersive visual interface and the traditional visual feedback interface differ in terms of setup components, a control experiment was conducted to assess the disturbance factor introduced by each component of the interface.

For a typical pick-and-place task, users should grab five blocks of different shapes and insert them into a box through the respective shape hole. This task was accomplished standing in front of the workspace, using the right hand to manipulate the blocks and using different media interfaces that introduced displacements between visual and haptic frames.

3.5.3 Apparatus

The study was conducted using two types of visualization interfaces to perceive the workspace: a *traditional interface* and a *immersive interface*.

The *traditional interface* consisted of one 22" LCD monitor, the Samsung SyncMaster 2233RZ with a native resolution of 1680 $\times$ 1050 and vertical refresh rate of 120 Hz. This monitor displays images acquired through a webcam with a wide field of view, in the remote workspace from a single fixed point of view. The webcam used is the HD Logitech C270 with an optical resolution of 1280 $\times$ 960 pixels (1.2MP), configured to capture video (4:3 SD) with 800 $\times$ 600 pixels, a focal length of 4.0 mm, and a field of view (FOV) of 60°.

The *immersive interface* consisted of one HMD, the Oculus Rift DK2, with a display resolution of 960 $\times$ 1080 per eye, an OLED screen of 5.7″, a refresh rate of 75 Hz, a persistence of 2 ms, 3 ms, full and viewing optics with a field of view of 100°. Internal tracking includes gyroscope, accelerometer, and magnetometer sensors with an update rate of 60 Hz. It refines HMD's pose with a positional tracking sensor based on a near-infrared CMOS sensor. The HMD controls the pan and tilt unit coupled with a webcam, the full HD Logitech C920, with an optical resolution of 1920 $\times$ 1080 pixels, configured to capture video with 800 $\times$ 600 pixels and an FOV of 78°. The pan and tilt unit supports this camera and orients it according to the HMD's pose orientation, that is conforming to the user's head orientation. The Computer-Controlled Pan-Tilt Unit (PTU 46-17.5) from Directed Perception has two-step motors for the pan and tilt movements: load capacity over four lbs, speeds over 300°/s, resolution of 3.086 arc minutes (0.0514°).

3.5.4 Evaluation Procedure

To analyze the effect of immersive and egocentric visual feedback, participants were asked to perform three hand-eye coordination tasks using two visual interfaces: traditional interface: fixed monitor, a single point of view with a wide view scene ($Fix_{Display} + Fix_{Cam}$) vs. the immersive interface: HMD and controllable point of view ($Rift_{Display} + Mov_{Cam}$).

Task 1, touch: Users should press on several key blocks using a 1 m stick according to a random sequence defined by the computer (the time to accomplish the task was measured and block touch mistakes) (see Figure 3.3).

Task 2, pick-and-place: Users should grab five blocks of different shapes using a manual gripper (80 cm long) and insert them into the respective hole of a wood box (the time to accomplish the task was measured; insertion difficulties due to spatial perception errors or handling were interpreted as delays) (see Figure 3.4 where blocks with different shapes like cylinders, cubes, and triangular prisms only enter in the right hole shape; time recording starts with the first block grab and ends with the last insertion).

Task 3, path following: Users should follow a predefined 3D path with a metal loop at the tip of a stick (the time to accomplish the task was measured and the number

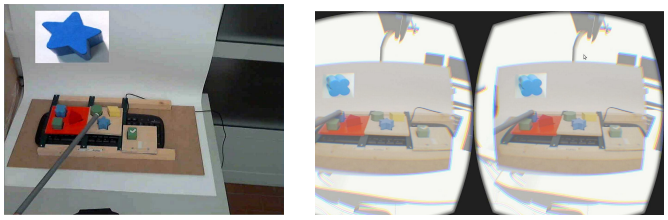
of hits loop/wire recorded). This setup consists of one metallic pipe with curved and straight paths, where the user should move the metal loop along the pipe, avoiding electric contact between both (see Figure 3.5).

Participants started randomly, either with Interface 1 or 2, to mitigate the effect of the learning factor. Task 2 and Task 3 were performed in a random order for the same reason.

The evaluation consisted of analyzing a set of performance-related parameters, which were collected during the experiments, and the answers given to a short questionnaire after each trial. The collected parameters, the questionnaire, and their analysis are presented in the remainder of this section.

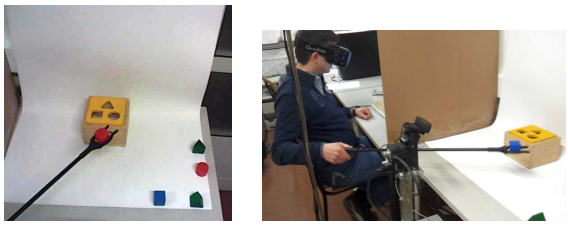
The procedure can be summarized as:

1. The participant is instructed about the task objectives and procedures.
2. Execute the trial (Tasks 1, 2, 3) with one of the two setups randomly selected.
3. Fill in the questionnaire about the user experience.
4. Repeat until four trials are complete.



(**a**) (**b**)

Figure 3.3: Task 1 setup used to compare performances. The user using a hand stick presses a sequence of key blocks defined by the computer at random. Interface view shows the next block to touch at the upper left corner of the display, after each correct touch. (**a**) Traditional interface vs. (**b**) immersive interface.



(**a**) (**b**)

Figure 3.4: Task 2 setup used to compare performances: The user grabs blocks with a hand gripper and places them into a wood box through the corresponding hole, using (**a**) the traditional interface ($Fix_{Display} + Fix_{Cam}$) vs. (**b**) the immersive interface ($Rift_{Display} + Mov_{Cam}$).

(**a**) (**b**)

Figure 3.5: Task 3 setup: The user follows a predefined 3D path holding a stick with a metal loop through a thick metallic pipe avoiding contact. An LED light signals the electric contact. (**a**) Traditional interface vs. (**b**) immersive interface.

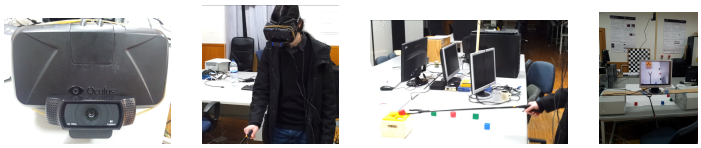
The experiments consisted of a typical pick-and-place task using different media interfaces randomly following the repeated measures approach (within subjects); see Figure 3.6. Users were asked to perform the task using:

- Direct vision (stereo binocular) and his/her bare hand (the baseline);
- Direct vision (stereo binocular) and a manual gripper;
- HMD indirect vision (mono biocular) and his/her bare hand;
- HMD indirect vision (mono biocular) and a manual gripper;
- Monitor indirect vision (mono biocular) and a manual gripper;

In the *"direct vision (stereo binocular)"* setup, users used direct visualization to grab the blocks and insert them into the respective holes of a wood box, either using their hand or a manual gripper (80 cm long).

In the *"HMD indirect vision (mono biocular)"* setup, users used an HMD with a monocular video camera fixed in front of it, creating a video see-through HMD. The users used their hand or a manual gripper to manipulate the blocks and could move their head freely (position and orientation). Because the camera and HMD display have different resolutions and FoVs, the 1:1 magnification was not suitable. To reflect the true size of the objects, the user adjusted the viewing image to half of the original size (looking either directly at the scene or through the "video see-through HMD").

In the *"monitor indirect vision (mono biocular)"* setup, the user performed the pick-and-place task standing in front of the workspace using a manual gripper to manipulate the blocks and while looking through a fixed LCD monitor that was displaying a video streaming of the scene using a monocular camera. The users were seated with a LCD monitor at their eyes' height (monitor hanged) so they could manipulate under the monitor with the manual gripper.



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

Figure 3.6: (**a**) "HMD indirect vision (mono biocular)": a video see-through HMD; (**b**,**c**) user performing a pick and place task using the setup "HMD indirect vision (mono biocular) + gripper"; (**d**) user performing a pick and place task using the setup "fixed monitor vision (mono biocular) + gripper".

3.5.5 Measurements and Questionnaires

The usability evaluation was performed in two parts: performance-related measures and user subjective evaluation using a questionnaire.

Regarding the analysis of the performance, we measured the following variables directly from the instrumented object and/or using a third observer to keep records.

Time: task completion time for the procedure. This integrates delays due to depth perception errors.

Hits: collisions with objects due to erroneous spatial perception.

Path following precision: 3D path following precision with the tip of the stick (electric contact and time to accomplish)

The goal is to identify task performance manipulation errors due to impaired visualization, shaking, etc.

For the subjective evaluation, a questionnaire was created inspired by the IBM Computer Usability Satisfaction Questionnaire [282], NASA-task load index (NASA-TLX) [197] and based on the presence questions of Slater, Usoh, and Steed [464, 499]. The participant feedback classified, on a seven point Likert scale, factors like usability, easiness, control precision, fatigue, realness, telepresence, and embodiment feeling. Questions were translated into the Portuguese language to simplify their understanding. The eight questions to answer were divided into two groups as follows:

Usability and task load questions:

Q1: I visualized the workspace ... (1=without any difficulties, 7=with difficulties)

Q2: Was the task tiring? (1=Not tiring, 7=Very tiring)

Q3: I managed to manipulate objects quite accurately (1= Not at all, 7=Very much)

Q4: The workspace visualization did not difficult object manipulation (1=Disagree, 7=Agree)

Immersion presence questions:

Q5: I forgot that I used an indirect technological visualization device (1=Disagree, 7=Agree)

Q6: I had a clear perception and total control of stick's movements? (1=Not at all, 7=Yes totally)

Q7: I perform better when: (1=I move my head, 7=I do not move my head)

Q8: I know where the objects are because I can touch them. (1=Disagree, 7=Agree)

3.6 Results and Discussion

3.6.1 Results

The performance of the user is measured in terms of time spent and mistakes occurring while executing the task, for each of the proposed interfaces. Additionally, the questionnaire score results enabled a qualitative evaluation.

Task performance related measures:

Figure 3.7 depicts the mean task-time performance and standard deviation for the key block sequence touches while using the traditional interface and the immersive interface, Task 1 setup. Keyb_Seq. 1 corresponds to touching the first four blocks, Keyb_Seq. 2 to touching the second four blocks, and Keyb_Seq. 1+2 the time spent to correctly touch the eight blocks. The second trial Keyb_Seq. 2 is faster due to the learning process.

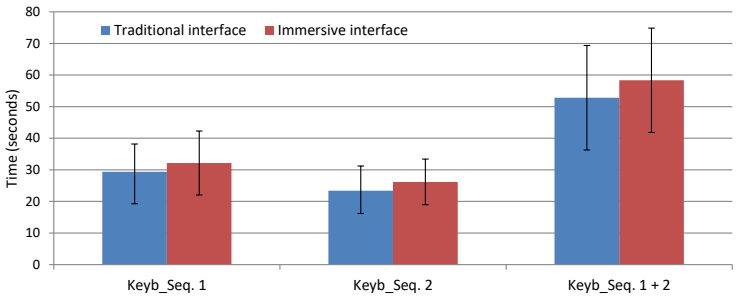


Figure 3.7: Task 1: mean task-time performance of participants while pressing on a sequence of key blocks determined by the computer: Keyb_Seq. 1 (first round), Keyb_Seq. 2 (second round), and Keyb_Seq. 1+2 (sum of both sequences).

These task performance measures are also available in Table 3.3, where for Task 1, μ stands for mean task-time performance in seconds and σ is the standard deviation, and the respective subscript t stands for traditional and i for immersive. According to the one-way repeated measures ANOVA (analysis of variance) tests, the round comparisons of Task 1 are not statistically significant:

Keyb_Seq. 1	$F_{1,8} = 0.212$, $p = 0.657$;
Keyb_Seq. 2	$F_{1,8} = 0.343$, $p = 0.574$;
KeybSeq. 1+2	$F_{1,8} = 0.281$, $p = 0.609$;

where F stands for the F-statistic, p the p-value and should be $<$0.05 for significant comparisons. Basically, in Task 1, the immersive interface has a poor performance when compared with the fixed camera setup, although it is not a significant difference. Even in trials where participants had some manipulation training first with the $Fix_{Display} + Fix_{Cam}$ setup, there were no changes. An explanation of this fact is that the tested workspace is small, fits all in the field of view of the user, and he/she does not feel the need to move his/her head to get better views.

Figure 3.8 depicts the mean performance times (seconds) and standard deviation for the Task 2 setup. It shows that users perform the pick-and-place task faster while using the immersive interface than when using the traditional interface. The values are also presented in Table 3.3, and the comparison involving 20 participants is statistically significant. It was assessed using the one-way repeated measures (within subjects) ANOVA analysis: $F_{1,19} = 7.95$, $p = 0.0109^*$ (the asterisk indicates a significant comparison, $p < 0.05$).

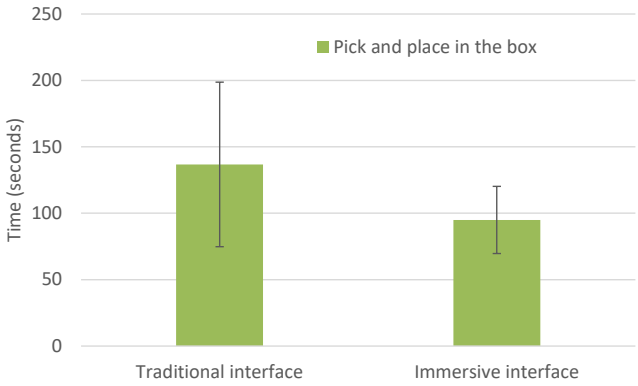
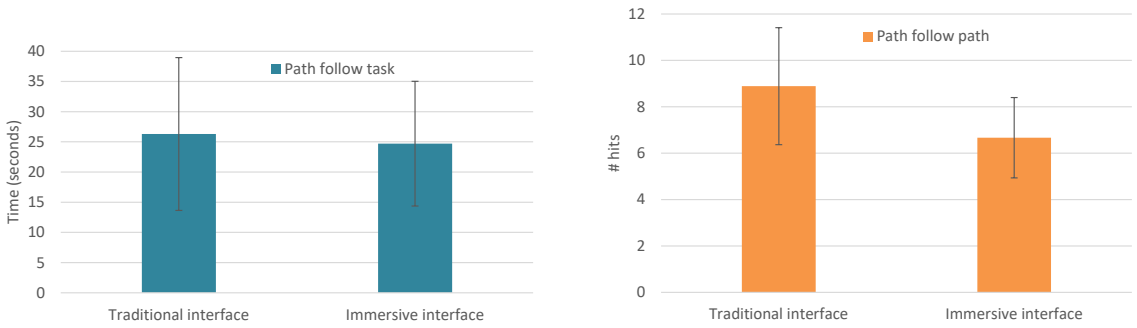


Figure 3.8: Task 2: mean task-time performance for picking and placing blocks in the box's holes.

Task 3's mean time performance measurements are presented in Figure 3.9a and in Table 3.3. The mean time performance using the traditional interface is similar to the immersive interface, with $F_{1,14} = 0.037$ and $p = 0.85$. Nevertheless, for this task, a lower number of hits indicates better performances, and it occurs while using the immersive interface in opposition to the traditional interface; see Figure 3.9b. It has a statistically significant difference: $F_{1,16} = 4.747$, $p = 0.044^*$.



(**a**) (**b**)

Figure 3.9: Task 3: (**a**) mean task-time performance to follow a 3D path with a metallic loop avoiding contact with the guiding pipe; (**b**) mean hits.

Table 3.3: Mean performance measures summary of Tasks 1, 2, and 3.

			Traditional Interface	**Immersive Interface**
Task 1	mean time performances	Keyb_Seq. 1	$\mu_t = 29.39, \sigma_t = 8.78$	$\mu_i = 32.15, \sigma_i = 10.13$
		Keyb_Seq. 2	$\mu_t = 23.39, \sigma_t = 7.82$	$\mu_i = 26.18, \sigma_i = 7.22$
		KeybSeq. 1+2	$\mu_t = 52.78, \sigma_t = 16.55$	$\mu_i = 58.33, \sigma_i = 16.51$
Task 2	mean time performances	*	$\mu_t = 136.75, \sigma_t = 61.93$	$\mu_i = 94.95, \sigma_i = 25.28$
Task 3	mean time performances		$\mu_t = 26.75, \sigma_t = 14.31$	$\mu_i = 25.50, \sigma_t = 11.52$
	mean hits	*	$\mu_t = 8.8, \sigma_t = 2.52$	$\mu_i = 6.6, \sigma_i = 1.73$

Control task:

Comparisons between individual disturbance factors introduced by each mediation technology (e.g., visual, haptic) while in a pick-and-place box task (Figure 3.6) are presented in Figure 3.10 regarding the mean task-time performance.

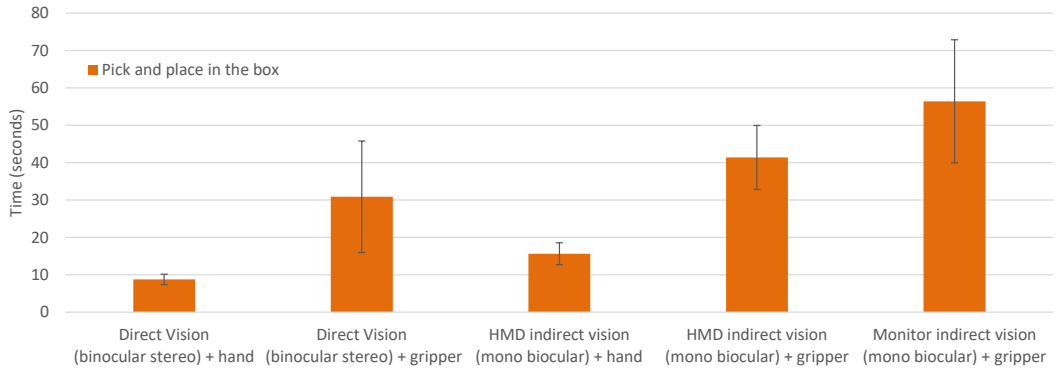


Figure 3.10: Mean task-time performance: pick and place in the box task. Comparison of individual disturbance factors introduced by each mediation technology (visual, haptic, shift between kinesthetic and visual feedback).

The one-way ANOVA test for the five factors shows a statistically significant com-

parison: $F_{4,32} = 24.05$, $p < 0.001^*$. The effect of the four disturbance factors on the performance time of eight participants was examined, regarding setup conditions "direct vision (binocular stereo) + hand" (F1), "direct vision (binocular stereo) + gripper" (F2), "HMD indirect vision (mono biocular) + hand" (F3) and "HMD indirect vision (mono biocular) + gripper" (F4). Its one-way repeated measures ANOVA analysis reveals that the comparison is statistically significant: $F_{3,21} = 23.62$, $p < 0.001^*$. Additionally, post-hoc tests and the pairwise multiple comparisons show which factors' means are significantly different and are summarized in Table 3.4.

Table 3.4: Characterization of individual disturbance factors introduced by each mediation technology.

F1	**F2**	**F3**	**F4**	**F5**
Direct vision (binocular stereo) + Hand	**Direct vision (binocular stereo) + Gripper**	**HMD indirect vision (mono biocular) + Hand**	**HMD indirect vision (mono biocular) + Gripper**	**Monitor indirect vision (mono biocular) + Gripper**
$\mu = 8.75$	$\mu = 30.875$	$\mu = 15.625$	$\mu = 41.375$	$\mu = 56.4$
$\sigma = 1.39$	$\sigma = 14.91$	$\sigma = 2.92$	$\sigma = 8.56$	$\sigma = 16.47$
$F1F2F3F4$: $F_{3,21} = 23.62$, $p < 0.001^*$				
	$F1F2$: $F_{1,14} = 17.45$, $p < 0.001^*$	$F1F3$: $F_{1,14} = 113.023$, $p < 0.001^*$	$F1F4$: $F_{1,14} = 36.07$, $p < 0.001^*$	
		$F2F3 : F_{1,14} = 8.05$, $p = 0.013*$	$F2F4 : F_{1,14} = 2.98$, $p < 0.106$	
			$F3F4$: $F_{1,14} = 64.71$, $p < 0.001^*$	

Qualitative evaluation based on the user questionnaire:

Figure 3.11 presents a comparison of the scores for each question while performing Task 2 and Task 3 ("pick and place in the box task" and "path following task") using either the traditional interface ($Fix_{Display} + Fix_{Cam}$) or the immersive interface ($Rift_{Display} + Mov_{Cam}$).

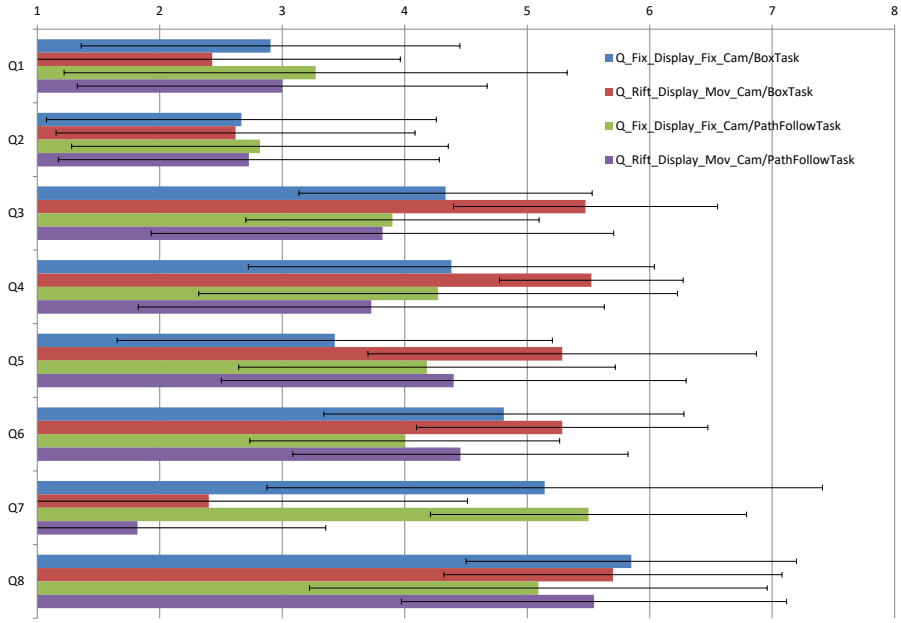


Figure 3.11: Comparison of mean scores from the user questionnaire feedback for the pick and place in the box task and 3D path following task, Likert scale: one to seven. Q1: I visualized the workspace (without any difficulties/with difficulties); Q2: Was the task tiring? (Not tiring/Very tiring); Q3: I managed to manipulate objects quite accurately (Not at all/Very much); Q4: The workspace visualization did not difficult object manipulation (Disagree/Agree); Q5: I forgot that I used an indirect technological visualization device (Disagree/Agree); Q6: I had a clear perception and total control of stick's movements? (Not at all/Yes totally); Q7: I perform better when: (I move my head/I do not move my head); Q8: I know where the objects are because I can touch them. (Disagree/Agree).

One-way ANOVA tests, without repeated measures, reveal the following statistically significant scores marked with an *:

Q1: $F_{3,60} = 0.7$, $p = 0.541$; Q2: $F_{3,60} = 0.044$, $p = 0.987$;
Q3: $F_{3,59} = 5.7$, $p = 0.0016^*$; Q4: $F_{3,60} = 4.04$, $p = 0.010^*$;
Q5: $F_{3,59} = 4.23$, $p = 0.0088^*$; Q6: $F_{3,60} = 2.49$, $p = 0.068$;
Q7: $F_{3,38} = 6.99$, $p = 0.0007^*$; Q8: $F_{3,58} = 0.64$, $p = 0.59$;

Given that all of the 21 participants performing Task 2 filled in the questionnaires and just a part of them performed Task 3, we opted to detail Task 2. Figure 3.11 (red and blue bars) summarizes the obtained scores for each question while performing Task 2 with both interfaces. These results were validated through one-way repeated measures ANOVA analysis, and the statistically significant scores are marked with an * :

Q1: $F_{1,20} = 2.016$, $p = 0.171$; Q2: $F_{1,20} = 0.022$, $p = 0.883$;
Q3: $F_{1,20} = 11.294$, $p = 0.003^*$; Q4: $F_{1,20} = 6.981$, $p = 0.015^*$;
Q5: $F_{1,20} = 26.544$, $p < 0.001^*$; Q6: $F_{1,20} = 2.016$, $p = 0.171$;
Q7: $F_{1,25} = 8.432$, $p = 0.007^*$; Q8: $F_{1,19} = 0.516$, $p = 0.481$;

3.6.2 Discussion

The experiments aim to understand the influence of an immersive visualization interface in relation to spatial and movement perception. To study these factors, we designed several tasks where users had to indicate a 3D position in space based on visual feedback mediated by technological means: a single monitor with a wide single view of the remote scene (traditional interface) versus multiple partial views of the scene naturally acquired while moving their head.

The experiment described in Task 1 demonstrates that, if the workspace of interest is all within the field of view of the HMD, there is no benefit using an immersive interface. The participants did not feel the need to move their head to accomplish the task of pressing key blocks, not gaining depth perception. The time performance was similar for both visualization interfaces.

Task 2 experiments show that an immersive interface outperforms the traditional interface when the workspace of interest is larger than the HMD field of view. Comparing a fixed wide view of the scene with a set of partial views of the scene provided by the head user movement, the last approach improves depth perception due to motion parallax. There is an enhancement of spatial and movement perception demonstrated by the accuracy and speed to accomplish the pick-and-place task. A consequence of the limited FOV of the remote camera (immersive interface condition) was that users moved their head more frequently. Nevertheless, this active spatial perception allowed users to focus their attention on one region of interest, minimizing the workload. All participants took advantage of the touch sense to localize objects and to perceive the height of the gripper tip. Users' common procedures consisted of moving the gripper in contact with the table until reaching the object. This way, the user compensates the lack of height perception through vision while moving the gripper or the tip of the stick.

In Task 3, we tried to evaluate the accuracy of the movement without the help of haptic feedback. Here, the time performances to follow a 3D path did not present a significant difference while comparing both interfaces; however, there was a significant improvement concerning the accuracy of the 3D movement. This is shown by the lower number of hits occurring while using an immersive interface.

Concerning the latency, there was no communication delay due to the proximity of the devices. However, special care was dedicated to tuning the two-step motor lag responsible for the pan and tilt movements of the camera, because faster movements of the head were tracked by the HMD, but were not properly executed by the motors (faster DC motors are advisable). The cameras of the system enabled the video frame rate ($>=$30 fps), thus fulfilling the requirements for handling the tasks.

The experiment referred to as the "control task" was initially executed to evaluate the weight of each mediation component in the immersive system. The no mediation task "direct vision (binocular stereo) + hand" became our ground truth. Users handled the block with their hands, looking directly with their eyes, and took $\mu_t = 8.75$ s ($\sigma_t = 1.39$) to accomplish the task. With the introduction of a tool ("direct vision (binocular stereo) + gripper"), users took $\mu_t = 30.87$ s ($\sigma_t = 14.9$). The gripper disturbance caused an increase of 252$\%$ in time performance. Movement perception of the gripper in relation to the objects became an issue. Using see-through HMD only ("HMD indirect vision (mono biocular) + hand"), users took $\mu_i = 15.62$ s ($\sigma_i = 2.92$). The see-through HMD disturbance increased 78$\%$ in time performance. The limited FoV contributed to such disturbance, and participants quite frequently used their other hand to self-position their body in relation to the workspace. With the introduction of the gripper and see-through HMD ("HMD indirect vision (mono biocular) + gripper"), users took $\mu_i = 41.3$ s ($\sigma_i = 8.5$). The combined disturbance increased 372$\%$ in time performance.

Questionnaires showed that:

(Q1) users found it quite "easy to use both visualization interfaces", traditional and immersive ($\mu_t = 2.9$, $\sigma_t = 1.54$ vs. $\mu_i = 2.42$, $\sigma_i = 1.53$) (the subscript t stands for traditional and i for immersive).

(Q2) users did not consider the pick, insert, and place task (Task 2) as a "tiring" one using either interface ($\mu_t = 2.66$, $\sigma_t = 1.59$ vs. $\mu_i = 2.61$, $\sigma_i = 1.46$). Q2 did not present significant differences because the question addressed a common task.

(Q3) users felt that their "movement action" was more precise using the immersive interface (HMD providing an active point of view) ($\mu_t = 4.33$, $\sigma_t = 1.19$ vs. $\mu_i = 5.47$, $\sigma_i = 1.07$). This significant difference results from the gain in depth perception, a consequence of motion parallax. Better visual depth feedback helps to calibrate the arm proprioception system, enabling finer movements of the tool. By moving and seeing our arm, or seeing the tool as an extension of our limb in an unknown 3D space, it helps us to perceive the spatial dimension and localize objects.

(Q4) Inquiring about which interface enabled a better visualization to manipulate objects, it was clearly stated that it was the immersive interface ($\mu_t = 4.38$, $\sigma_t = 1.65$ vs. $\mu_i = 5.52$, $\sigma_i = 0.74$). Subjective scores were statistically significant, and quantitative time mean performances measurements confirmed these results. Task 2 itself is an easy one, and in terms of usability, the preference was for the immersive interface.

(Q5) Inquiring about if users were aware that "visualization" was being supported through an "indirect technological device", they answered that they forgot more easily when they were using the immersive interface ($\mu_t = 3.42$, $\sigma_t = 1.77$ vs. $\mu_i = 5.28$, $\sigma_i = 1.58$). It is understandable that users answered this way, as the immersive interface provides a more natural point of view. People tend to forget that their own eyes are not really in the remote space.

This is a confirmation of the importance of the view point transfer proposal as a key element for achieving telepresence.

(Q6) Users also thought that with the immersive interface, they had "a clear perception and total control of stick/tool movements" ($\mu_t = 4.80$, $\sigma_t = 1.47$ vs. $\mu_i = 5.28$, $\sigma_i = 1.18$). This result might require more samples to become statistically validated; notice, however, that some of spatial perception arises from monocular cues already available in the traditional interface ($Fix_{Display} + Fix_{Cam}$). Pictorial cues in a single image and movements in the scene can provide depth information. This type of cue provides information about occlusion, relative height, relative size, perspective convergence, texture gradient, and shadows, enough to perceive the space.

(Q7) The vast majority of users were unanimous in stating that they "they perform better when they move their" heads ($\mu_t = 5.14$, $\sigma_t = 2.26$ vs. $\mu_i = 2.40$, $\sigma_i = 2.11$). Although this question does not make sense with relation to the traditional interface, because a fixed monitor ($Fix_{Display} + Fix_{Cam}$) does not provide a dynamic point of view, it still provides monocular cues related to depth. On the other end, by moving the user's head, the immersive interface adds to the mentioned monocular cues and the motion-produced cues, like motion parallax, accretion, and deletion. Besides the benefit of depth information from motion parallax, users can play with occlusion to match their mental models.

(Q8) Inquiring about the importance of the sense of touch during an interaction, users confirmed the importance of haptic feedback. The requirement of this type of feedback was evident for both interfaces, especially when depth information was unavailable (similar scores, $\mu_t = 5.85$, $\sigma_t = 1.34$ vs. $\mu_i = 5.7$, $\sigma_i = 1.38$). Visual localization of the objects was frequently confirmed through the sense of touch while reaching it. Visual images of the tool in movement enable knowing its position; however, the knowledge of its height can be difficult. To overcome this limitation, some users moved the tip of the tool in contact with the table, refining in this way the reference plane with haptic information provided by arm proprioceptive sensors.

Presence Question Q5-Q8 show that the immersive interface easily becomes transparent for the user, letting him/her feel that he/she is naturally perceiving the remote environment. Visual feedback is transparently mediated and, combined with motor-actions, contributes to the sense of telepresence.

The findings also show that the performances are better when the task workspace is in front of the user, in opposition to setting it on the right side. A visuomotor task where the natural frame of reference of vision is shifted with relation to the common frame of reference for hand/arm movement increases mental workload and consequently decreases task performance. Operators are required to do mental transformation to compensate for their manipulation inputs with relation to the corresponding tool action observed in the displays. The gesture coordination ends up relying more on visual feedback. For example, in our experiments, the pick-and-place task using a hand gripper (Task 2) with the traditional interface had the workspace on the right side; whereas, in the control experiment, the

same task was performed using the traditional setup “monitor indirect vision (mono biocular) + gripper" (F5) with the workspace in front of the user. Comparing the mean task-time performance for the traditional interface in both experiments, we found that participants performed faster with the workspace in front than with it on their right side ($\mu = 56.4$, $\sigma = 16.47$ vs. $\mu_t = 136.75$, $\sigma_t = 61.93$). It is a significant difference with one-way ANOVA $F_{1,23} = 8.03$ and $p = 0.009^*$. Thus, users are used to the consistency between visual, proprioception (sense of body position), kinesthesia (sense of body movement), and vestibular feedbacks, and any inconsistency implies new skills.

The research on perceptual factors affecting user's control behavior is useful to improve direct control teleoperation interfaces, minimize control workload, improve task performance, and maximize safety. Additionally, it can contribute to designing semi-autonomous functionalities based on cognitive human-robot interaction architectures to assist operators. For example, during direct interaction such as telesurgery, semi-autonomous systems can prevent dangerous movements of the robotic arm, adapt the responsiveness of the system to the variability of perceptual factors, or adapt the interface to different users. Moreover, the interfaces or robot's behaviors can be adapted to the context.

3.7 Conclusions

This research shows that people can experience and perform actions in remote places, through a robotic agent having the illusion of being physically there. The sensation can be compelled through immersive interfaces; however, technological contingencies can affect human perception. Considering the results from studies on human factors, we provide a set of recommendations for the design of immersive teleoperation systems aiming to improve the sense of telepresence for typical tasks (ex. Table 3.5). The mitigation of issues like system latency, field of view, frame of reference, or frame rate contribute to enhancing the sense of telepresence. The presented example of the evaluation methodology allows analyzing how perceptual issues affect task performance. By decoupling the flows of an immersive teleoperation system, we start to understood how vision and interaction fidelity affects spatial cognition.

Task experiments with participants using traditional vs. immersive interfaces allowed quantifying the disturbance introduced by each component on the system. For example, taking as a reference a simple manual pick-and-place task, the introduction of a visual see-through HMD increased the time to perform it by 78%; the introduction of a manual gripper tool increased that time by 252%; and the combination of visual and tool mediation increased the overall time by 372%. Decoupling the flows of an immersive teleoperation system allowed a separate analysis of visual feedback disturbances (e.g., limited FOV) without the influence of other factors that affect the frame of reference for motor-action. Our findings show that misalignment between the frame of reference for vision and motor-action or the use of tools that affect the sense of body position or sense of body movement leads to higher mental workload and has a higher effect on spatial

cognition. Misalignment between kinesthetic and visual feedback increases the mental workload and compromises the sense of telepresence and the embodiment feeling. The mental workload to control the suggested video feedback component is considerably lower (in the immersive interface); however, the combination of both requires a higher effort (i.e., active visual mediation plus tools). Thus, a recommendation is to keep activities at skill-based behavior levels, where familiar perceptual signals are essential to lower the cognitive effort. Future work includes the evaluation of the traditional interface setup, considering the control of the remote camera orientation with a joystick, and the evaluation of the proposed immersive interface to control a robotic arm with haptic feedback.

Table 3.5: Human capabilities vs. human capabilities through mediated technologies.

	Task	Analyzed Criteria	Ref	Resolution	Frame Rate (FR)	Latency	Field of View (FoV)	Frame of Reference (Camera Perspective)	Depth Cue	Display Type	Results
Human capability	Multi-purpose	-	Table 3.1	>60–200 pixels/ degree	>1800 Hz	<7–15 ms	210° (H)×135° (V)	Egocentric	Pictorial, motion parallax, binocular cues	-	
Teleoperation	Placement	Accuracy, speed, and performance	[99, 101]	-	>15Hz	-	-	-	-	-	-
	Placement and grasping	Accuracy	[512, 513]	-	> 25 Hz	-	-	-	-	-	-
	Tracking	Accuracy, perceived control, and stability	[139]	-	> 12 Hz	-	-	-	-	-	-
	3D Tracking	Accuracy and speed	[291]	-	> 33 Hz	-	-	-	-	-	-
Telemanipulation	Telesurgery: cutting, stitching, knotting	Accuracy, precision, and performance	[305]			<300 ms	-	-	-	-	-
Telemanipulation	Laparoscopy surgery	Usability and performance of experienced surgeons	[264]			<105 ms	-	-	-	-	-
Telepresence	Telepresence robot	Performance, usability, workload	[496]			<125 ms	>170° (H), wide or with pan/tilt	Egocentric	Pictorial, monocular, parallax motion cues	1 × monitor or HMD	Navigation and social interaction
Driving	A 6 wheel all terrain rover of 6.800 kg	Avg. speed and avg. time stopped	[423]	40 pixels/degree, 5 × 1600 × 1200	>25 Hz	<480 ms	200° (H) × 30° (V)	Egocentric	Pictorial, monocular, and motion parallax cues	5 × high-res LCD monitor, side-by-side, true size	Operator's situational awareness and perception of the vehicle's position and motion
Driving	A car driving on city roads at 30 km	Tracking line, obstacle detection, performance	[179, 216]	5 × 640 × 480	>25 Hz	<550–600 ms	240° (H)	Egocentric	Pictorial, monocular, and motion parallax cues	3 × high-res LCD monitor side-by-side, true size, and HMD	–

Chapter 4

Presence and Telepresence in HRI/HMI

The chapter showcases a series of experiments that prove how telepresence, embodiment, natural interaction, and immersive interfaces can improve robot teleoperation, making it easier to perform tasks while minimizing cognitive workload. It comprises extended versions of papers published in IEEE Human–Robot Interaction (HRI) and Human–Machine Interface (HMI) conferences [27][23][18], and a journal paper published in Computer Graphics and Applications [174].

4.1 Be the robot: Human Embodiment in Tele-Operation Driving Tasks

This experiment proposes new and natural interacting strategies to support operators challenging task of driving a robot and simultaneously control the perception camera. This new concept consists in mixing telerobotics, telepresence and physical embodiment, resulting in virtually transferring the operator to the remote robot. Four interaction styles were experimentally compared and results show that gestures and contactless body intention methods improved the user tele-operation performance task. The human visual perception of the remote environment is enhanced by having the remote camera viewpoint controlled by head movements (see figure 4.1). Better metric perception results from viewing the robot body and spatial perception can be augmented by fusing remote and local visual cues. Moreover, this study suggests that, when a person is focused on the task, achieving the ownership illusion towards remote body, there are autonomic responses that correspond to what would be expected in events that take place in reality (collisions avoidance postures).

Figure 4.1: Navigation task performance comparison while using different teleoperation interaction styles designed to enhance embodiments sensations

4.1.1 Introduction

The evolutions of communications and robotics have created the perfect opportunities for new tele-operated robots. These are expected to play an important role in maintenance or exploration tasks to be performed in remote or hazardous environments.

In these teleoperation tasks, the operator usually controls the movements of the remotely located robot by observing the evolution of its position on a map-like representation, or by using live video streams to perceive the remote environments. In the latter case, the robot has a camera that becomes the "eyes of the operator". It is common to have joystick-like interfaces to control both the robot motions and the camera orientation. Apart from being unnatural, this type of interaction styles can overload operator's handling capability and require extra attention to multiple controls, decreasing the main task performance [99][539][26][392][232].

To address such challenges we proposes a new strategy for the multitasking problem of having a human simultaneously driving a robot and controlling a remote camera to perceive the remote environment. To maximize the task performance and minimize the operator's physical and cognitive workload, we propose to exploit the operator induced sensation of being at the remote environment [459][463]. Moreover, the generation of remote physical embodiment feeling is explored where the user can perceive the robot's structure as his/her own body [528].

This new principle consists in mixing teleoperation, telepresence and physical embodiment, resulting in virtually transferring the operator to the remote robot. A physical embodiment already envisioned in AVATAR film. The perception of ourselves can be fooled if synchronous actions and sensations are perceived, like in the case of "the rubber hand illusion" [111]. Our goal is close to this, in the sense that we want the operator to perceive the remote environment as if he/she is really there, and control the robot as if it is his/her own body (figure 4.1).

Mel Slater [463] found that when embodied in a virtual body, the perspective

position and point of view provided to the user minimizes the importance of visual-tactile synchronization. Such important finding leads us to implement a viewpoint transfer mechanism using an HMD, to remotely replicate the human visual perception strategies. We argue that viewing the remote environment as one being controlling the robot from inside of it, reduces the required mental workload to compute, the otherwise needed, view and control transformations.

Based on the findings of our previous research [27], the present robot architecture addresses the operator metric perception issue, by broaden the field of view (FOV), and letting the operator to view the robot body for reference purposes. A new visualization strategy is also suggested to enrich the user spatial perception. It consists in fusing remote and local visual cues in one synchronous video feed. By enabling the user to see his own feet, while looking down through the HMD, we expect to enhance the sense of tele-presence.

In the remaining of this section we present the proposed interaction mechanism, the developed interaction support software architecture and the experimental tasks for evaluating our proposed approach and compare it to traditional remote operation controls. Finally the results of the experiments are presented and discussed.

4.1.2 The human in the teleoperation loop

To better understand the possible influence of the various interaction mechanisms proposed on this paper, we start by proposing a simplified model of the teleoperation problem. Several authors have studied the teleoperation problem. Some propose models where the human operator is part of a control loop, sending commands through a delayed transmission channel [232], and receiving the manifestations of the robot motion through the same delayed channel. The relevance of such models comes from the fact that the transmission delays have important effects on the controllability of the overall system.

In our study we are concerned with the relationship between the human ability to control the remote robot and the interaction mechanisms in use. Concerning this, figure 4.2 presents a diagram that models the tele-operation process. This model is composed of an outer tele-operation control loop, that utilises an inner perception control loop.

Tele-operation Loop - As already stated, tele-operating a robot can be modelled as a control loop. In this model the human compares a given goal with the position of the robot in the remote environment. From the perceived difference, he/she develops an intention that is then translated to robot commands through some interaction mechanism. This is represented in figure 4.2, where block A represents the perception of the error and production of an intention. This intention (X_i) is transformed into a command through block B which represents the human action over some interface that produces the adequate command (X_m). After a certain time delay, the command (X_m), now referred as command (X_d), is executed by the robot at the remote environment. This is possible if the loop can be closed by having the user perceiving the pose of the robot on the remote environment. The

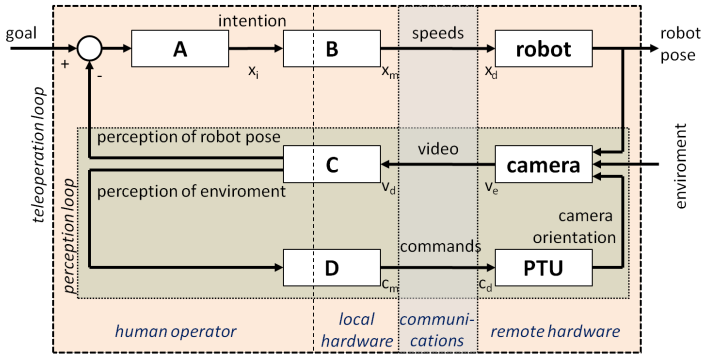


Figure 4.2: Teleoperation and perception as control loops

new robot pose is reflected to the operator through environment video images (V_e), affected with time delay (V_d) due communication latency.

Perception Loop - In this experiemnts we consider teleoperation systems which the point of view of the operator is from inside the robot, i.e. using an embedded camera. Ideally the user should be able to perceive both the environment and the robot motions as if he was driving the robot from the inside. Nevertheless, due to technical limitations this is not the common case.

This perception process can also be described as a control loop. Here the human controls the robot camera orientation and uses the visual feedback to compensate his scanning actions required to pursue a goal. Controlling the camera, the user can scan the environment, track objects, etc. Similarly to the previous case, the camera captures images (V_e) that are viewed by the user after some time (V_d). These results in the perception of the remote environment and the relative pose of the robot in it, and it is modelled by block C. Tracking, searching of objects, or snooping can be made by controlling the camera orientation. This can be done through the actuation of the user onto some interaction mechanisms represented by block D, whose commands (C_m) after certain time delay (C_d), are sent to the Pan-and-Tilt unit (PTU).

4.1.3 From teleoperation to remote embodied operation

Considering the presented model, we are now going to present the correspondences between different perception and control interaction mechanisms and the referred blocks A,B,C, and D (see figure 4.2).

The traditional teleoperation approach - In the traditional teleoperation systems, both blocks B and D represent joystick-based operation for respectively the robot motions and the PTU. In this case block C corresponds to viewing the remote environment images in a screen, and block A to the transformation of the perceived error into the intention to move the robot .

Viewpoint transfer using an HMD - Aiming at solving the problem of controlling the remote viewing camera, our approach uses the user's head movements, tracked through a mounted display (HMD), to control a remote camera on the robot. It controls a PTU that holds the camera (block C and D). Thus, the camera is

controlled in an egocentric way, as the user has the view centered on the camera and can control it by rotating his head in the desired direction.

This aims at being the first step for giving the user the sensation of being physically embodied on the remote robot, as "the user will see what the robot can see". Here, block C corresponds to viewing the remote environment images through the HMD which enables a wider FOV. The operator scanning strategy is transformed into commands to the PTU by block D either to perceive the environment and/or to refine the perception of the robot pose. To accomplish this we use the HMD inertial measurement unit (IMU), which is composed of a three axial accelerometer, gyros and magnetometer. From this unit we estimate the user head pose, which are then transmitted to control the remote PTU that supports the camera. Using this system, whenever the user looks up-or-down, left-or-right, the remote PTU will replicate these movements giving the remote camera a direction equivalent to the user's gaze. By consequence, the user can "look around", "track a moving object", or just look down to see his own belly and feet, now replaced by the robot structure. This enables the user to have an egocentric perception of the remote surrounding environment, as if he/she was at the robot position and orientation. This will enable him/her to naturally explore the environment and navigate as if he/she was really there.

Deictic gesture-based control of the robot

Looking for natural strategies to tele-operate a mobile robot, other than the classic joystick, the present work was focused on the development of a gesture-based framework (block B). Aiming at creating a natural interaction mechanism, the initial idea was to use deictic gestures to control the robot movements.

Instead of using the motion transfer principle that is exploited by the X-BOX® games, where the movements of the user are transferred to some virtual character that appears on the screen, the idea here is give the user that he/she is controlling his/her own motion, as it he/she was at the remote location. One again, the idea is that of exploiting the physical embodiment sensation by making the user to perceive the robot structure as his own body. If accomplished the user will see him as being the robot, or inside the robot, and his pointing gestures will be used to control "his own" motions on the remote environment.

On a typical mobile robot these gestures should be mapped to control the linear and angular velocities of the robot. Here we propose to recognize gestures that mean forward, back, turn right, turn left, which are similar to the ones used by a person to help another to navigate or maneuver a car.

The operator expresses his intention to move (block A) through deictic gestures which are interpreted by block B into speed commands. The perception of robot pose loop feedback that is block C and robot camera enables the operator to tune his gestures range.

For this we used a Kinect® sensor and the OpenNI™ [382] library to track the body posture. The produced output is an estimate of the coordinates of 17 points (joints)

that define the 3D configuration of a skeleton-like model. On figure 4.3 on the left there is a representation of this model, and on right there is a frame of a tracking sequence where the skeleton is superimposed on the user silhouette.

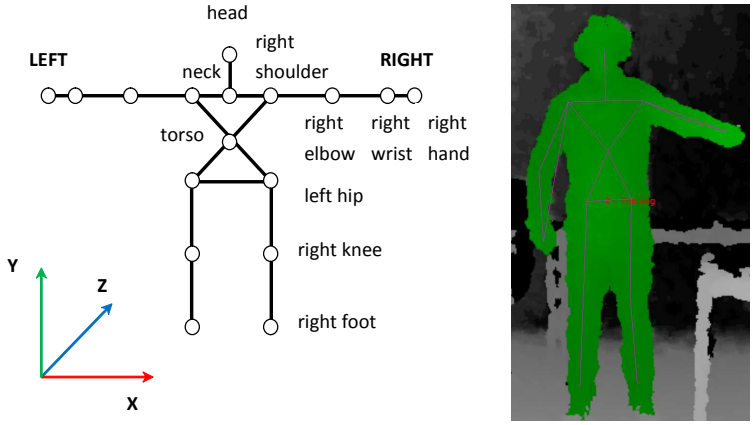


Figure 4.3: Skeleton Model Joints

The deictic gestures can then be extracted from selected joints of this model. These gestures were defined from the basic set of pointing forward, backward, left or right and can be directly mapped into going forward, going backward, rotating left and rotating right. We can define that gesture along the sagital plane (see figure 4.4) of the operator (or very close to it) correspond to going straight forward or backward, and that gestures along the operator's coronal plane (or very close to it) will represent pure rotation, whose rotating direction depends to which side of the sagital plane it is done. Intermediate positions of the hand will result in combinations of forward/backward movements and rotations.

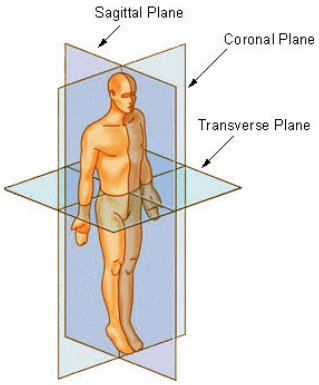


Figure 4.4: Body planes representation [515]

The skeleton is composed by 17 joints, namely head, neck, torso, right hand, left hand, right wrist, left wrist, right elbow, left elbow, right shoulder, left shoulder, right hip, left hip, right knee, left knee , right foot, left foot. Figure 4.3 describes these joints, as the sensor referential.

Now we are able to define robot commands based on gestures, by extracting parameters from some joint combinations. The angular velocity Ω, can be related with the angle between the arm and the sagital plane, meaning that if the arm

points to the front we get 0 degrees (no angular velocity), and if the arm points to the right we get 90 degrees (maximum positive angular velocity). Similarly, we relate the linear velocity V (forward/backward velocity) as the angle between the arm and the coronal plane. Therefore the maximum linear velocity is obtained when the arm points to the front and zero linear velocity when the arm is down (hand near the hip). To simplify the conversion of those gesture into linear and angular velocities to be applied to the robot, we consider a referential located at the intersection of the body planes referred on figure 4.4, where the X axis is at the intersection between the coronal and transverse planes, the Y axis the intersection between the sagital and transverse planes, and Z axis the intersection between the sagital and coronal planes. Now the linear and angular velocities are computed from the projection of the hand into the Z axis and X axis respectively, and then normalized by the arm length. There resultant values are the fraction of the maximum defined velocities (eq. (4.1) and (4.2)) .

$$V = \left(\frac{rightShoulder.Z - rightHand.Z}{\|rightShoulder - rightHand\|} \right) V_{max} \tag{4.1}$$

$$\Omega = \left(\frac{rightShoulder.X - rightHand.X}{\|rightShoulder - rightHand\|} \right) \Omega_{max} \tag{4.2}$$

Body intention-based control of the robot

We have defined another way of controlling the robot movements based in the body expressed intentions (block B). We define body intention as changes in the body posture, with respect to the rest position, that may be used to express intention to move. For example to walk forward or backward we displace one foot forward or backward respectively. To rotate left or right we rotate the shoulder in the corresponding direction (eq. (4.4)). The operator expresses his intention to move (block A) through body postures which are interpreted by block B into speed commands. The perception of robot pose loop feedback, that is, block C and robot camera, once again, is used by the operator to tune his body postures range.

To simplify this method of interacting we have marked a square on the floor with marks inside for both feet rest positions. Now it becomes simple to compute the linear velocity (eq. (4.3)) from the projection onto the Y axis of the displacement of the moving foot with respect to the rest position [27].

$$V = \left(\frac{movingFoot.Z - restingFoot.Z}{\|movingFoot - restingFoot\|} \right) V_{max} \tag{4.3}$$

$$\Omega = \max \left(1, \left(\frac{leftShoulder.Z - neck.Z}{leftShoulder.X - neck.X} \right) \right) \Omega_{max} \tag{4.4}$$

Mixed reality video streaming

While large field of view (FOV) has an important role in tele-operation or navigation tasks, namely user metric perception, it does not solve the scale problem that can exist between operator local site and the remote environment (the operator can be controlling a micro robot in a micro environment without realizing that). For example, a car driver usually perceives the outside environment through the windshield, although when he looks inside, he sees simultaneously the steering wheel, the dashboard, his legs and the outside environment. Such references cue allows the human driver to establish scale references between inside and outside worlds.

A new visualization strategy is suggested to enrich the user spatial perception. It consists in fusing the remote video streaming with a local video streaming resultant from a camera mounted on the HMD. Both remote and local camera points the same direction, although the mixing process only happens when the user looks down. The implemented solution (see algorithm 1) is based on a croma key segmentation and composition. The operator is asked to stand on a small green carpet, only visible when the camera is pointed down. Feet and legs are segmented and superimposed on to the robot video stream (see figure 4.5). The functionality was programmed using OpenCV library to achieve faster frame rates. By enabling the user to see his own feet, while looking down through the HMD, we expect to enhance the sense of tele-presence.

Algorithm 1 Remote and local video stream fusion when the operator looks down (to green carpet)

1: **Input:** remote video stream (robot camera)
2: **Input:** local video stream (HMD camera)
3: **Output:** display video stream (HMD LCD)
4: **for all** frames $\mathbf{rvid}$ remote video stream & frames $\mathbf{lvid}$ local video stream **do**
5: $\mathbf{hsv}$ = convert $\mathbf{lvid}$ to HSV color model
6: $\mathbf{mask}$ = filter $\mathbf{hsv}$ green pixels
7: $\mathbf{maskc}$ = find biggest contour of $\mathbf{mask}$
8: $\mathbf{maskh}$ = find convex hull of $\mathbf{maskc}$
9: $\mathbf{maski} = \mathbf{maskh} \cap \mathbf{maskc}$
10: $\mathbf{dvid} = \mathbf{lvid} \cap \mathbf{maski}$
11: $\mathbf{maski}$ = invert $\mathbf{maski}$
12: $\mathbf{dvid} = \mathbf{rvid} \cap \mathbf{maski}$
13: insert $\mathbf{dvid}$ in display video stream
14: **end for**

4.1.4 Experiments and Results

Feeling embodied in a tele-operated agent means acting as naturally as no mediation exists. To understand the influence of different interaction style on the tele-operation of a mobile robot and to asses how natural can a user interact and

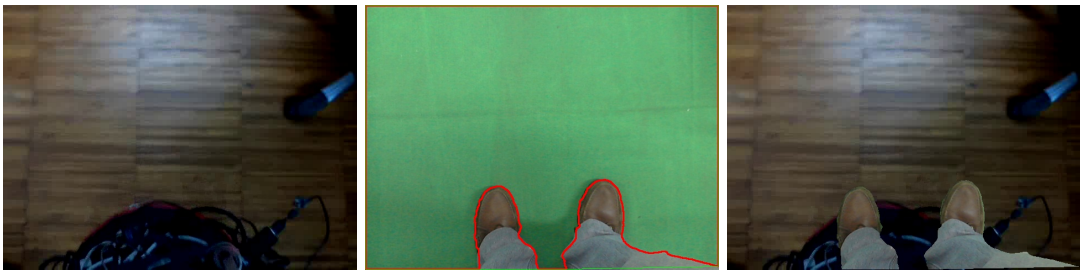


Figure 4.5: Enhance the sense of tele-presence through video stream fusion when the operator looks down: remote robot video stream, local video stream and fused display video stream

Figure 4.6: Hilario Robot (left) and Remote Control Station (right): (1) Scout, (2) Laptop, (3) PTU Camera, (4) Joystick, (5) RGB-D Camera, (6) Head-Mounted Display

perceive the remote robot structure as his own body, a quantitative and subjective task performance analysis were carry out involving several participant in a driving task. The mediation interfaces sets described as first, second, third and four experimental setups, enabled comparison between standard and natural view point scan strategies, joystick and more natural robot control forms, like deictic gestures and body intention.

Implementation

A setup composed of a mobile robot and a control station was built, both represented on figure 4.6.

The robot platform is a Scout II, onto which a structure was built to support a Directed Perception® PTU with a camera at about 1.60m height. Both the robot and the PTU are controlled from a module that receives the commands from the remote control station. The camera view is transmitted to the remote station via Skype®, what enables a good video quality with minimal bandwidth loss.

In what concerns the remote control station, its configuration varies according to the experiment. As will be described later, depending on the case, different combinations of joystick controller, screen monitor, RGB-D sensor and head-mounted display with IMU, will be used.

Architecture:

The architecture shown on figure 4.7 was designed to enable the testing of different configurations by creating different mixes of the software modules developed for both the robot and the control station. In fact, as different interaction strategies were to be tested and evaluated, the active modules that compose the setups vary from experiment to experiment, as described hereafter. To simplify the development the communication between modules at both sites is done via wireless TCP/IP based connections.

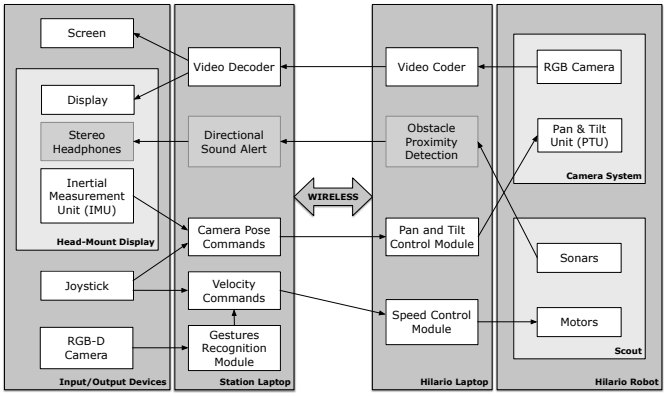


Figure 4.7: Architecture Diagram

Experimental Design

The objective of the experiments is to understand the influence of different interaction styles on the teleoperation of a mobile robot. To assess how natural can a user interact and perceive the remote robot structure as his own body, four experiments were designed to test and compare the embodiment effect on task performance. The designed task consisted of teleoperating a robot through a path (figure 4.8), avoiding obstacles and finishing in the shortest period of time. The participants repeated this task using all the four experimental setups enabling comparison

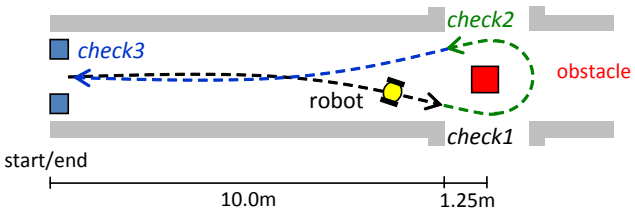


Figure 4.8: Experimental task path, divided in 3 section (check 1, check 2 and check 3) with one obstacle (red box).

between standard and natural view point scan strategies, visual feedback, joystick, deictic gestures and body intention based robot control.

The four interaction setups developed are:

- **Setup 1:** Both the robot motions and camera orientation are controlled using two joysticks.
- **Setup 2:** The robot motion is controlled using a joystick, and the camera orientation is directly controlled using with the orientation of the HMD worn by the user.
- **Setup 3:** The robot motion is controlled using deictic gestures and he camera orientation is directly controlled using with the orientation of the HMD worn by the user.
- **Setup 4:** The robot motion is controlled using changes in the body posture and the camera orientation is directly controlled using with the orientation of the HMD worn by the user

In all cases the participants were standing controlling the robot and the camera as described. For the first case, the participants could view the remote camera video steam through a screen located just in front on the user. For the remaining cases the video was displayed on the HMD.

Procedure

Thirteen persons (2 female, 11 male), whose ages have a mean value of 26.46 and standard deviation of 5.03, participated in the experiment. These participants were informed about the purpose of this experiment, the procedures involved, that their cooperation was voluntary and that the personal data was going to be kept confidential. They had no prior knowledge about the experiment or the involved technologies. They were instructed that the main goal of this experiment was tele-operate the robot through a path, avoiding obstacles and finish in the shortest period of time. Then they tried to execute it using each of the above explained setups. The execution times were measured, and at the end they were asked to answer a short questionary.

Results

As stated above, in these experiments we tried to evaluate the performance changes and the subjective satisfaction of the users. The performance of the user is measured in terms of time spent for executing the task, for each of proposed interfaces. Instead of using a single time measure, the trajectory was divided in three sections (check 1, check 2, check 3). This enabled us to detect variations in performance between each of them, and to separate the initial adaptation and learning of the interaction mechanism from the rest of the experiment.

Figure 4.9 depicts the performance mean times (sec) and standard deviation for the three trajectory segments of the four setups.

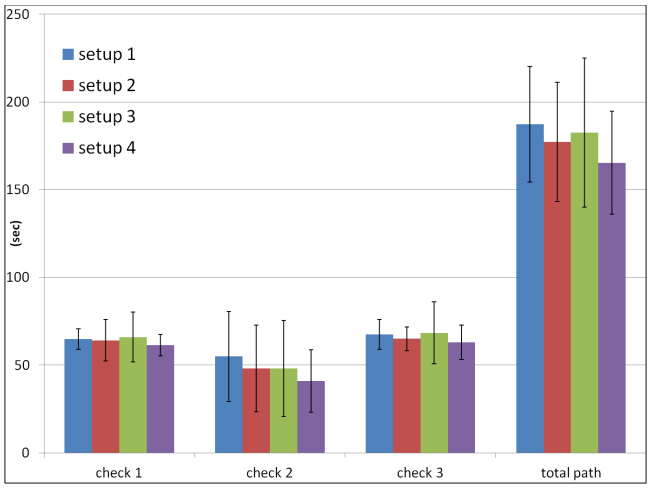


Figure 4.9: Users mean time spent to perform the task on each experiment (seconds). A measurable comparison effect on task performance while using different teleoperation interaction styles designed to enhance embodiments sensations

A subjective evaluation was developed, based on questionnaires that the users had to fill at the end of the experiment. The participant feedback questionnaires were based in 7 point Likert scale and comments. The questionnaire contents follows the IBM Computer Usability Satisfaction Questionnaire [283] and NASA-task load index (NASA-TLX) [197], in which users compared their experiences with different visual feedback models and robot control interfaces concerning subjective measures like consciousness, easiness, embodiment feeling, usability and perceived used time. The questions to be scored from 1 to 7 were:

- Q1: It was intuitive to use.
- Q2: It was easy to learn.
- Q3: I could get enough visual feedback.
- Q4: I felt embodied on the robot while performing the task.
- Q5: I completed the task quickly.

Figure 4.10 summarizes the obtained scores.

Discussion

Objective measure results demonstrate that when the operators controls the camera point of view with his head pose (freeing their hands and minimizing the cognitive workload), obtaining a more natural visual feedback through an HMD, they improve significantly their task performance. The introduction of this visual strategy was done in setup 2 and maintained through setup 3 and setup 4. In

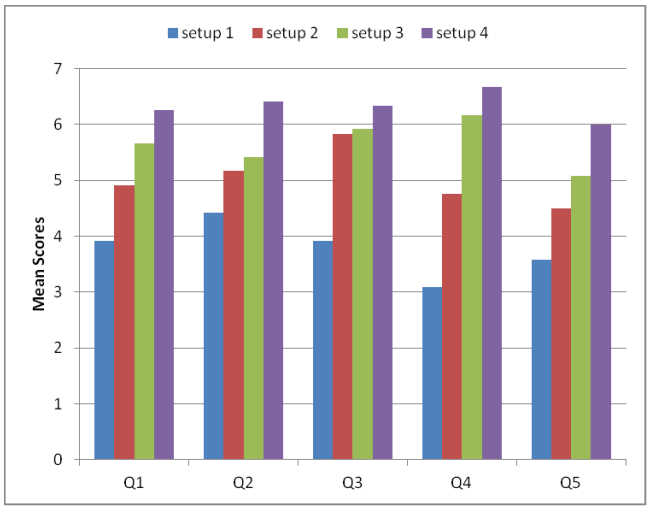


Figure 4.10: Mean scores from user questionnaire feedback, scale : 1- Strong Disagree to 7-Strong Agree

Setup 1, while manipulating camera and robot with joysticks, operators took mean times of 65s (SD=5.9), 55s (SD=25.6), 67s (SD=8.5) to perform path sections 1, 2, 3 respectively.

The visual feedback control strategy, introduced in setup 2, enabled to outperform the setup 1 time values and obtain means of 64s (SD=11.8), 48s (SD=24.7), 65s (SD=6.7) to perform path sections 1, 2, 3 respectively). The evidence is more noticeable in path section 2 (check2), where the operator had to scan the environment and plan a trajectory to turn around the obstacle. Due some visual feedback delays and limited field of view, part of path section 2 trajectory had to be mentally anticipated by the operators. Small collisions with the obstacle were registered, and the number of operators that collided with the obstacle were less in setup 2 than in setup 1 (4 in opposition to 6).

When questioned about gains in visual feedback (Q3), operators have scored high, setup 2, 3, and 4 (mean scores 5.83, 5.92 and 6.33 respectively) in opposition to setup 1 (mean score 3,92)

The introduction of natural deictic gestures based robot control (setup 3) presented gains in task performance when compared with setup 1, specially where higher skills were required (check 2 section)(setup 1 check 2: 55s (SD=25.6) and setup 3 check 2: 48s (SD=27.4)). Operators spent less time in setup 4 (setup 4 check 2: 41s (SD=17.7)). Setup 2 and setup 3 had equivalent times (setup 2 check 2 : 48s (SD=24.7) and setup 3 check 2: 48s (SD=27.3)).

Body intention-based robot control (setup 4) was the operators choice in all questions, confirmed by the time performance measures (means of 61s (SD=6.1), 41s (SD=17.1) and 63s (SD=9.7) to perform path sections 1, 2, 3 respectively). Notice that the standard deviation of setup 4, in section 2 check were smaller than in other equivalent section setup, meaning that operators were more regular performing this section.

All the operators felt comfortable operating the robots with their shoulders. In setup

4, practically there weren't operators colliding with the obstacle in check 2 section, as they were more familiar with path and had also a precise orientation control.

When questioned how easy is to use the control interfaces, operators scored high setup 3 and setup 4 (mean scores 5.6, 6.2 respectively), that is, the natural interaction styles. They felt easy to adapt to the interfaces of setup 4 (scoring 6.4). Once the controller of the visual feedback was identical in setup 2, 3, and 4 operators have scored with similar maximum values of 5.8, 5.9, and 6.3 respectively. The embodiment feeling high scores were associated to natural interaction styles implemented on setup 3 and setup 4. Without knowing exactly the chronometer time when filling the questionnaires, the operator felt that they have performed setup 3 and setup 4 faster than others experiments (mean scores of 5 and 6 respectively), even that objective measure point setup 2 as the fastest. An interpretation of this result might indicate that operators were immersed in their task, and that did not have a correct time perception.

Initially, operators complained about limited field of view when looking down, as they could not see the ground and the robot base simultaneous, making difficult to turnaround the obstacle. Another comment was related with the lack of wider field of view and the knowledge of robot size. Corridors with white walls, without reference cues, were pointed as complexity factors. Commands execution delay was also referred as a relevant factor. Issues addressing the field of view (FOV) are presently solved by enlarging the FOV, either with wide the use of angle lens, either with carefully camera positioning (avoiding occlusions with robot parts).

From the experiment external video records, it was possible to observe operator reactions, to critical events, similar to ones as they would behave in the remote environment. For instance, when the robot was about to collide, the user felt the moment and tend to get away from the control station.

To sum up, the use of natural teleoperation styles presents higher embodiment feeling, being demonstrated either by the analysis of self-questioners and by the task time performance. By using operator's head pose to control his visual feedback and body motion to control the robot movement, operator seems to agree that the physical and cognitive workload decrease from setup 1 to setup 4.

4.1.5 Summary

The present study suggests that, when a person is in an remote body, and has the ownership illusion towards remote body, apparently substituting their own, there are autonomic responses that correspond to what would be observed in events that take place in reality. This study results show that introducing new interaction and view control mechanisms that improve the physical embodiment sensation in tele-operation tasks can improve both the user satisfaction and performance. It is clear that viewing the remote environment as one being controlling the robot from inside of it, reduces the required mental workload to compute, the otherwise needed, view and control transformations. In future experiments we intend to explore the introduction of auditory feedback as represented in figure 4.7, to map the proximity of obstacles.

4.2 Design and Evaluation of a Natural Interface for Remote Operation of Underwater Robots

Nowadays, an increasing need of intervention robotic systems can be observed in all kind of hazardous environments. In all these intervention systems, the human expert continues playing a central role from the decision making point of view. For instance, in underwater domains, when manipulation capabilities are required, only Remote Operated Vehicles, commercially available, can be used, normally using master-slave architectures and relaying all the responsibility in the pilot. Thus, the role played by human-machine interfaces represents a crucial point in current intervention systems. In this paper a User Interface Abstraction Layer is presented, which allows a non-expert user to control an underwater robot vehicle by using a new intuitive and immersive interface. Furthermore, the user will receive only the most relevant information about the current mission. Finally, some experiments have been carried out to compare a traditional setup and the new procedure, demonstrating reliability and feasibility of this new approach.

4.2.1 Introduction

The Fukushima nuclear disaster in 2011 had a strong impact on the international community and, in particular, on the vision of the ways that robots should operate in this kind of new challenging missions. Inspired by this terrible accident, a lot of new activities have been started out, like DARPA Challenge in USA, or the Eurathlon competition in Europe, to name a few. This kind of hostile scenarios that preclude the presence of humans, are making mandatory a new generation of intervention robotic systems able of performing the missions that, in other conditions, would be developed by humans experts. Currently, the control of these robots is requiring the supervision of human experts, enabling the possibility of hybrid systems, usually tele-operated and, only in particular situations, with autonomous response. This paper addresses the issue of developing a Virtual Reality (VR) -based interface for providing an immersive experience for the operator that controls the robot's mission. The aim, hereafter presented and discussed, is to provide all the necessary ingredients for, through achieving a compelling sensation of telepresence [27, 393], having the operator control the robot as if he was inside of it, on some kind of cockpit.

Robots can play important roles in many different types of missions, such as maintenance, surveillance, exploration, or search and recovery/rescue (SAR), especially in hazardous environments. In particular, the need for intervention in underwater environments, has been significantly increasing during the last years (e.g. oil and gas industry, SAR, deep water archaeology, oceanography research). These tasks are usually performed making use of work class Remote Operated Vehicles (ROV) launched from support vessels, and remotely operated by expert pilots, through umbilical cables and by using very complex human-robot interfaces.

Besides ROV commercial systems, the Autonomous Underwater Vehicles (AUV), were introduced mainly for inspection tasks. The need for the inclusion of manipula-

tion capabilities gave birth to the Autonomous Underwater Vehicles for Intervention (I-AUV). Since the pioneering works in the 90s, these robots have been used in two main types of interventions: search and recovery (SAUVIM, RAUVI, FP7-TRIDENT) and panel intervention (ALIVE, TRITON, FP7-PANDORA). In all of them, the user is still in the control loop selecting the intervention, supervising the mission, or controlling the robot.

Recently, some autonomous behaviors (e.g. open/close a valve) have been developed in order to increase the autonomy level, reducing so the user cognitive fatigue associated with the ROV systems. Despite the evolution from ROV to AUV in terms of Human-Robot Interaction (HRI), the interfaces in use are still very complex. This is due to the large number of sensors to be monitored and the difficulties of the operations in the underwater domains.

4.2.2 Autonomous versus Teleoperated UVs

We could start here a discussion on if the underwater vehicles (UV) should be autonomous or teleoperated. The fact is that this environment poses several limitations to both approaches, as we will summarize hereafter.

The first and major problem comes with the difficulty in propagating radio-waves in this medium, that voids both the use of wireless communications for teleoperation and the use of GPS-based localization for attaining some autonomy. The choice of strategic sensing strategies of seafloor features for localization is very difficult due to the lack of detectable features, their constantly changing nature, and the limited range of operation of these sensors. As an example of these difficulties we can notice that a camera is able to capture very distinguishable images of the seafloor at very close distance from it, but when these images are taken at a distance greater than 2-3 meters, they are totally blurred by the microscopic elements in suspension. The alternative is to sonar devices that can only provide low resolution representations, captured at a low rate due to the need to mechanically sweep the area of interest.

Teleoperation on the other side, also has a number of problems that start with the umbilical connection required, and that limits the range of operation. Other important problems are mostly related with having a user controlling a large set of variables of the robot, like thrust, direction, orientation of cameras or other sensors, and probably a robotic arm, using only a huge set of information distributed along a set of screens and/or numerical displays, but having a limited view of the task to execute.

Figure 4.11 shows an example of a ROV control room of MBARI Ridges 2005 Expedition [329].

4.2.3 Human in the loop: pros and cons

Teleoperation concept has evolved since the initial remote control experiments of the late 1800s. More, than remotely switching on and off devices, the operator is

Figure 4.11: A typical ROV Control Room. Courtesy of Monterey Bay Aquarium Research Institute.

asked to control systems using his/her ability to interpret the available information and take the most appropriate decisions. This ability to decide, react, and adjust operation in the presence of noisy and incomplete data, makes the human operator still a fundamental and irreplaceable part of many systems. By consequence, as in may other areas, the human factors analysis gained a prominent place. This has led to human-centered approaches in the designing of new systems, aiming at simultaneously increase the performance of the system controlled by the human operator, and reduce the number of failures due to to operator faults. Task performance is frequently measured in number of accomplished tasks per unity of time, which is just the reciprocal of time taken to accomplish a single task. So "doing tasks faster" typically conflicts with "doing tasks well", i.e. without failures. Fortunately, this is not necessarily true as reducing mental workload, providing more natural interaction mechanisms, may simultaneously increase the operator's performance and reducing the number of committed faults.

Starting with the analysis of the typical human errors that may have an impact in teleoperation, we can list three types: issuing a wrong command, issuing a command too late, or not issuing a command at all. These errors can be produced by: (1) the lack of knowledge on how to act in the presence of a given information, (2) the time needed to interpret the received information, or (3) not having received the information at all. The first case which is the lack of knowledge is related with the need to train specialised operators to operate the robots. The second case can be related with mental fatigue [449] that makes the operator take an increasing time to interpret the received information or the time to perceive the received stimuli. The last case of not having received the information may be due to the fact that the user was paying attention to some detail of the task or the interface and did not see or hear the information coming.

Knowing this, we need to search for solutions to help in reducing the number of failures that the operator is responsible for. The first solution is to make the systems more robust to human faults, knowing that they may exist. This implies that these systems have increased intelligence and dispose of additional sources of information that enables them to "adapt" their responses to the user commands by

weighting it by the "sensed danger" they represent. A typical use of these principles is the electric wheelchairs adapted for people that suffer from Parkinson's disease, cerebral palsy or other, so that tremors or imprecise actuations on a joystick does not make the user fall down the stairs or crash against a wall [130]. Other approaches may rely on increased autonomy [173] of the robotic systems so that the user only issues higher-level commands. This reduces the mental workload of the operator, that becomes more a spectator to detect any situation that needs his/her intervention. Mixed situations exist where the operator is asked to do a fine control of some of the robot movements, while simultaneously the robot autonomously is in control of others. Examples of the latter can be found in surgical robots where the robots guarantees that the movements are restricted to a predefined area or volume. Other examples are the flight control of some planes, where an automatic maintains the stability of the plane, helicopter or drone, as the pilot is in control of the flight moves.

A complementary solution to the previous may be in trying to reduce the number of user injected faults. This requires a deeper understanding on the human cognitive and physical factors that may influence the operator ability to execute the expected operations.

From this understanding, special care is put in designing interfaces that into account the user dexterity, induced physical fatigue, required mental workload, attentional mechanisms, etc. The objective is that the systems are developed so that the operator (surgeon, pilot, or other) receives the necessary information to perform the task without the need to search for it in unusual places. Vital information is always placed in visible and in a salient way so that the user can perceive it immediately. And, the controls are adapted to be manipulated in an easy, simple and effective way.

4.2.4 Contributions and research organisation

Guided by these principles, in this research we present a solution for the teleoperation problem based on exploring an immersive system. Such system is used to induce a telepresence feeling so that the operator acts as if he/she was aboard of the robot, reducing the mental workload induced by third person views.

The recent introduction on the market of devices like Kinect™ and Leap Motion™ , which are able to track and estimate the pose of the human body and hands, seems to create an excellent opportunity to replace the traditional joysticks, keyboards and mice. This motivated the study of their benefits by measuring some parameters related to task performance achieved by a group of users and analysing their subjective evaluation in terms of usability, perceived task load and immersive feeling.

The rest of the section is organized as follows: Subsection two presents the proposed architecture and implementation details. Subsection three shows and discusses about the user experience evaluation. Subsection four presents the conclusions.

4.2.5 Designing an immersive teleoperation system

Traditional remote control setups, that typically are composed of multiple displays and controls, when applied to remotely operated vehicles or robots require at least one specialized operator, for successful operation.

The complexity and the number of variables to be observed and/or the number of controls, frequently require that the number of operators increases.

Our proposal aims at simplifying the remote operation control setup, by exploring the principle of telepresence. Our assumption is that if, by the use of some devices, the operator can experiment the sensation of being inside the robot, disposing of a wide field of view (windshield like), then the control task becomes as natural as driving a car. This can be achieved by transforming some of the existent explicit controls into implicit ones, e.g. by controlling the orientation of a camera using head rotation instead of using a joystick or other control for that, reducing both the required dexterity and implied cognitive workload.

Virtual Cockpit: From explicit to implicit controls

Teleoperating any kind of vehicle in some remote environment where the operator cannot have a third person view of it, requires the use of an embedded camera that will be the operator's eyes. To have the ability to perceive the remote environment the operator has to be able to rotate the camera left, right, up and down. If the vehicle cannot perform these rotations about a fixed point, then a pan-and-tilt unit is needed for that purpose. This means that the operator has to control the two degrees of freedom of the camera in addition to those required to pilot the vehicle. This represents increased demands in terms of effort and concentration from the operator.

The solution we have designed addresses this problem, and consists in creating a virtual cockpit (VC) for the operator. This can be achieved by using a Head-Mounted Display (HMD), whose orientation is used to control the PTU, so the user's head movements are implicitly transposed into camera movements. This will enable the user to browse the surroundings of the vehicle, enjoying the sensation of being aboard. By superimposing virtual elements over the camera view, it is possible to create the perception of a cockpit with its instruments.

The chosen position on the robot to fixate the PTU and its camera defines the location of the virtual cockpit. This location has to be carefully chosen as it has to be adequate for proportionating the best view for the task to be performed, e.g. navigation and maneuvering the AUV, or controlling a robotic arm. Having a set of inputs to control the movements of the AUV and/or its robotic arm, the user can operate them enjoying the sensation of being there. This fulfills our goal of providing the user a perception similar to that of driving a car, control an excavator, piloting an helicopter, etc

A more immersive interface

To achieve the aforementioned goal of creating a simpler and more natural user interface for tele operating robots, in particular UVs, we have designed a system that takes the user aboard of the remote vehicle inside a virtual cockpit. This should overcome the limitations, of having a single camera view whose orientation is manually controlled, that normally result in higher demands in terms of concentration, attention, etc. Instead of this, and taking advantage of the already presented UIAL, we have designed system based on the use of an head mounted display (HMD), that enables the user to look in any direction. By proportionating a first person wide field of view, it should induce a sense of presence on the operator, enabling him/her to pilot the UV as if being aboard of it. For instance, the real robot and UWSim have exactly the same communication interfaces (same topics). This means that the real robot can be controlled by "publishing" the commands in a named topic, that can also be used for the simulator to mimic the robot behaviour and enable the operator to visualize the expected behaviour of the robot. This in fact, enables the interface to display any information made available by the sensors or camera streaming, by simply subscribing the corresponding "topic" and processing the received data.

The head movements, captured by the inertial sensors of the HMD, are translated into commands and sent a the remote PTU (pan-and-tilt unit) that supports the viewing camera. This enables the user to browse the remote environment, from side to side, up and down, seeing through the camera.

Being the communications supported by cables, with most of the data flowing from the vehicle to the control station, we can expect that no important delay be introduced in the commands sent to the PTU. Concerning the PTU response, commercial PTUs can have very high performance, exhibiting speeds of 100 degrees per second and more. This is, in fact, below the maximum rotational speed of that the human head can attain, and that can be as high as 365 ± 96 degrees/s [419]. Nevertheless, these higher rotational speeds are normally attained in response to frightening events and not in normal condition of operating or driving a vehicle like a car. In these cases, neck rotations at speeds of tens of degrees/s are used for browsing the view field or visually tracking moving targets. This should enable the user to behave as if he is aboard of the remote robot, and attain a better level of control.

For the control of the robot we have tested both a joystick and a Leap Motion™ (*LM*-device). A noted limitation was the lack of perception of the relative position of the hand with respect to the *LM*-device. Another aspect is the need for some reference frame for the operator, so that he can perceive in any instant if he is looking in the forward direction of the robot, up, down or elsewhere. For this reason an Augmented Reality approach was taken by adding two virtual elements on a fixed position with respect to the user: a virtual table and a virtual joystick on it. The table acts as the reference object that enables the user to know to where he is looking at. The virtual joystick shows the control that is being applied through the device in use (the real joystick or the *LM*-device). To improve the perception of the *LM*-device location, a small fan was placed close to this device. With this the

user can sense the air flow and not only perceive its position, but also the vertical distance between the hand and the device by the airflow intensity. A second approach was to enable the user to see himself in the virtual cockpit and perceive the relative position between his hands and the *LM*-device device. This was done by using an additional RGB-D sensor to capture the 3D point cloud that represents the user body and introduce it on the virtual environment.

In summary the proposed teleoperation setup aims at addressing the problem of simplifying the teleoperation of an underwater robot, by taking the user virtually aboard. This setup has the possibility of integrating different interaction modalities and devices always aiming at reducing the number and complexity of controls required for the operation. This justifies for instance the inclusion of the *LM*-device as it can detect a large spectrum of hand poses and configurations that can be expected to be mapped to robot controls. As will be shown later, this type of device is not as precise as we could expect and introduces other types of problems.

4.2.6 Implementation

The development of the above presented ideas was made upon UWSim [406], an open source simulator for underwater robotic missions, that is under development at the Interactive & Robotic Systems Lab citeirs, at Universitat Jaume I. This software package is currently used in a few ongoing projects funded by European Commission (e.g. PANDORA) to reproduce real missions from captured logs, for user training, to test algorithms, to monitor the robot or as a 3D simulation tool for benchmarking. The adopted architecture is based on ROS [421], that enables the rapid substitution of the simulator by a real robot, or use the simulator to, in parallel with the robot, enhance the user interface with predictive information.

Being one of the purposes of our research to provide insight on which are the best interaction styles and modes for use in the teleoperation of remote robots, and as everyday appear on the market new devices for human interaction, we realised that each of them has to be adapted for each particular use. As an example, two joysticks with different shapes, or a joystick and a yoke-like input, may require a different mappings between the device "axes" and the intended commands. This mapping is not only in terms of establishing pairings, but also on the definition of calibration functions that may vary, between two joysticks, due to shape differences.

Although Virtual-Reality Peripheral Network (VRPN) libraries [1] enables us to extend the range of devices that can be used, it is the User Interface Abstraction Layer (UIAL) that is responsible for providing the correct mappings. In addition it introduces the flexibility in performing online activation or deactivation of interaction devices. As an example, it enables the rapid change of the remote PTU control from a joystick to an HMD's IMU, or the control of the thrust and direction of the robot from a joystick to a *LM*-device, or other. Apart from performing the adequate mapping between devices and the real robot or the simulator, the UIAL is also responsible for the verification of some safety measures that try to minimize some drastic consequences of human errors. These safety measures are guided by

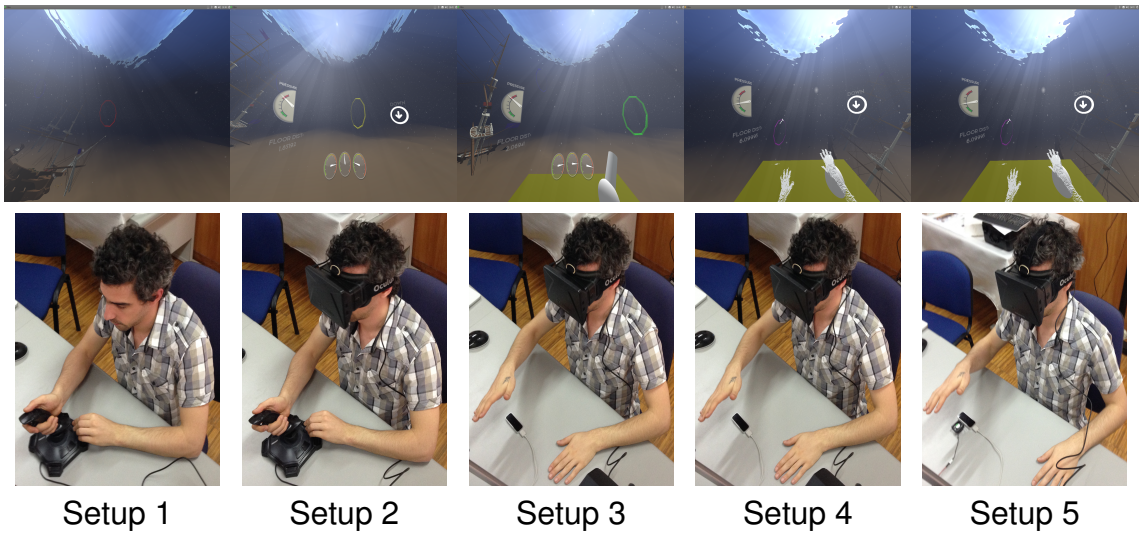


Figure 4.12: Experimental setups: (1) Traditional control; (2) VC with joystick and virtual joystick; (3) VC with LM and virtual joystick; (4) VC with LM and point cloud for arms representation; (5) VC with LM and airflow haptics.

a set of rules which are related with the information provided by some of the on board sensors, like pressure sensors, proximity sensors, etc.

4.2.7 Evaluating the user experience

Although the presented interaction mechanism was developed having in mind the control of real robotic platforms, for the sake of safety and given that the interest is in evaluating the interface, all the tests described hereafter were performed using solely the simulator UWSim.

Methodology

To evaluate the benefits of the proposed changes in the interaction mechanisms for teleoperating remote robots, we simulated a teleoperated underwater vehicle performing a simple obstacle navigation task. We compared our proposed immersive teleoperation approach based on a virtual cockpit with natural egocentric view, against the traditional teleoperation interfaces that use manual camera camera control and visualisation of mission related information through a set of monitors. Two control devices, joystick and *LM*-device, were tested in terms of usability. Figure 4.12 shows a user in different phases of of the test and an example of the scene that is visualized.

The evaluation consisted in analysing a set of performance related parameters, which were collected during the experiments, and the answers given to a short questionnaire after each trial. The collected parameters, the questionnaire and their analysis are presented in the remaining of this section.

Evaluation Procedure

For the purpose of evaluating the effect of immersive technologies on the teleoperation of underwater robots, we have designed an evaluation procedure where participants are invited to control a simulated underwater robot with the objective of completing a trajectory. That trajectory includes passing in order through 5 rings that are not collinear and have different orientations, in minimum time and without colliding with the rings or other underwater structures. For each experiment there is a "warmup" from the starting position until reaching the first ring. The measuring process is started immediately upon passing the first ring. The process is repeated for each of the 5 control setups listed on table 4.1. For each participant the sequence of the setups is random to avoid effects of learning the trajectory that normally improves the performance for the laters to be executed. In fact the setups vary in terms of the type of support for visualisation of the remote environment, the control of the remote camera orientation, and the robot navigation controls.

Test	**Display**	**Camera Orientation**	**Robot Navigation Control**
1	Traditional 2 Monitors	Fixed	joystick
2	Immersive Virtual Cockpit via HMD	Head orientation from HMD IMU	joystick
3			LeapMotion
4			LeapMotion + point cloud
5			LeapMotion + air flow

Table 4.1: The five different test setup combinations for the navigation task

The procedure can be summarised as:

1. Participant is instructed about the task objectives and procedures.
2. Execute trial with 1 of the 5 setups.
3. Fill questionnaire about user experience.
4. Repeat to step 2 until the 5 trials are complete.

Measurements and questionaries

The usability evaluation was performed in two parts: objective task performance related measures and user subjective evaluation through a questionnaire.

Concerning the analysis on performance we measured the following variables directly from the simulator and/or using a third observer to keep records.

- *Time*: navigation time for each of the 4 path segments between rings.
- *Traveled distance*: The length of the executed trajectory for each path segment.

- *Number of collisions*: number of times the robot collided with the elements of the underwater environment, including the rings.
- *Number of steering compensations*: number of issued steering commands for each path segment.

For the subjective evaluation a questionnaire was created, that was inspired on the IBM Computer Usability Satisfaction Questionnaire [282], NASA-task load index (NASA-TLX) [197], as well as on Slater, Usoh and Steed [464, 499] presence questions. The participant feedback was given by classifying on 7 point Likert scales subjects like: usability, easiness, control precision, fatigue, realness, telepresence and embodiment feeling. The 8 questions to answer were divided in two groups as follows (table 4.2 and table 4.3):

Q_i	**Question**	**Ordinal scale to ancor responses**
1	The interface to control the robot was...	(1=Easy to use, 7=Hard to use)
2	How tiring was the task?	(1=Felt tired, 7=Didn't feel tired)
3	How precise was the robot control?	(1=Not precise, 7=Precise)
4	Performing the experiment was ...	(1=Frustrating, 7=Rewarding)

Table 4.2: Usability & Task load questions

Q_i	**Question**	**Ordinal scale to ancor responses**
5	I had the impression of being...	(1=In the lab, 7=Aboard the vehicle)
6	How close I felt from the obstacles?	(1=Felt close, 7=Didn't feel close)
7	How real was the experience?	(1=Close to real, 7=Far from real)
8	The perceived motion sensation was:	(1=I was moving, 7=The scenery was moving)

Table 4.3: Immersion presence questions

Participants

The experiments were performed both at the University of Coimbra, Portugal (UC) and Universitat Jaume I, Spain (UJI), with 13 participants from UC and 13 from UJI. The participant group included students and researchers in fields such engineering and computer science, with an overall average age of 30.12 years. All participants reported normal or corrected to normal vision, where 17 had experience with video games. the subjects had no prior knowledge of the experience or involved technologies. Participation was voluntary, and research ethical principles were attained.

4.2.8 Results

We can divide the participants results in two groups: the quantitative results taking into account the performance in each setup, and the qualitative evaluation. As

previously stated, each participant was asked to execute a set of 5 teleoperation experiments (in random order), and performance related values were collected during each of them. At the end of each experiment, the participants were asked to fill a short questionnaire related with the subjective perception of usability and immersion.

Task performance related quantitative measures

The results are resumed in the following plots for the captured parameters, which are: trajectory time (Fig. 4.13a), traveled distance (Fig. 4.13b), and number of steering commands or compensations (Fig.4.13c). In these figures w_n represents the trajectories between the n and $n+1$ rings, for each of the 5 setups.

Figure 4.13a presents the mean times and variances obtained by the whole set of testers for each path segment (w_n), and for each of the setups. The ANOVA (analysis of variance) test was applied and showed that the results are statistically significant (marked with an asterisk).

($F_{4,95} = 8.57$, $p < 0.0001^*$ on $w1$ time, $F_{4,95} = 2.94$, $p = 0.0242^*$ on $w2$ time, $F_{4,95} = 6.44$, $p = 0.0001^*$ on $w3$ time, $F_{4,95} = 17.79$, $p < 0.0001^*$ on $w4$ time).

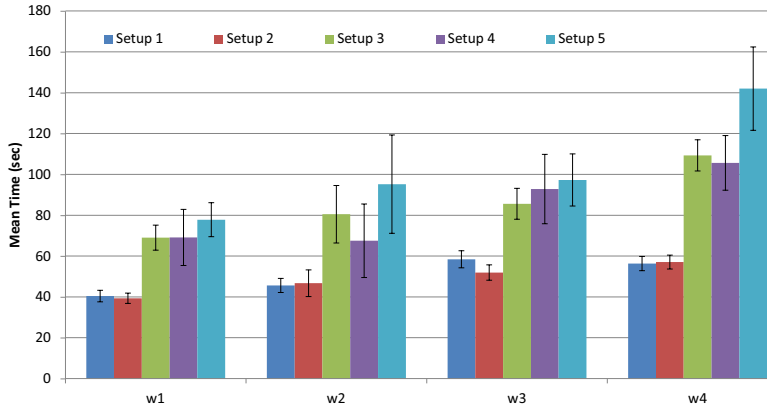
Figure 4.13b presents the mean values for the traveled distances between way points, and the significance analysis gives that only $w1$ and $w4$ results are statistically significant. ($F_{4,95} = 2.77$, $p = 0.0314^*$ on $w1$ dist., $F_{4,95} = 0.87$, $p = 0.4798$ on $w2$ dist., $F_{4,95} = 0.67$, $p = 0.6141$ on $w3$ dist., $F_{4,95} = 2.82$, $p = 0.0290^*$ on $w4$ dist.).

Figure 4.13c depicts a graphic for the number of collisions, although the ANOVA tests show these are not significant from the statistical point of view. ($F_{4,95} = 0.46$, $p = 0.7609$ on $w1$ Col., $F_{4,95} = 0.99$, $p = 0.4168$ on $w2$ Col., $F_{4,95} = 0.87$, $p = 0.4806$ on $w3$ Col., $F_{4,95} = 3.55$, $p = 0.0095$ on $w4$ Col.).

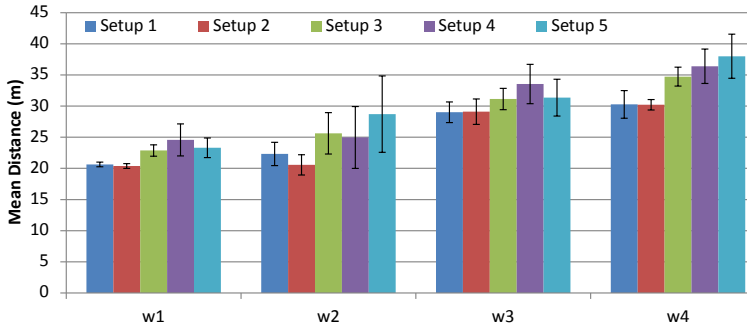
Finally, figure 4.13d presents the mean values of steering commands, and the ANOVA one-way test showed that all the results, except for trajectory segment $w2$, are statistically significant, as follows: ($F_{4,95} = 15.91$, $p < 0.0001^*$ on $w1$ Ord., $F_{4,95} = 3.76$, $p = 0,0068$ on $w2$ Ord., $F_{4,95} = 7.69$, $p < 0.0001^*$ on $w3$ Ord., $F_{4,95} = 15.90$, $p < 0.0001^*$ on $w4$ Ord.) .

Qualitative evaluation based on user questionnaires In what concerns the questionnaire presented in section 4.2.7, the results are presented in figure 4.14.

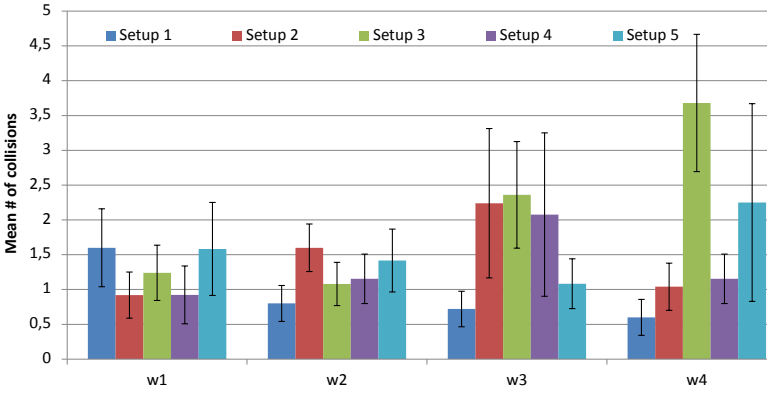
The ANOVA one-way test results are as follows: ($F_{4,95} = 31,59$, $p < 0.0001^*$ on question Q1, $F_{4,95} = 37,97$, $p < 0.0001^*$ on question Q2, $F_{4,95} = 18,45$, $p < 0.0001^*$ on question Q3, $F_{4,95} = 24,57$, $p < 0.0001^*$ on question Q4, $F_{4,95} = 32,19$, $p < 0.0001^*$ on question Q5, $F_{4,95} = 2,86$, $p = 0,0274^*$ on question Q6, $F_{4,95} = 14,06$, $p < 0.0001^*$ on question Q7, $F_{4,95} = 7,46$, $p < 0.0001^*$ on question Q8).



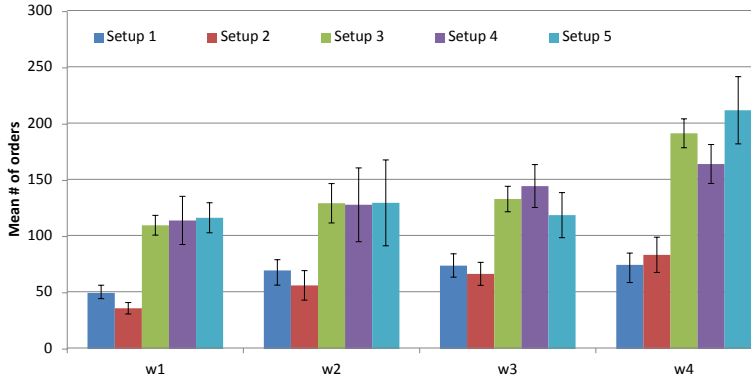
(a)



(b)



(c)



(d)

Figure 4.13: Obtained mean values and standard deviation for: (a) trajectory time, (b) trajectory length, (c) number of collisions, and (d) number of steering commands, per trajectory segment and per setup.

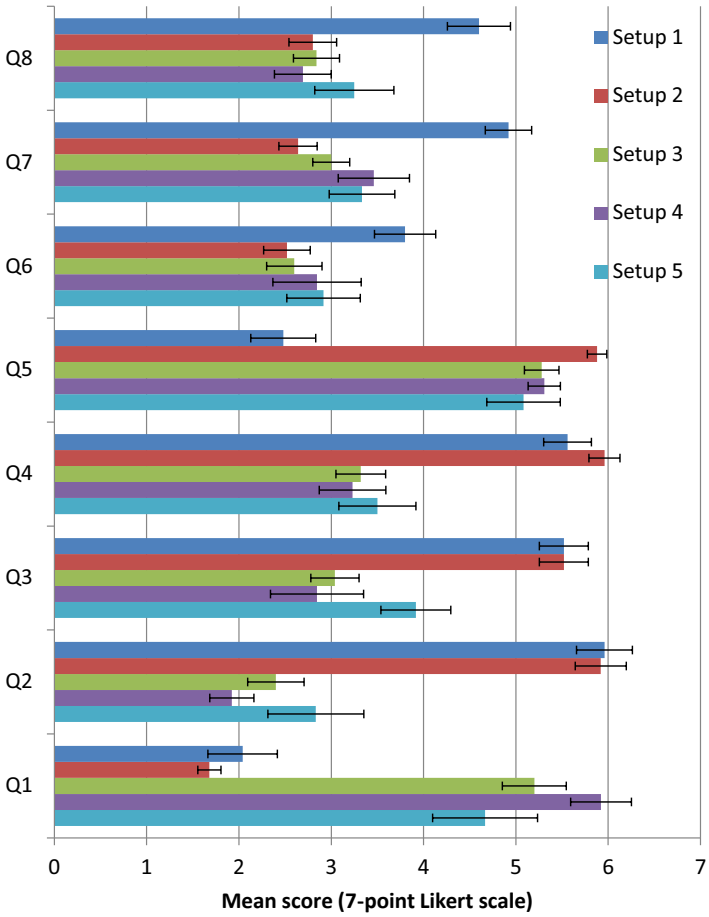


Figure 4.14: Mean scores for the five interfaces obtained from the users answers.

Discussion

The navigation task performance measures, correspond to the execution of a trajectory, divided in 4 segments ($w_k, k = 1..4$), for each of the 5 setups presented. The task was to drive the robot along pass through each ring that separates a path segment from the next. The user could adapt to the commands during the first part of the trajectory, i.e. till passing the first ring and then the all the measures were started for w_1, then after the second ring for w_2, etc.

Each of the trajectory segments had its own particularities and implied complexities, as follows: w_1 - straight forward, w_2 - simple curve, w_3 - hard curve, and w_4 - variations in altitude ending with a curve.

Recalling also the devices used in each setup:

Setup 1 - Conventional teleoperation setup using 2 monitors and joystick, with no camera control.

Setup 2 - Virtual Cockpit (VC) using an HMD for visualisation and controlling camera orientation, and joystick.

Setup 3 - Same as previous, replacing joystick by *LM*-device device.

Setup 4 - Same as previous, with representations of the user and *LM*-device inside the VC.

Setup 5 - Same as previous, replacing point cloud by air flow-based haptic for *LM*-device localization.

The best mean times, on almost all path ways resulted from using Setup 2, except for $w4$ where operators presented less time using Setup 1. Using *LM*-device had generally a negative effect on time performance. Its combination with the air flow solution (Setup 5) showed good results for controlling rotation on a plane, e.g. w_3. For more complex cases of changes on orientation and altitudes (ex: w_4), Setup 4 presented better results. In these cases, operators did not sense air flow changes when moving their hands up and down, but could visualize their hand representation.

The lower distances traveled, on w_1 and w_4 pathways resulted from using both Setups 1 and 2. *LM*-device based interfaces led to bigger traveled distances, as users also reported that it is less precise.

The smallest number of steering order are associated to Setup 2 on w_1 and w_3 pathways. When dealing with changes in altitude canges ($w4$), Setup 1 performed better. Again setups using *LM*-device led to higher number of steering orders.

Operators using Setup 2, the immersive virtual cockpit with joystick, presented better quantitative measures for task time performance, traveled distance and number of steering orders. Setup 1, the traditional approach played better on challenges involving orientation and altitude complexities, although the operators didn't change the camera point of view leading to higher the dexterity workloads

Analysing answers to questions 1 to 4 we can conclude that: (1) The immersive approaches, with the POV based on head orientation are preferred, and (2) the operator considerer important to use a precise device for robot navigation control, like the joystick. Other than precision, (3) having knowledge of range limit of the device and feeling some mechanical feedback from the device is important. This, and the fatigue induced by *LM*-device use made users tend to the joystick as the preferred control device.

From questions 5 to 8 results we can conclude the following. The virtual cockpit solutions are clearly a contribution for the immersion feeling as demonstrated by tele-presence question, Q5. Questions related with tele-embodiment (Q6 and Q8, i.e. virtual contact and self motion) also show a trend to higher immersion rates. Realism question, Q7, present a moderate trend while the operator perceives the simulator as a game and not as a real environment. The perceived realism of the simulated environment exhibits a correlation with the reported ease of use of the control device. This suggests that the simpler and natural is the interaction, the more immersive becomes the experience.

4.2.9 Summary

This study presented the principle that virtual reality-related immersive systems can be used to induce the telepresence feeling in remote operation of underwater robots and that this can be used to improve the performance task execution. To this end, a system was developed with the objective of virtually placing the operator aboard of the remote robot and let him/her do the driving tasks from there. The immersive system combines the images, obtained from an orientable camera on the robot, with virtual instruments. This combination is displayed to the user using a HMD, which tracks the user head movements to modify the POV camera. Additionally, the system can be improved adding to the scene the user own representation.

The evaluation results showed that the immersive system was preferred by the users. Furthermore, when compared with the traditional interfaces, the use of this immersive system has a positive effect in the teleoperation performance.

Concerning the replacement of joystick by a Leap Motion, the results pointed that the lack of touch on the latter has negative effect on the observed user performance and is not appreciated by the users, as they still prefer the former. Another disadvantage reported by the users for the Leap Motion is the fatigue that results from "keeping the hand in the air". Nevertheless, adding both an air flow-based haptic sensation to enable the user to locate the Leap Motion device, or the inclusion of a representation of it the user hand via a point cloud in the virtual environment, showed some improvements in the results, in particular for the first one.

To sum up, the combination of the immersive virtual cockpit, with implicit control of the remote camera orientation from the user head orientation, and joystick, has shown to produce the best results in terms of performance and is the preferred by the users.

4.3 Natural interaction in immersive reality with a cyber-glove

Over the past few years, virtual and mixed reality systems have evolved significantly yielding high immersive experiences. Most of the metaphors used for interaction with the virtual environment do not provide the same meaningful feedback, to which the users are used to in the real world. This research proposes a cyber-glove to improve the immersive sensation and the degree of embodiment in virtual and mixed reality interaction tasks. In particular, we are proposing a cyber-glove system that tracks wrist movements, hand orientation and finger movements. It provides a decoupled position of the wrist and hand, which can contribute to a better embodiment in interaction and manipulation tasks. Additionally, the detection of the curvature of the fingers aims to improve the proprioceptive perception of the grasping/releasing gestures more consistent to visual feedback. The cyber-glove system is being developed for VR applications related to real estate promotion, where users have to go through divisions of the house and interact with objects and furniture. This work aims to assess if glove-based systems can contribute to a higher sense of immersion, embodiment and usability when compared to standard VR hand controller devices (typically button-based). Twenty-two participants tested the cyber-glove system against the HTC Vive controller in a 3D manipulation task, specifically the opening of a virtual door. Metric results showed that 83% of the users performed faster door pushes, and described shorter paths with their hands wearing the cyber-glove. Subjective results showed that all participants rated the cyber-glove based interactions as equally or more natural, and 90% of users experienced an equal or a significant increase in the sense of embodiment.

4.3.1 Introduction

Virtual and immersive reality (VR) are technologies that can find many applications that go far beyond gaming, in areas such as rehabilitation, real estate promotion, education and medical training, museum exhibitions, showrooms, simulation of accident scenarios, police training, social training, etc. The potential of adapting any space to a dynamic new virtual world in which the user can move in, opens a whole of new challenges related to immersion. The effectiveness of VR environments strongly depends on providing the same stimuli as those experienced in the real world. In order to enhance user's immersion, embodiment and presence [82][472][450], we need to support a natural consistency between the vestibular and proprioceptive feedback in addition to the visual feedback, while enabling a precise tracking of body parts [459]. Interaction based on active movements contributes for the "sense of agency", that is, the sense of having "*global motor control, including the subjective experience of action, intention, control, motor selection and the conscious experience of will*" [74].

There are several low-cost commercial VR systems available that provide an effective user experience in large spaces (e.g., HTC Vive, Oculus Rift S), with a very reliable body tracking. For user interaction with VR, most systems use

hand-based controllers with click buttons and inertial sensors. However, the interaction is not always perceived as natural, because while users hold these controllers they cannot grab or touch real objects in a mixed reality interaction, or it compromises the embodiment in virtual reality interaction. In a previous paper [27] we explored the notion of tele-presence and physical embodiment. The aim was to virtually transfer the operator to the remote robot to improve teleoperation, maximizing task performance and minimizing the operator's physical and cognitive workload. One of the system's limitation was the lack of hand interaction with objects. Although human body parts were mapped through a skeleton representation, the hands and fingers were not tracked, compromising the manipulation tasks. Recent approaches based on glove systems can help achieve a more natural user interaction, freeing user's hands, allowing the detection of finger movements, haptic feedback and gesture recognition [132][211][231][333]. Cyber-gloves open a new range of applications in gaming, industry, surgery training, rehabilitation and education. Gloves with haptic feedback are being proposed for hand and finger rehabilitation [489][211], or surgery training [93]. These glove–based systems aim to provide feedback to the users to enable the perception of virtual objects. Several technological approaches to this problem have been proposed in literature, which include: the use of force sensitive resistors combined with vibro-tactile actuators to provide force feedback to the user [211]; fingertip contact pressure sensors, capable of providing vibratory and visual stimulation [489]; or twisted string actuation integrating force sensors and small-size DC motors [217]. The use of bend sensors and IMU (inertial measuring unit) is also commonly used to track, respectively, fingers and hand position and orientation [489]. The use of cyber-gloves for gaming is proposed in [10] and [227]. The work in [227] describes an exoskeletal VR glove that tracks the user's physical finger movement, and is capable of translating the movement to virtual fingers in a game environment. Haptic feedback is provided by attaching motors to each finger joint. A VR glove for falconry is proposed in [10], which is intended to give the player the illusory sensation of a falcon standing in their hand. All of these approaches are prototypes and yet lack the desired usability and wearability combined with reliable exteroceptive perception, which makes user interaction still not very natural.

In this paper we propose a cyber-glove to improve the immersive sensation and the degree of embodiment in virtual and mixed reality interaction tasks. In particular, we are proposing a glove system that can track wrist movements, hand orientation and finger movements. Most VR systems do not provide a decoupled position of the wrist and hand, although actual hand manipulation tasks require that these two movements are independent. Thus, the perception of these two different degrees of freedom significantly increases the embodiment perception in all kind of hand interaction tasks, for which the joint between the hand and the wrist is not rigid. Furthermore, the curvature angle of the fingers is detected, supporting an effective perception of finger movements. Cyber-glove has features comparable to other commercial products such as CyberGlove III [118]. However, the cost of the cyber-glove prototype is significantly lower (a few hundred of US dollar versus more than $10K). Additionally, our device has longer autonomy and enables a direct contact of the palm of the hand with objects. It has also a heart rate sensor and a range sensor on the back of the hand that detects nearby obstacles, alerting

the user with a vibro-tactile stimulation. The system is being tested in a virtual environment where users have to perform interaction tasks such as grab and rotate door handles, push or pull drawers. We compare the naturalness, usability, immersion and embodiment perception using the developed cyber-glove and the HTC hand trackers. These interaction tasks are part of a global set commonly used by a person when exploring a real house. This work is being developed in the scope of a major project (called HTPDIR) in partnership with a company that aims to do real estate promotion. The tests aim to understand how gloves can improve user interaction with the house being visited.

4.3.2 The System – Immersive room testbed

The present application builds a VR immersive scenario based on Unity 3D providing egocentric visualization and interaction. It is being developed targeting mixed reality applications where real objects are dynamically mapped in the VR scenario through several Kinect RGB-D sensors [178]. This paper is only focused on the design and development of an immersive cyber-glove and its use in VR environments that include manipulation tasks.

Immersive cyber-glove – hardware architecture

The immersive glove detects wrist, hand and finger movements with the purpose of complementing the immersive tracking system. The absolute position of the wrist is tracked with HTC Vive trackers, while the hand and finger movements are tracked with our own customized glove (see Fig. 4.15). The hardware architecture is shown in Fig. 4.16. The glove prototype is composed of the following modules: an IMU Sensor (BNO055) with ARM Cortex M0 incorporated, an wi-fi module ESP8266-07 with a Tensilica Cadence L106 embedded microcontroller, a PPG sensor (MAX30105) that is responsible for detecting the heart rate (that can be used to evaluate emotional reactions), a micro-laser range sensor (VL53L0X) responsible for measuring the distance to physical obstacles, 5 Flexible sensors (Spectra Symbol 2.2) that sense the curvature angle of the fingers, 5 force sensors (FSR 400) that provide touch information in mixed reality interaction, a vibro-motor that is actuated when the back of the glove is near a physical obstacle (for safety reasons) and a LiPo battery (BAT525). The BNO055 is a system in a package, that integrates a triaxial accelerometer, a triaxial gyroscope, a triaxial geomagnetic sensor and a 32-bit microcontroller. It merges the accelerometer, gyroscope and the magnetometer within 9 degrees of freedom, returning an absolute orientation with a high throughput without distortion of the magnetic field. The microcontroller of the ESP8266-07 module receives via I2C the data from BNO055, namely the rotation in quaternions and the acceleration data in x, y, z coordinates. The ESP8266-07 also interfaces all the remaining analog, digital and I2C sensors, computes the heart rate, and sends to BNO055 its initial configuration settings. The wi-fi module of the ESP8266 sends all glove's data in UDP/IP packets to the Unity application running on the host PC that does the data processing and visual rendering. To prolong the battery autonomy of the cyber-glove, several

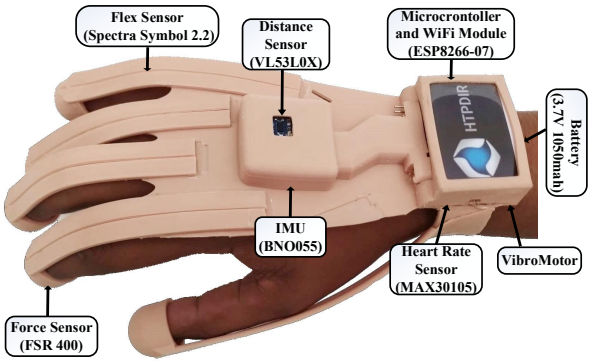


Figure 4.15: Photo of the developed cyber-glove prototype (HTPDIR). Description of main hardware components integrating the cyber-glove.

sensors were programmed to be in "sleep mode" whenever they are not being used. In normal operating mode the cyber-glove power consumption is about 95 mA, while in "sleep mode" it is about 30 mA. The minimum duration of the battery in the normal operating mode is approximately 8 hours. The glove was designed in SolidWorks and printed in a 3D printer Sigma using FilaFlex material, a very resistant and flexible material (elongation at break – 665%) that provided a good hand fit sensation. The palm of the hand is free which is quite important for the mixed reality experiments.

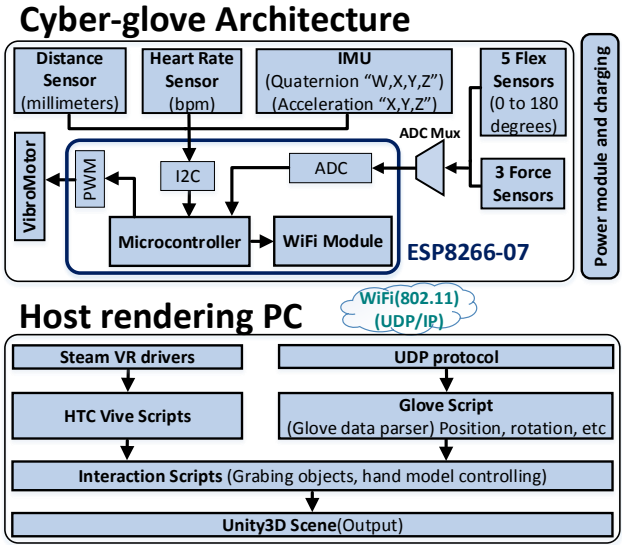


Figure 4.16: Cyber-glove and 3D Unity framework architecture.

Immersive environment - Unity 3D

Each task and effect created in the virtual environments is based on various Unity scripts and GameObjects. The script that interacts with the glove receives all the data via UDP packets every 20 ms, namely, angular position of the hand, acceleration, range sensor distance, heart rate, fingers curvature, pressure feedback, and sends commands to actuate the vibro-tactile motors that provide haptic feedback to the user. These data are forwarded to specific scripts attached to each object

to be manipulated. The hand's script receives the quaternions to animate the hand model (see Fig. 4.17). This hand model is attached to the wrist position detected by the HTC Vive Tracker. The wrist reference system $\{x, y, z\}_{wrist}$ has 6-DoF to which the hand reference system $\{x, y, z\}_{hand}$ is attached, having 3-DoF (roll, pitch, yaw). The reference system $\{x, y, z\}_{hand}$ is a translation of $\{x, y, z\}_{wrist}$ along the z-axis. Additionally to 3D position, we get linear acceleration values for each coordinate. Fingers curvature values are obtained from flexible sensors that return a bending angle in a range of 0º to 180º and feed a Unity's Hinge Joint component that couples two rigid GameObjects. This enables a rotation along one of the common axis, reproducing prehensile gestures. In the extremity of each finger there is a force feedback sensor to inform the VR application that thumb and index finger are touching each other (i.e. a gesture event) or that the fingers are in contact with a real object. Interaction with objects is managed through scripts that parents the object to the hand on pickup. The attachments between objects can be rigid or use a Joint to integrate force feedback for the physics engine. These scripts also define the pick-up and release methods. When a hand and finger models are in contact with virtual objects, the respective Collider triggers the vibrotactile motors to simulate the real touch feedback. The heart rate sensor provides information that can be used to assess the user's engagement and for example change the application dynamics.

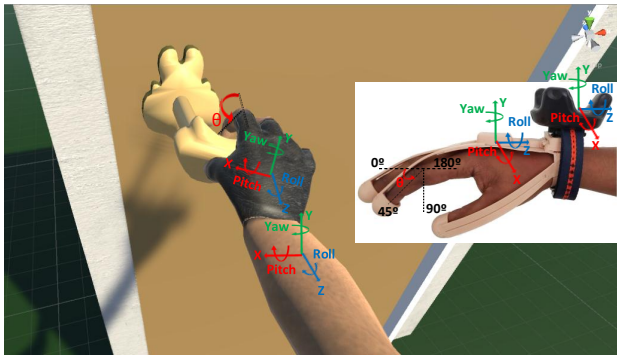


Figure 4.17: Picture of the immersive scenario during a hand manipulation task: door handle rotation. Reference systems of hand and wrist in the immersive environment from an egocentric view, and respective reference systems of the cyber-glove and wrist tracker in real world.

For this experiment, a virtual door scenario was created to compare the performance of the cyber-glove and the HTC Vive Controller. Two input methods were implemented to open the door: the "DoorViveController" method that enables door opening using the HTC Vive Controller and the "DoorGlove" that refers to the door being opened by the cyber-glove. In "DoorViveController" the user presses a controller's button to grab the door's handle while in "DoorGlove" the user closes/opens his/her hand (fingers) to grab or release the door's handler. Both door-opening methods use the parenting concept to rotate the door's handle and relies on a Hinge Joint. A "Quaternion.LookRotation" method enables the rotation of the door's handle regarding to the hand position. For the "DoorGlove", hand's position results from a translation of the wrist position, provided by Vive Tracker, however their orientations can be independent (see Fig. 4.17). For the "DoorViveController", the hand and wrist are a rigid body with position and orientation provided by the

HTC Vive controller.

The door's handle rotation is computed using the EulerAngles method and it is limited to predefined angles. Consequently, the script that rotates the door is enabled only if the door's handle rotation reaches a predefined angle. After the rotation of the door handle, the position of the hand is used to compute the door's rotation angle. This component is attached to the door frame through a Hinge Joint, which enables a rotation along a vertical axis.

To create a fully VR environment that integrates real objects of the room, we also developed a support framework that relies on HTC Vive controllers to pinpoint the object's corner coordinates.

4.3.3 Method

The evaluation of virtual or mixed reality applications requires an analysis of factors like naturalness, usability, immersion, embodiment and task performance, which can be assessed through user's actions such as VR 3D navigation and manipulation. This section describes the proposed evaluation methodology for this VR application and devices.

Participants

Twenty-two participants (7 women and 15 men) from the Polytechnic Institute of Tomar were invited to test the system. Participants were mainly students and researchers from engineering courses. Participants were aged between 20 to 46 years old (μ=26.56 , σ=6.66). Participation in the experiment was voluntary. Four of the 22 participants never had contact with video games technology and only 3 subjects had previous experience with interaction devices like the proposed cyber-glove.

Materials

Users viewed the virtual and the mixed reality environment through a head mount display (HMD), which is a fully immersive reality helmet that presents three-dimensional stereoscopic views. The HMD is part of the Vive VR System having two AMOLED screens, a resolution of 1080 x 1200 pixels per eye (2160 x 1200 pixels combined), a refresh rate of 90 Hz and a field of view of 110 degrees. Internal inertial sensors (accelerometers and gyroscope) and an outside-in laser tracking system provided user's head position and orientation to render the virtual world accordingly. User's hand movements and gestures were tracked either through HTC Vive controllers, or through a wrist-tracker strapped around the wrist. These trackers, when attached to a real object provided also its position. Additionally, the cyber-glove enabled the mapping of the movements of the wrist, hand and fingers in the immersive environment.

Experience design

One of the applications of this work is real estate promotion with virtual enviroments. A typical task commonly performed at home has been designed, *a person crosses the several divisions of a house and for this he/she has to open a door and pass through it*. One of the focus was on the embodiment and realism of the movement of the hand during the handle rotation, for which the detection of movement of the wrist and hand are essential. This simple task comprises a series of small steps that people are used to do in the real world. However, the recreation in the virtual environment presents some technological challenges. Thus we have designed a virtual door and a metaphor to transpose it. Each person, wearing a VR HMD system that provides an egocentric view, was invited to perform the following door opening based sub-tasks (see Fig. 4.18):

- A - Walk to the door
- B - Rotation of the door handle
- C - Push the door
- D - Pass through the door

These tasks were repeated by each subject 2 times:

1) holding the standard HTC Vive hand controller that has a click button interaction, and provides position and orientation. To grab the door's handle, the user had to press a button; and

2) wearing the developed cyber-glove, that frees the hand and enables grab functionalities. To grab/release the door's handle, the user had to close/open the hand (fingers).

Furthermore, qualitative and quantitative evaluations were performed, using the two hand-based input devices. Figure 4.19 shows a photo of a subject using the the cyber-glove while performing the immersive interaction task.

Experiment procedure

Subjects were invited individually to the lab and informed about the procedure to open the virtual door. One of the project's researcher was in charge of helping the subjects to wear the equipment, and a software application recorded the performance measures during task execution. Each participant performed the task only once in each condition. No training was provided to better assess the intuitiveness of the solution. The experiments involving the HTC Vive controller and the cyber-glove were performed randomly and at the end, users filled subjective questionnaires.

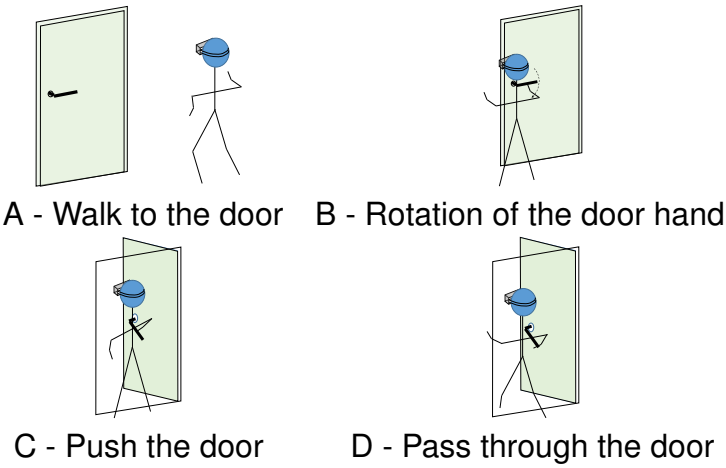


Figure 4.18: Steps to open the door in the immersive environment.

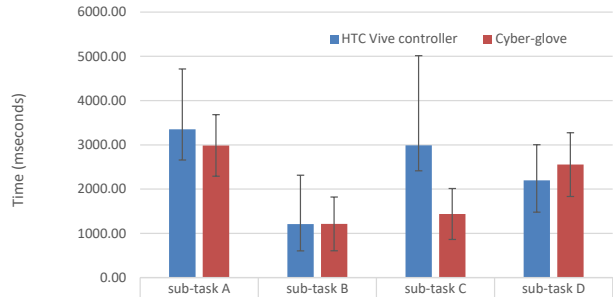
Evaluation metrics

Qualitative and quantitative measures were accessed for the two different hand-based input devices:

Efficiency measures

- Time - measures the time taken by a person to open a door and pass through it:
  - Total time: overall time to accomplish the task;
  - Sub-task time: measure the time taken by a person in each door opening tasks (A, B, C and D);
- Length - length of the path described by the hand in each sub-tasks (a shorter paths means less effort and better naturalness);

Figure 4.19: Photo taken at a national exhibition of a subject using the cyber-glove while performing the immersive interaction task.



	sub-task A	sub-task B	sub-task C	sub-task D
HTC Vive controller	3352.98	1210.83	2989.09	2199.14
Cyber-glove	2986.73	1213.22	1437.38	2553.24

Figure 4.20: Mean task-time performance of participants for each sub-task of the virtual door opening global task (HTC Vive controller vs cyber-glove).

Immersion and presence

In order to evaluate the experience qualitatively, the subjects were invited to fill a questionnaire on a scale from 1 to 7, where 1 meant *very weakly* and 7 *very strongly*, enabling to determine issues like naturalness (question Q1), usability (question Q2), immersion feeling (question Q3) and sense of embodiment (question Q4). These questions were adapted from IBM Computer Usability Satisfaction Questionnaire [283], NASA-task load index (NASA-TLX) [197], and from Usoh and Slater Presence Questionnaire [499].

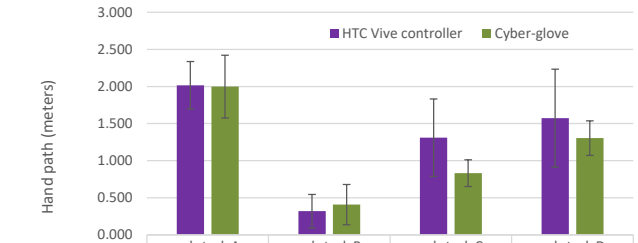
- Q1 - How natural was the interaction with the VR environment?
- Q2 - How easy was manipulating and moving objects in the VR environment?
- Q3 - How strongly was the immersion feeling in the VR environment?
- Q4 - Did I feel that my own hand was manipulating and moving objects in the VR environment?

4.3.4 Results - Task Performance

Objective Results

Figure 4.20 shows the mean of task-time performances in each sub-tasks of the virtual door opening task, using the HTC Vive controller or the cyber-glove. Statistical significance was assessed using repeated measures (within subjects) ANOVA (analysis of variance) test (asterisk mark indicates statistically significant). Results show that in sub-task C users pushed the door faster wearing the cyber-glove ($p < 0.05$):

- sub-task A: F(1,22)=0.69, p=0.4156
- sub-task B: F(1,22)=4.34E-05 , p=0.9948
- sub-task C: F(1,22)=6.51, p=0.0181*



	sub-task A	sub-task B	sub-task C	sub-task D
HTC Vive controller	2.016	0.320	1.310	1.574
Cyber-glove	1.998	0.408	0.832	1.304

Figure 4.21: Mean length of the path described by the hand of participants, for each sub-task of the virtual door opening global task (HTC Vive controller vs cyber-glove).

- sub-task D: F(1,22)=1.29, p=0.2680
- Total task-time : F(1,22)=1.59, p=0.2204

Figure 4.21 shows the mean length of the path described by the hand in each sub-tasks of the virtual door opening task, using the HTC Vive controller or the cyber-glove. Statistical significance was assessed using repeated measures ANOVA test:

- sub-task A: F(1,22)=0.014, p=0.9054
- sub-task B: F(1,22)=0.751, p=0.3952
- sub-task C: F(1,22)=9.012, p=0.0065*
- sub-task D: F(1,22)=1.781, p=0.1956
- Total task-length: F(1,22)=3.575, p=0.0718

Metric results revealed that 83% of the users performed faster door pushes, and described shorter paths with their hands wearing the cyber-glove ($p < 0.05$).

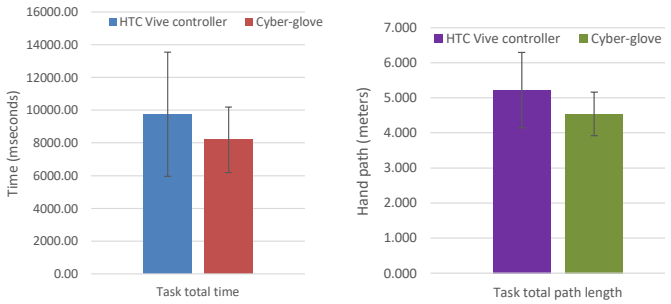
In Fig. 4.22 we present the total mean task-time results, and the total mean length of the path described by the hand of participants. The time and length of the global task are smaller for the cyber-glove. The statistical significance test for the total path length presents a p=0.0718.

Subjective Results

Figure 4.23 illustrates the results of the questionnaire.

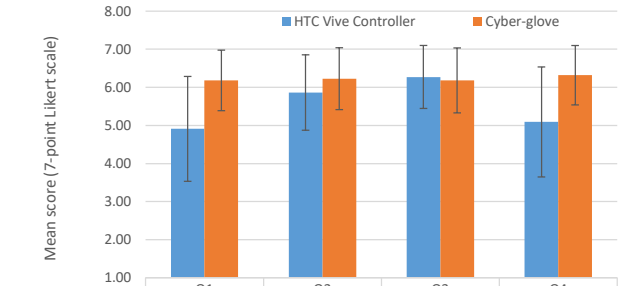
The within subjects ANOVA one-way test for each question shows its statistic significance (asterisk marks):

- Q1: F(1,42)=14.10, p=0.00052*



a) b)

Figure 4.22: a) Total mean task-time, and b) total mean length of the path described by the hand of participants while performing the virtual door opening global task, HTC Vive controller vs cyber-glove.



	Q1	Q2	Q3	Q4
HTC Vive Controller	4.91	5.86	6.27	5.09
Cyber-glove	6.18	6.23	6.18	6.32

Figure 4.23: Results of subjective questionnaires to participants to evaluate the virtual door opening task.

- Q2: F(1,42)=1.77, p= 0.19018
- Q3: F(1,42)=0.13, p=0.72144
- Q4: F(1,42)=12.29, p=0.00109*

100% of the participants rated the cyber-glove based interactions as equally or more natural, and 90% of users experienced an equal or a significant increase in the sense of embodiment.

Discussion

Objective results related with task-time performance in sub-task A (“Walk to the door”) shows that there is not a significant time difference at reaching the door, neither there is any differences in the length of path described by the user’s hand, i.e. the hand device does not influence this sub-task. Concerning the time performance in sub-task B (“Rotation of the door handle”) the participants took the same time to perform this sub-task with both devices. Users with the HTC controller described a path slightly shorter than with cyber-glove while rotating the door’s handle. According to our expectations and in relation to sub-task C (“Push the door”), users pushed the door to an angle of 60º faster with the cyber-glove

than with the HTC controller. The length of the trajectory described by the hand while wearing the cyber-glove was also shorter. Users reported that releasing the door's handler was easier with the cyber-glove because they just had to open the hand/fingers, and they were not concerned about the instant to release the HTC controller button. Thus, users performed sub-task C better with the cyber-glove, being both task-time and hand's length path statistically significant. Concerning sub-task D ("Pass through the door"), users were faster transposing the door's frame holding the HTC Vive controller than when they were using the cyber-glove, however they described a longer path holding the HTC Vive controller. For this sub-task, the start matches the moment when the user removes their hand from the door handler and moves himself to a certain distance from the door.

Qualitative evaluation based on questionnaires to the users shows that the cyber-glove contributes for a greater naturalness. Factors like usability and immersion feeling seem similar for both devices, and the sense of embodiment is significantly improved with the cyber-glove. Several participants reported orally that with the cyber-glove they had more freedom of movement, while with the HTC Vive controller they felt hand movement constraints, and were afraid of dropping the controller during the release of the door handle. Overall, the results of the ongoing cyber-glove prototype are very promising. Yet, some limitations were identified during the experiments, namely, the size of the glove is not suitable for every users due to different hand sizes, and the movement of the virtual fingers exhibits a small latency in relation to the real fingers movement, due the animation model.

4.3.5 Summary

This work describes a cyber-glove system that tracks wrist movements, hand orientation and finger movements, aiming to improve the immersive sensation and embodiment in virtual and mixed reality environments. The proposed system is capable of decoupling the position of the wrist and hand, contributing for the sense of embodiment in 3D VR manipulation. The comparative study between the cyber-glove and the HTC Vive controller showed that the task-time performance of pushing a virtual door's, is faster when wearing the cyber-glove, and additionally the hand of the users describes shorter paths. The subjective questionnaires show that the cyber-glove contributes for a greater naturalness, present similar degrees of usability and immersion feeling, but improves significantly the sense of embodiment. Future work includes glove refining to better adapt to the different hand sizes, and improve VR models of the fingers. More complex tasks in virtual and mixed reality will be carried out to validate the overall system.

4.4 Conclusions

The present study suggests that, when a person is in an remote body, and has the ownership illusion towards remote body, apparently substituting their own, there are autonomic responses that correspond to what would be observed in events that take place in reality. Results show that introducing new interaction and view control mechanisms that improve the physical embodiment sensation in tele-operation tasks can improve both the user satisfaction and performance. It is clear that viewing the remote environment as one being controlling the robot from inside of it, reduces the required mental workload to compute, the otherwise needed, view and control transformations. The study demonstrated that virtual reality-related immersive systems can be used to induce the telepresence feeling in remote operation of underwater robots and that this can be used to improve the performance task execution. To this end, a system was developed with the objective of virtually placing the operator aboard of the remote robot and let him/her do the driving tasks from there. The immersive system combines the images, obtained from an orientable camera on the robot, with virtual instruments. This combination is displayed to the user using a HMD, which tracks the user head movements to modify the POV camera. Additionally, the system can be improved adding to the scene the user own representation. The evaluation results showed that the immersive system was preferred by the users. Furthermore, when compared with the traditional interfaces, the use of this immersive system has a positive effect in the teleoperation performance.

Concerning the replacement of joystick by a Leap Motion, the results pointed that the lack of touch on the latter has negative effect on the observed user performance and is not appreciated by the users, as they still prefer the former. Another disadvantage reported by the users for the Leap Motion is the fatigue that results from "keeping the hand in the air". Nevertheless, adding both an air flow-based haptic sensation to enable the user to locate the Leap Motion device, or the inclusion of a representation of it the user hand via a point cloud in the virtual environment, showed some improvements in the results, in particular for the first one.

To sum up, the combination of the immersive virtual cockpit, with implicit control of the remote camera orientation from the user head orientation, and joystick, has shown to produce the best results in terms of performance and is the preferred by the users. Additionally, was developed a cyber-glove system to enable hand natural gestures interactions. it tracks wrist movements, hand orientation and finger movements, aiming to improve the immersive sensation and embodiment in virtual and mixed reality environments. The proposed system is capable of decoupling the position of the wrist and hand, contributing for the sense of embodiment in 3D VR manipulation. The comparative study between the cyber-glove and the HTC Vive controller showed that the task-time performance of pushing a virtual door's, is faster when wearing the cyber-glove, and additionally the hand of the users describes shorter paths. The subjective questionnaires show that the cyber-glove contributes for a greater naturalness, present similar degrees of usability and immersion feeling, but improves significantly the sense of embodiment.

Chapter 5

Co-Presence - a Fast 3D Model Acquisition

The chapter proposes a fast 3D model acquisition framework contributing to copresence in human-centered mediated communications, HRI and HMI. It includes published works in HRI, HMI, computer graphics conferences [28][26][22], in the 3D Research journal [14], and unpublished work.

5.1 Incremental 3D Model Building

The work presents a solution for one of the problems of creating a teleconferencing system that goes beyond traditional video-conferencing systems. The possibility of having meetings between people or teams at long distances without the need to travel has become attractive in terms of time and economic savings. Nevertheless, these are frequently less than satisfactory as the sense of presence does not really exist. One can imagine two teams trying to convince each other of (or blaming each other for) a given subject. When more than one person is on both sides of a table, deictic gestures or eye contact are frequently used to simplify communication or develop some empathy. Typical video conferencing systems do not provide this type of communication, which is one of the disadvantages most pointed out by the users. With this in mind, we propose a communication system that explores the concept of telepresence so that users have the true feeling of being in the presence of others. For this to be possible, we must be able to illude the human senses like sight and audition. Smell and touch could improve the presence feeling, but let's restrict it to situations where people are not that close. The sound source's position is essential, but solutions exist that can provide a good approximation for that problem. Given this, we remain with the issue of giving the participants the visual sensation of being in the presence of each other. For this, if one participant moves w.r.t. the other, he will see the second person from a different perspective, e.g. glancing to one's side. For this to be possible, complete 3D models must be sent between communication endpoints.

There are various possibilities for 3D model acquisition, from 3D laser scanners

to vision stereo rigs. The former provides dense and accurate data; however, they are generally slow, cumbersome, and, more importantly, too expensive for a consumer application. The latter requires considerable processing power and does not behave well in regions with low texture profiles. The recent introduction to the market of consumer RGBD sensors opens an opportunity for tele-immersion applications for the general public.

Some notable works realistically exploit the user's appearance for tele-immersion, like those developed at UC Berkeley [266] and GrImage at INRIA [399]. Both use video cameras array to perform real-time full-body 3D reconstructions. There are indeed some areas for improvement that can be identified in those approaches, like: as reconstruction problems due to the lack of accuracy in low-texture or repeated pattern regions, high-cost acquisition data setups, high power computational requirements, and their unsuitability for domestic use.

As an improvement opportunity to overcome those weaknesses, we propose to exploit these new RGBD sensors and use natural human motion as an ally to incrementally obtain a 3D model of the persons involved in the communication. It is a framework with a low computational cost suitable for real-time applications. The 3D mesh models integrate new data as the user shows more of their body while moving. For this, we propose a new incremental version of the Crust algorithm [30] that incrementally builds a surface mesh from the registered 3D point clouds maintaining the topological correctness converging to the original surface. On the other hand, data should be integrated into the mesh if, and only if, it can contribute to refining the model; otherwise, it should be discarded. For this, we use local entropy measures to decide if a given part of the model should be improved by integrating new information. The approach reduces the computational load to the necessary, as receiving 3D data for regions of the body that were "already seen" will not force the integration of new points on the mesh and subsequent dimension reduction. In these cases, such data will be used solely for pose estimation, which is required for the telepresence application.

Figure 5.1 depicts the application concept goal in which displays, video cameras, a depth sensor, microphones and speakers enable users to communicate and interact remotely, experiencing the benefits of a face-to-face meeting in full size. If user B moves his pose from C_0 to C_1 he will see user A through a new point of view.

Figure 5.2 presents an overview of the reconstruction algorithm that aims to continuously generate a realistic body model, transfer the model and reconstruct it on a remote standard display or virtual environment according to each user's viewpoint by a tracking process (view-dependent synthesis).

Accurate tracking of the viewer's head and rendering view-dependent images on standard screens (ex: TVs and LCDs) enables us to create the illusion of an actual window.

The remainder of this section is organized as follows: the next subsection 5.1 presents an overview of related works, provides background about 3D reconstruction and emphasizes our contribution regarding the incremental 3D model building. Section 5.1.1 describes the suggested methodology, and section 5.2.7

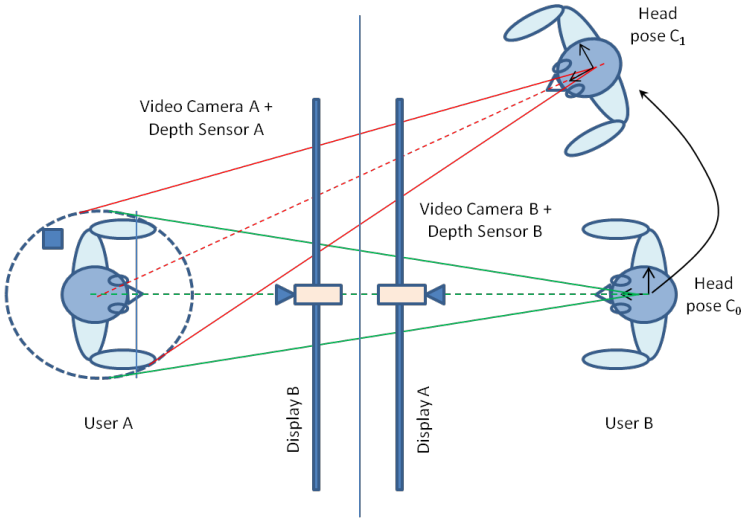


Figure 5.1: Face to face meeting through technology mediation, line of sight preserving method

presents implementation, calibrations and experimental results. Finally, section 5.2.8 presents future work and conclusions.

Background

Virtual view synthesis and modelling are the potential graphic tools to create the eye-to-eye contact illusion on tele-presence communications [229][81][501][14][20]. Real-time human body 3D reconstruction approaches can be divided into three categories: silhouette-based reconstruction, voxel-based methods with space sampling and image-based reconstruction with dense stereo depth-maps. Usually, the body surface is reconstructed by merging sensor data from different views. Two types of information are required: depth data and sensor pose data.

When there is no prior information about depth and pose, the reconstruction technique relies on structure from motion. In such cases, the sensor ego-motion estimation is based on corresponding features found in consecutive images. The depth information, without absolute scale, is then computed using the obtained ego-motion information.

When depth information is available a priori, but sensor pose is still unknown, using data resulting from a ToF or structured light depth camera, a laser scanner or a stereo camera without inertial sensors, the reconstruction techniques are usually based on the Iterative Closest Point (ICP) algorithm [64]. 3D point clouds acquired from different views are registered onto the same referential by iteratively matching overlapping surfaces. This method is computationally heavy for real-time applications.

When depth and sensor pose data are known a priori, no registration procedure is required to merge the data into a global referential. The precision of depth measurements and sensor pose estimation influences the final surface reconstruction quality.

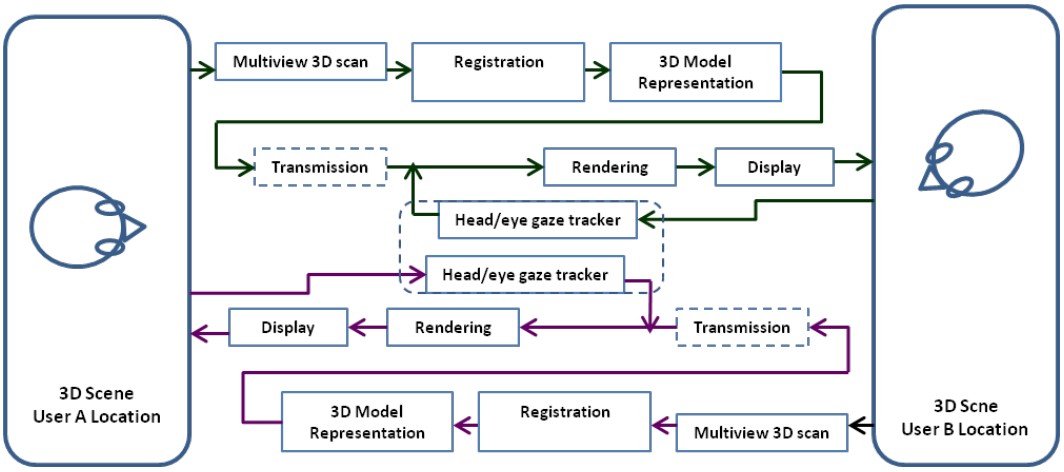


Figure 5.2: Overview of the reconstruction algorithm that aims to continuously generate a realistic body model, transfer the model and reconstruct on a remote common display or virtual environment according, each user's viewpoint by a tracking process. The proposed real-time 3D full reconstruction system combines visual features and shape-based alignment between consecutive point clouds while the mesh model representation is updated incrementally using a new Crust based algorithm.

Recent depth sensor devices provide 3D measurements and RGB data, enabling 2D image algorithms. It is possible to improve the 2D feature mapping between consecutive RGB images, associating the respective depth data and creating a 3D feature tracking. 2D image features mapping approaches are generally based on Kanade-Lucas-Tomasi (KLT) method [452][304][494], Scale-Invariant Feature Transform (SIFT) method [302] or Speed Up Robust Features (SURF) method [57]. Several works use these techniques to track 3D pose sensor changes, either for object detection, path planning, gesture recognition or reconstruction purposes [205] [347] [9] [328] [28] [334].

Our incremental reconstruction approach based on confidence measures aims to update a pre-built mesh surface with new scanned 3D points. The proposed strategy adds new data to the reconstructed model, refining model parts and discards redundant data information, lowering the computational requirements. Online applications benefit from this incremental approach that minimizes registration errors and filters scanning noise. The object surface reconstruction solution aims to incorporate some desirable properties like:

- *Incremental and independent updating:* incremental reconstruction enables incorporating new scanned data information into the reconstructed model without recompiling all models. It is an essential characteristic for real-time applications with limited computational resources and, at the same time, enables further refining while providing global object visualization.

- *Range uncertainty representation:* acquired 3D points present different confidence levels according to the sensor pose and position. Typically the distance from where the sensor acquires the data and from where it stands

in relation to the surface normal is inversely proportional to data information confidence. Accuracy measures depend on the incident angle.

- *Efficient use of all range data:* redundant surface observations leads to scanning noise reduction. It enables to refinement of the model and incorporates time object changes.
- *Geometric and topological surface structure representation:* the surface data representation should take advantage of the underlying geometric and topological structure (ex: curvature, point adjacency, surface orientation and data confidence). Mesh triangles-based surface representations are ideal for performing data reduction without losing the geometric structure (mesh simplification) and enable the synthesis of any 3D model point targeting realistic visual rendering purposes.
- *Discard redundant information:* by discarding 3D points that do not add information to the surface topological structure, fewer data need to be processed. For example, mesh simplification procedures minimize the number of triangles representing a plane.
- *Robust against scanning noise, registration errors and outliers:* ambient lights variation and sensor introduces noise on measures that can be mitigated through redundant surface measures. Wrong pair-wise image point feature correspondences can cause problems, so outliers should be removed to avoid object motion estimation errors. Such error can also be compensated a posteriori through a confidence-based weighting function that changes spatial point positions on overlapping regions.

3D modelling consists mainly on four main phases:

1. Scan object surface from different views
2. Register the views
3. Integrate the views
4. Render the integrated data

Bernardini et. al. [63] makes an extensive survey about the pipeline operations to create a 3D model: data acquisition, range image registration, line-of-sight uncertainty, mesh integration, surface resolution and colour, and texture mapping.

Related work approaches concerning the third phase, *integrate the views*, follow the three classes: Mesh Integration, Volume based integration and Point based integration.

- *Mesh Integration*

  Two unstructured point clouds from a scanned object are converted into two polygonal meshes (ex., triangles). Such representation enables the use of

the surface' topological geometric information. Overlapped mesh regions are detected and discarded. The remaining meshes are reconnected to construct the global surface [481][497][434].

Turk and Levoy [497] delete the overlap triangles regions until they intersect only along a boundary, being later *zippered* together. Soucy et. al. [470] uses the canonic view concept. The integrated surface model is piecewise estimated by triangulations modelling of each canonical subset of the Venn diagram of the set of range views. These triangulations are subsequently connected to yield a global surface. Rutishauser et. al. [427] triangulate the overlapped mesh by growing it at its contour.

Pito [402] defines the concept of co-measurements based on the position and orientation of the range scanner. Only the most confidently acquired measurements are kept. The redundant triangles are removed, and then the patches of triangle meshes are seamed together.

- *Volume based integration*

  The overlapping area is detected, interpolated, and the surface is extracted using implicit volumetric reconstruction methods. It can cope with any topology on a bounded volume. Sampling noise leads to mesh noise. It is needed to provide exact surface topology due to the intersection interpolation between the implicit surface and the voxels. Common approaches use marching cube algorithms to extract triangular meshes [117][209].

- *Point based integration*

  The Cartesian 3D space is initially decomposed into multiple equally sized voxels. All points that fall into the same voxel are then integrated as a consensus point without considering the topology between points. Volume integration mainly differs from point-based integration in the way how it extracts the triangular mesh. Volume integration uses marching cube algorithms, while point-based integration considers the intersection between voxel edges and a plane perpendicular to the orientation at the consensus point. Large registration errors and changes in the density of points in 3D space may lead to algorithm fails [426].

  Recent RGB-D reconstruction-related works are using alignment and integration approaches based on SLAM sparse methods [205][361] [514]. Henry et al. [205] combine visual feature matching with ICP-based pose estimation to build a pose-graph which they optimize to create a globally consistent map. The resulting point cloud map is post-processed to generate a surfel model that significantly reduces the map storage requirements whilst providing a visually smoother representation of the environment. Newcombe et al. [361] presented an improved accurate solution known as KinectFusion, which uses a new algorithm for real-time dense 3D mapping. KinectFusion integrates depth maps from the Kinect into a truncated signed distance formula (TSDF) representation. The pose estimation required to fuse the depth maps is based on a fast iterative closest point (ICP) GPU implementation. The TSDF is discretized into a voxel grid, typically 512x512x512, representing a physical space volume (e.g. a 3 m cube). Each voxel contains two numbers: a signed

distance d indicating how far that cell is from a surface and an integer weight w representing confidence in the accuracy of the distance. If $d < 0$ then v is "inside" a surface; if $d > 0$, then v is "outside" the surface. Only depth values within a truncation band $-T < d < T$ are stored (a typical value is T = 0.03m); the remaining voxels are sentinels with either w = d = 0 (uninitialized) or $d = T$ (empty space). The actual world surfaces are encoded as the zero crossings of the distance field and can be extracted by ray casting or marching cubes. The KinectFusion original representations restriction (a single cube grid of voxel (e.g. 5x5x5m)) was overcome using cyclic buffers on recent Whelan et al. [514] work Kintinuous enabling long-range data integrations while sensor translates.

Contributions beyond state-of-art

Henry's [205] approach is not completely real-time while requiring a post-processed step to generate a surfel model. The KinectFusion [361] solution has a high computational requirement being impractical without a high-end GPU. The iterative characteristic of ICP algorithm and the requirement to store the scene volumetric data on GPU memory imposes high requirements. Our proposal differs rather strongly from the KinectFusion, because the estimation object pose is performed using full RGB-D measurements instead of depth maps only. Planar surfaces, for which the KinectFusion fails to track the object, might be minimized this way. On the other hand, using a robust feature detector avoids the feature-matching ambiguity problems associated with homogeneous areas and repetitive patterns such as occurring on walls, tables, etc. The system has lower computational requirements using a closed-form solution instead of an iterative alignment method like ICP. Rather than using a static volumetric representation that limits scene representation either in detail, as in extension, we propose an incremental dynamic mesh representation that incorporates in the faces, the entropy confidence measures to robust the model and integrates new meshes using an incremental version of Crust Algorithm. Our work aims a real-time incremental body modelling.

5.1.1 Modeling the telepresence 3D video conference tool

The present section details the algorithmic solutions suggested by the flow chart of figure 5.2. The global 3D reconstruction model framework describes the mesh generation process used to synthesize virtual views, from the data acquisition stage (depth + RGB), through the registration, tracking, mesh model generation, incremental integration and refining stages (figure 5.10), required to render views consistent with a real face-to-face meeting.

The mediation goal is to place one user in front of the other in a shared mixed virtual space (see Figure 5.7).

Geometric Definitions

This subsection introduces some geometric definitions

Point Sets: Let P denote the set of sample points in the 3D Cartesian space.

$$P := \{p_1, ..., p_n \in \Re^3\} \tag{5.1}$$

Point Neighborhood: Let $N_R(q, P)$ denote the ballpoint neighbourhood or euclidean point neighbourhood of a query point $q \in \Re^3$ within a point set P with the neighbourhood radius R.

$$N_R(q, P) := \{p \in P \mid d_p^2(q, p) \leq R^2\} \tag{5.2}$$

The function $d_p(q, p)$ is the unsigned distance between two points.

Triangle Meshes

Edges and Triangles: Let $\overline{ab}$ denote the line segment or *edge* connecting the two points $a, b \in \Re^3$ with a $a \neq b$

$$\overline{ab} := \{x \in \Re^d \mid \mathbf{x} = \lambda_1\mathbf{a} + \lambda_2\mathbf{b}; \lambda_1 + \lambda_2 = 1; \lambda_1, \lambda_2 \geq 0\} \tag{5.3}$$

Following such approach, let $\triangle(a, b, c)$ define a *triangular face* connecting three non-collinear points $a, b, c \in \Re^3$

$$\begin{aligned} &\triangle(a, b, c) := \\ &\{x \in \Re^3 \mid \mathbf{x} = \lambda_1\mathbf{a} + \lambda_2\mathbf{b} + \lambda_3\mathbf{c}; \\ &\lambda_1 + \lambda_2 + \lambda_3 = 1; \lambda_1, \lambda_2, \lambda_3 \geq 0\} \end{aligned} \tag{5.4}$$

Mesh Definition: A triangle mesh M is a piecewise, linear approximation of an unknown surface S by triangular faces and is described by the triple

$$M := (V_M, E_M, T_M) \tag{5.5}$$

where V_M is a set of n vetices, $V_M := \{v_1, ...v_n\} \in \Re_3$, E_M a set of m edges $E_M := \{e_1, ...e_m\}$, T_M a set of k triangles $T_M := \{t_1, ...t_k\}$ under conditions that

$$\forall e \in E_M : \exists a, b \in V_M : e = \overline{ab} \tag{5.6}$$

and

$$\forall t \in T_M : \exists a, b, c \in V_M : t = \triangle(a, b, c) \wedge \overline{ab}, \overline{bc}, \overline{ac} \in E_M. \tag{5.7}$$

In an actual face-to-face meeting, when one of the users moves his point of view, he starts to see new views from his interlocutor or the surrounding scene. For example, when someone is talking and looking directly at his interlocutor's face, he cannot see the interlocutor's ear, although moving one step aside, he might see it (Figure 5.3).

The virtual window concept tries to simulate the same visual perception that someone gets when looking at a real scene through a glass window, but replaces it with an artificial display system. The challenge is to generate scene image views in the display plane precisely as could be observed by a person while moving his head or body in front of the real window (Figure 5.3). When the user moves in front of a window, new object views are seen, object parts become occluded, and new parts appear. Such visual perception can be simulated through technology mediation in which the real scene is acquired, represented as a 3D model, and 2D virtual views are synthesized according to the user's point of view and presented on an artificial display (virtual window) (Figure 5.4).

Changing the observer B pose from $\mathbf{C_0}$ to $\mathbf{C_1}$ is equivalent to a pose change of the 3D data acquisition sensors (video camera A + depth sensor A) which provides new 3D map perspectives (Figure 5.3). Once new images are computed on sensor A, according to the observer view point, it is possible to establish a homography between sensor image plane A and the display A plane (virtual window).

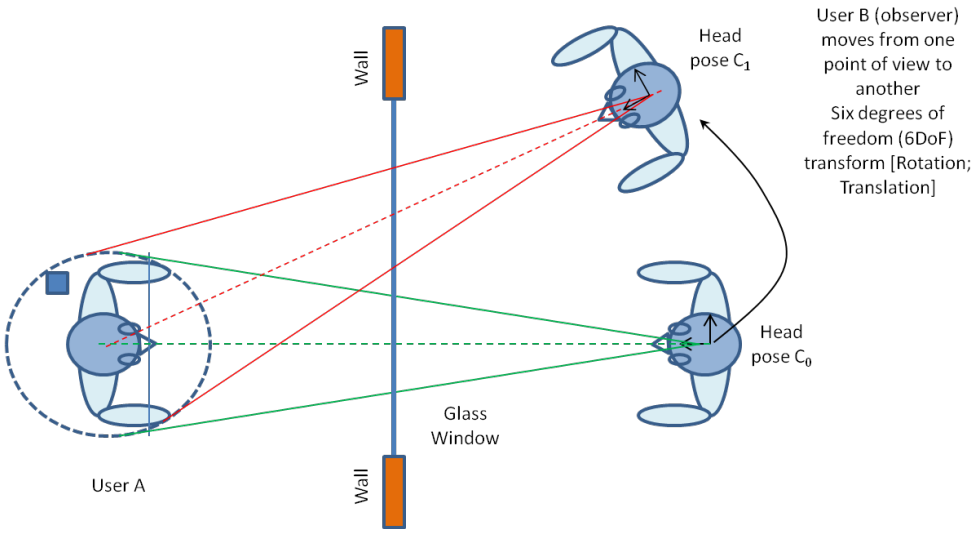


Figure 5.3: Face to face meeting through a glass window

View dependent image synthesis

The projection of a real 3D object model point X_i on the virtual window plane π, denoted by X_i' and observed from the point of view C_0 is illustrated in figure 5.5. If the observer assumes a new point of view C_1, the X_i projection will be seen at virtual window location X_i''.

The virtual window geometry problem can be stated as follows:

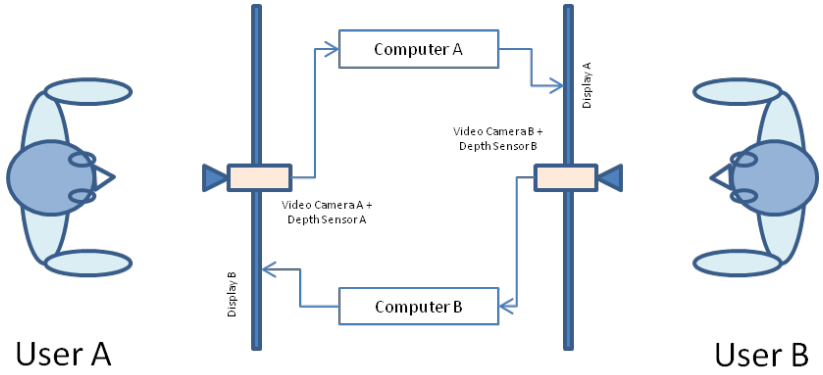


Figure 5.4: Technology mediation setup: Video cameras, depth cameras, lcd displays and computers

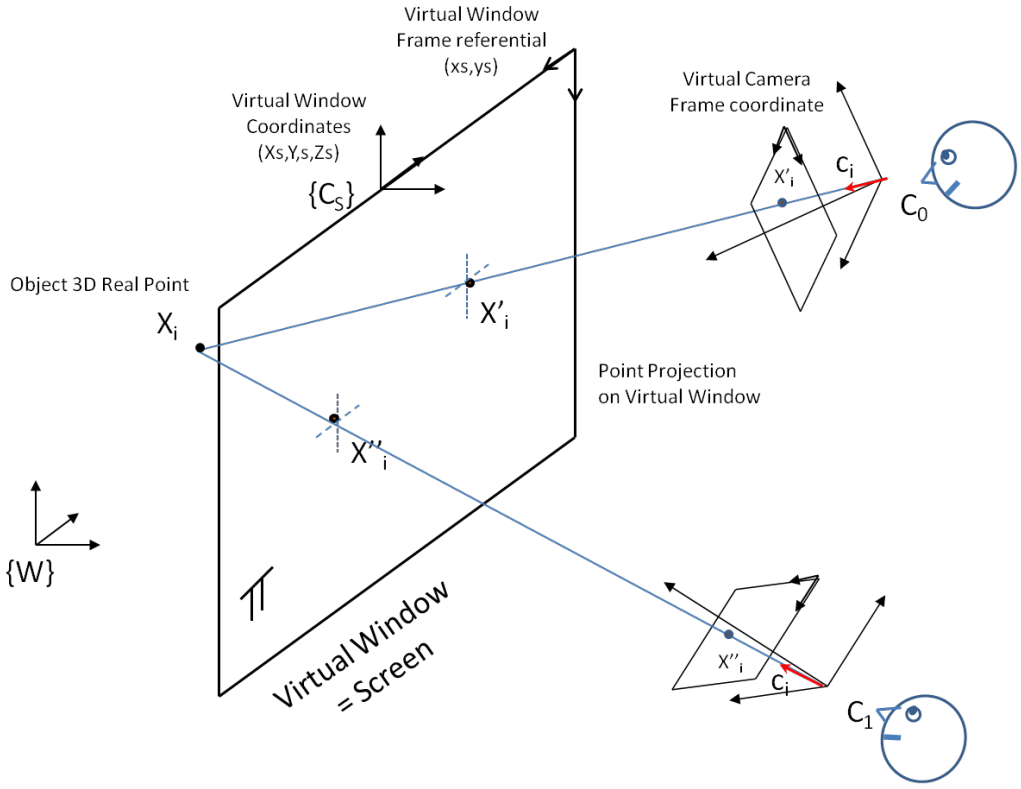


Figure 5.5: Virtual Window Geometry Concept

Given a set of 3D points $\mathbf{X_i}$*=*$(X_i, Y_i, Z_i)^T$ *(reconstructed points), a display plane* π *specified by its coordinates in word coordinate* $\{\mathbf{W}\}$ *and a center of projection* $\mathbf{C_i}$ *(point of view), determine the projection of* $\mathbf{X_i}$ *on plane* π*,* $\mathbf{X_i^{'}}$*=*$(X_i^{'}, Y_i^{'}, Z_i^{'})^{T}$.

It is the point that is supposed to be seen on the window plane when the users look at the real 3D point from a position vector $\mathbf{c_i}$. Let's assume that the user's head pose $\mathbf{c_i}$ is known through a head tracking technique.

The intersection of the line defined by space point $\mathbf{X_i}$ and $\mathbf{C_i}$ with the display plane π results on a space point $\mathbf{X_i^{'}}$ that it is supposed to elude the user's eyes in the absence of the real point.

A plane can be represented as a set of point $X_i^{'}$ for which

$$(\mathbf{X_i^{'}} - \mathbf{X_0^{'}}) \cdot \mathbf{n} = \mathbf{0} \tag{5.8}$$

where $\mathbf{n}$ it is a vector normal to the plane and $\mathbf{X_0^{'}}$ is a point on the plane.

The vector equation for the intersecting line is:

$$\mathbf{X_i^{'}} = \delta \mathbf{c_i} + \mathbf{C_0} \tag{5.9}$$

where $\mathbf{c_i}$ is a vector in the direction of the line and $\mathbf{C_0}$ is a point of the line.

Replacing $\mathbf{X_i^{'}}$ on the plane equation by the line equation, the intersection equation result on

$$(\delta \mathbf{c_i} + \mathbf{C_0} - \mathbf{X_0^{'}}) \cdot \mathbf{n} = \mathbf{0} \tag{5.10}$$

, which due to the distributive property can be written as follows

$$\delta \mathbf{c_i} \cdot \mathbf{n} + (\mathbf{C_0} - \mathbf{X_0^{'}}) \cdot \mathbf{n} = \mathbf{0} \tag{5.11}$$

Solving in order of δ

$$\delta = \frac{(\mathbf{X_0^{'}} - \mathbf{C_0}) \cdot \mathbf{n}}{\mathbf{c_i} \cdot \mathbf{n}} \tag{5.12}$$

So replacing δ on $\mathbf{X_i^{'}} = \delta \mathbf{c_i} + \mathbf{C_0}$ result on the virtual point $\mathbf{X_i^{'}}$.

Thus, when the user moves to a new viewpoint position $\mathbf{C_1}$, a new virtual view $X_i^{''}$ must be synthesized.

Central Projection Mapping and Homography

The projected image $\mathbf{X_i^{'}}$, of a 3D point $\mathbf{X_i}$, on a plane π (virtual window plane) result from the intersection of a line that contains the point $\mathbf{X_i}$, the projection center

$\mathbf{C_i}$ and the plane. In order to avoid computing all the intersections separately, it is possible to model a virtual pinhole camera at $\mathbf{C_i}$ location, obtain the $\mathbf{X_i}$ projection on its image plane, and through a homography process, map those image's points onto plane π (virtual window plane). Such a virtual pinhole camera simulates a cyclop eye at the observer's head for a given point of view.

Pinhole camera geometry

The pinhole camera is represented by its *optical center* $\mathbf{C}$ (camera projection center) and the *image plane* [198][535]. The distance between $\mathbf{C}$ and the image plane is called *focal length f*. The line from the camera center, perpendicular to the image plane, is the camera *principal axis* or *optical axis*. The camera *principal plane* is the plane parallel to the image plane containing the optical center. Let the center of projection be the origin of a Euclidean coordinate system where z-axis is the principal axis. The image plane z=f is also called *focal plane*.

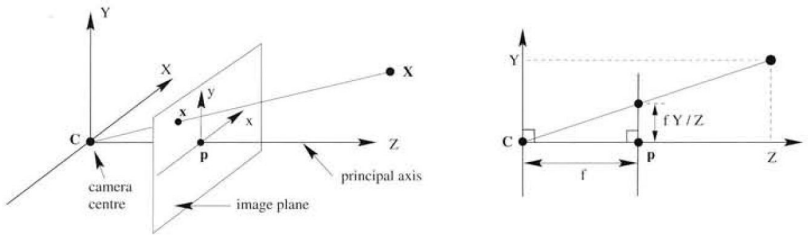


Figure 5.6: Pinhole camera geometry [198]. The left figure represents the projection of a 3D point $\mathbf{X}$, on the image plane result from the intersection of a line containing the point and the the projection center $\mathbf{C}$.

According the pinhole camera, a 3D point with coordinates $\mathbf{X} = (X, Y, Z)^T$ is projected on the image plane where a line containing the point and the optical center intersects the image plane (figure 5.6).

The relation between the 3D coordinates and the coordinates of its projection onto the image plane is described by the *central* or *perspective projection*. By triangle similarity, it is seen that a 3D point $(X, Y, Z)^T$ is mapped to the point $(fX/Z, fY/Z, f)^T$ on the image plane. Representing world and image coordinate through homogeneous vectors enables us to define perspective's projection as a matrix multiplication:

$$\begin{pmatrix} fX \\ fY \\ Z \end{pmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{5.13}$$

The matrix representing the linear mapping is called the *camera projection matrix P* and equation (5.13) can be written compactly as:

$$\mathbf{x} = P\mathbf{X} \tag{5.14}$$

where $\mathbf{X} = (X, Y, Z, 1)^T$ are the homogeneous coordinates of the 3D point and $\mathbf{x} = (fX/Z, fY/Z, 1)^T$ are the homogeneous coordinates of the image point.

The projection matrix P in equation (5.13) encodes only the focal distance f, which represents one of the elementary possible cases. The camera projection matrix is a 3x4 full rank matrix and, being homogeneous, it has 11 degrees of freedom. Matrix P factorization, using QR factorization, is expressed as:

$$P = K[R|\mathbf{t}] \tag{5.15}$$

where K is upper triangular (nonsingular), $\mathbf{t}$ is a translation vector and R is a rotation matrix. K is the *camera calibration matrix*, as it encodes the transformation from camera coordinates to pixel coordinates. It includes the *intrinsic parameters*, namely *focal length f*, image center coordinates in pixels o_x, o_y and pixel size in mm s_x, s_y along the two axes of camera sensor [535]:

$$K = \begin{bmatrix} f/s_x & 0 & o_x \\ 0 & f/s_y & o_y \\ 0 & 0 & 1 \end{bmatrix} \tag{5.16}$$

The *camera extrinsic parameters* are the translation vector, $\mathbf{t}$, and the rotation matrix, R, which specify the transformation between camera and the world coordinate system.

Homography

A set of 3D points and a camera center define a cone of rays, while the projected images results from the intersection of these rays with the plane. Considering that such cone of rays intersects two planes, this means that exists a perspective map relating the two images. In Figure 5.5, for example, the images x_i on the virtual camera image plane and $X_i^{'}$ on the virtual window plane are related by a perspective projection map. Given the image points on a plane and having the same camera center, such points can be mapped to one another plane by a plane projective transformation [198]. The two images x_i and $X_i^{'}$ are related by a *homography*. To obtain the homography formula, lets consider two cameras with the same center $\mathbf{C}$ (equation 5.17):

$$P = KR[I|-\mathbf{C}] \quad , \quad P^{'} = K^{'}R^{'}[I|-\mathbf{C}] \tag{5.17}$$

Since both cameras shares a common center, the following relation can be written, $P^{'} = (K^{'}R^{'})(KR)^{-1}P$. This means that the images of a 3D point $\mathbf{X}$ on the two cameras are related as:

$$\mathbf{x}^{'} = P^{'}\mathbf{X} = (K^{'}R^{'})(KR)^{-1}P\mathbf{X} = (K^{'}R^{'})(KR)^{-1}\mathbf{x} \tag{5.18}$$

In summary, the corresponding image points are related by a planar homography (a 3x3 matrix) as:

$$\mathbf{x}^{'} = H\mathbf{x} \tag{5.19}$$

where $H = (K'R')(KR)^{-1}$.

Returning to the problem of generating view-dependent images in the virtual window plane (figure 5.5), it is possible to acquire a scene image using a real camera at the user's point of view location and, through a calibration process, compute the planar homography that relates the camera image plane and the display plane (virtual window plane).

Sharing a virtual space

The goal is to place one user in front of the other in a shared mixed virtual space (see Figure 5.7).

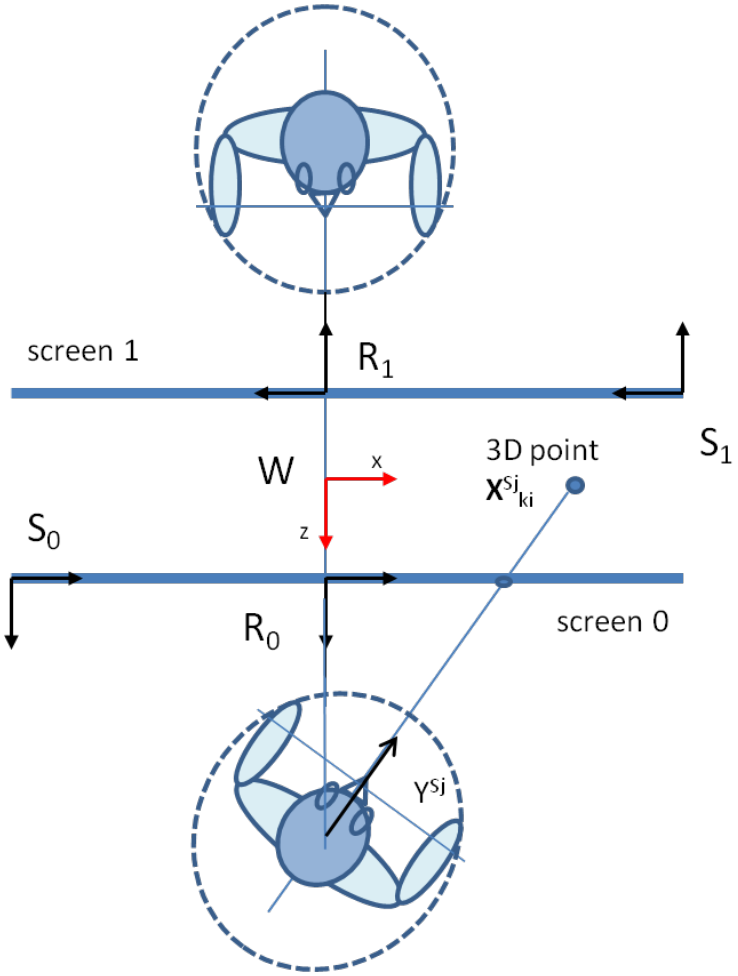


Figure 5.7: Face to face geometry

The reconstructed point's clouds are in their respective sensor device coordinate, respectively $\mathbf{R_0}$ and $\mathbf{R_1}$.

Let the virtual world coordinates be $\mathbf{W}$. The sensor referential coordinates relates to virtual world coordinates through a rotation and translation and for a particular 3D point $\mathbf{X_i^k}$, where i is the index of the points in the reconstructed point clouds from the k sensor referential:

$$\mathbf{X_{ki}^{W}} = {}^{W}R_k\mathbf{X}_i^k + {}^{W}t_k \tag{5.20}$$

and the points will be placed in the correct location in the virtual world. ${}^{W}R_k$ and ${}^{W}t_k$ represents the rotation and translation from k sensor coordinate system to the virtual world coordinate system.

During rendering, given the viewer's eye location in the virtual world's coordinate system and the screen's display region in the same coordinate system, all point clouds or meshes can be correctly rendered. As referred before, the user eye point of view is tracked to provide a sense of motion parallax. The suggested eye-tracking approach is performed by detecting the 2D eye's position using Haar's cascade method and associating that location to the corresponding 3D point on the reconstructed point cloud, which enables a head/eye gaze position vector. Such information is already in the sensor coordinate system.

For the j_{th} user, S_j denotes the local coordinates system of the screen, and C_j, is the local coordinate of the tracking sensor (RGB and depth cameras from Kinect). Let the relation of the user's eye position with C_j be Y_j . The following equation transforms it to world coordinates:

$$\mathbf{Y_{j}^{W}} = {}^{W}R_j\mathbf{Y}_j + {}^{W}t_j \tag{5.21}$$

where ${}^{W}R_j$ and ${}^{W}t_j$ are the rotation and translation from the sensor (Kinect) local coordinate system to the virtual world's coordinate system. The rendering process can be computed in the screen coordinate system. To transform the point cloud and the viewer position to the screen coordinates system S_j, the following equations are developed:

$$\mathbf{X_{ki}^{S_j}} = ({}^{W}R_k)^{-1}\mathbf{X}_{ki}^{W} - {}^{W}t_{S_j} \tag{5.22}$$

$$\mathbf{Y^{S_j}} = ({}^{W}R_k)^{-1}\mathbf{Y}_{ki}^{W} - {}^{W}t_{S_j} \tag{5.23}$$

here ${}^{W}R_k$ and ${}^{W}t_{S_j}$ are the rotation matrix and translation vector from the screen's local coordinates to the virtual world's coordinates.

Eye to Eye contact

Eye-to-eye contact is important in human conversation, and conference mediation technology should preserve such cues. In typical 2D video conferences, like skype, one user cannot simultaneously look at the camera acquiring their image and look at the remote user image displayed on his screen. To accomplish such an illusion, one must consider a virtual acquisition sensor located in the screen at the remote user image eyes middle point (cyclop virtual sensor)

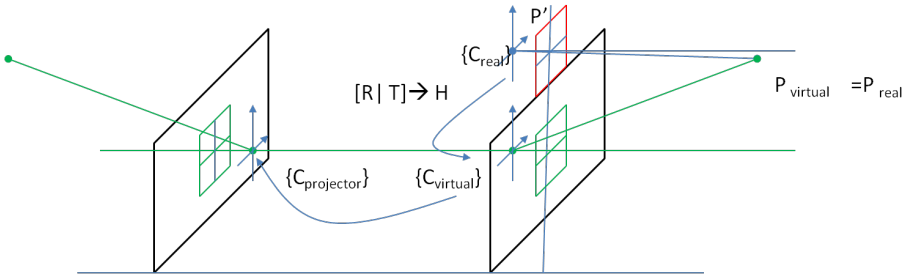


Figure 5.8: 2D conference with eye contact

5.2 Mesh Generation and Virtual View Synthesis

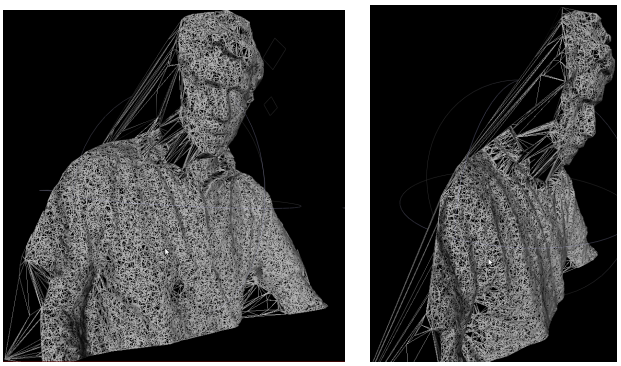


Figure 5.9: Mesh model using Crust triangulation

An incremental adaptation of the Crust algorithm is proposed and enables the addition of new 3D points without recomputing previously generated meshes. The stitching process [28] relies on integrating new mesh poles as new vertices on the triangulation step and computing triangles only where both surfaces share vertices.

Given a set o registered points $X \in R^3$ sampled from an object surface S, it is possible to approximate its shape by a triangle mesh. Based on a modified Crust algorithm, the approach uses a set of points P from the medial axis (polos) to extract a subset from the Delaunay triangulation of X that approximate S. The polo points, obtained from the Voronoi vertex or triangle's average outer normal's, are positive (p^+) if they lie on the convex side of the surface and negative (p^-) otherwise. Once the Delaunay triangulation of $X \cup P$ is computed, the surface mesh is estimated by extracting the set of simplices whose vertices belong to X. The proposed approach adds an incremental characteristic to the Crust algorithm as it is efficiently viable to add new vertices to a Delaunay triangulation.

Assuming that the Crust algorithm already processed a set of points X_t, the set of poles P_t and the Delaunay triangulation of are also available. To add a new set of sample points X_{t+1} to the mesh surface and to avoid a complete mesh recalculation, the following steps are performed:

1. P_{t+1}=poles of X_{t+1}

2. Add $P_{t+1} \cup X_{t+1}$ as new Delaunay triangulation vertices
3. Extract triangles whose vertices belong to $X_t \cup X_{t+1}$

The procedure can be applied repeatedly to accommodate any number of point sets X_i, nevertheless to avoid progressive growth in the number of mesh vertices, points closest to the mesh vertex, that is, under a given Euclidean distance threshold, are deleted from the input point cloud before the incremental Crust step. The surface mesh extraction process is described on algorithm 2 and generates results like in Fig. 5.9.

Algorithm 2 Surface mesh extraction algorithm

Input: X {3D points}
Output: 3D_triangle_mesh_surface_model
Compute Voronoi Diagram of X
for x_i = all points in X **do**
  $V_i =$ Voronoi cell containing x_i
  if x_i belongs to the convex hull **then**
    p^+ = average of the outer normals of triangles adjacent to x_i
  else
    p^+ = farthest Voronoi vertex from V_i
  end if
  p^- = farthest Voronoi vertex from V_i with negative projection on n^+
end for
$P =$ set containing all poles p^+ and p^-
D = Compute Delaunay triangulation of $P \cup X$
Return the subset of D containing only the triangles whose vertices are contained in X

In a 3D video conference, real eye contact is preserved while each participant observes the others from their current perspective. Users' viewpoints change according to their movements around the shared meeting environment. Therefore, new perspective's views must be presented at each instant depending on the viewer's pose in front of the display Fig. 5.1. This requires a precise estimation of the viewer's pose in 3D space, which a head/body tracking module can accomplish [519][437]. The selected approach is based on a facial feature tracker using eye feature [505]. The purpose of using Haar-like features is to meet the real-time requirement. The resulting 2D position of the eyes can then be associated with 3D points for calculating the 3D position of the head.

5.2.1 Multiview 3D Scan

Considering that from the sensor we obtain at instant t an image I^t and the scene depth D^t, we can consider that (after appropriate calibration) we have for each pixel $I^t(i,j)$ the corresponding depth information $D^t(i,j)$. From D^t we obtain a set of 3D points described in the sensor frame

$$\{\mathbf{x}_{ij}\} = f(i,j,D^t(i,j)), i = 1 \ldots W, j = 1 \ldots H$$

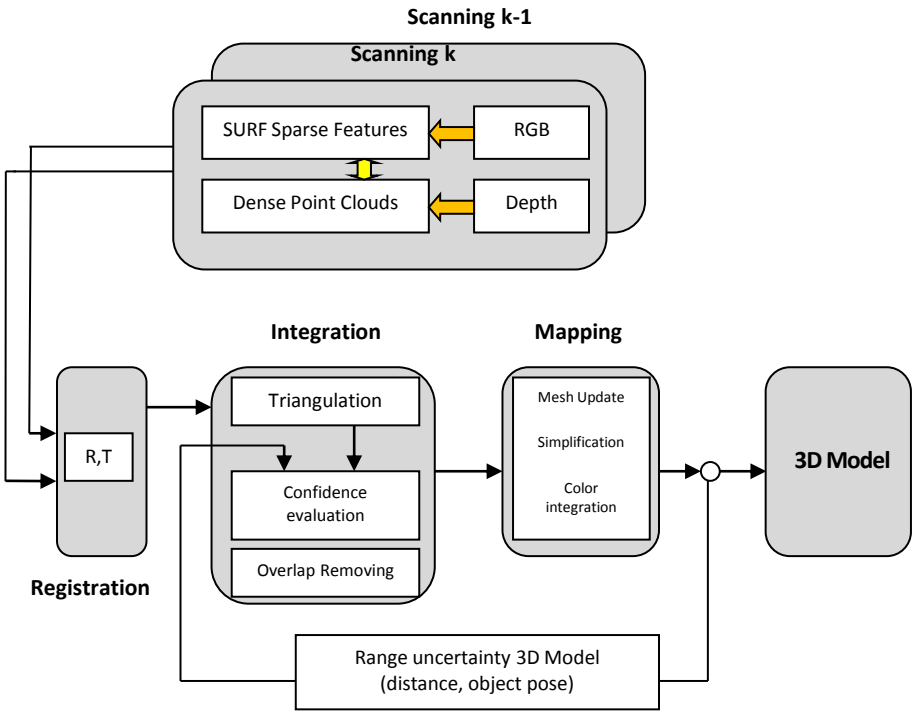


Figure 5.10: Algorithm overview modules

From the image I^t we extract a set of features

$\{S_k\}, S_k = (i_k, j_k, descriptor)$

where each feature is represented by its pixel coordinates and descriptors. We can then associate each feature with the corresponding 3D point. In other words we have 3D point coordinates, its image coordinate and the associated descriptor.

Denoting the 3D points as $\mathbf{P} = (x^t, y^t, z^t)$, the corresponding image point as $p = (x^t, y^t)$, we can establish the following tuples for the set of detected features $\{S'^t_k\}, k = 1 \dots N$,

$S'^t_k = (P^t_k, p^t_k, descriptor^t)$.

Considering that at a posterior acquisition we obtain another set

$\{S'^{t+1}_l\}, l = 1 \dots M$,

we expect to establish pairs between the elements of

$\{S'^t_k\}$ and those of $\{S'^{t+1}_l\}$.

Their association is made by comparing the descriptors between both sets.

Having established the correspondence we have to estimate its rigid body transformation.

5.2.2 Registration

Registration of a Segmented Body: considerer the motion of a rigid body in front of a scanner and the estimation of the rigid transformation (rotation and translation) where β_p refers to each different segmented body part referential. This information is important to register the body points on the same referential and create a global model.

Suppose the existence of two corresponding 3D points sets $\{^{\beta_p}\mathbf{x}^{\mathbf{t}}_{\mathbf{S_{ki}}}\}$ and $\{^{\beta_p}\mathbf{x}^{\mathbf{t+1}}_{\mathbf{S_{ki}}}\}$, $i = 1..N$, from consecutive t and $t+1$ scans, related through the following equation (5.24) :

$$^{\beta_p}\mathbf{x}^{\mathbf{t+1}}_{\mathbf{S_{ki}}} = ^{\beta_p}\mathbf{R}\,^{\beta_p}\mathbf{x}^{\mathbf{t}}_{\mathbf{S_{ki}}} + ^{\beta_p}\mathbf{t} + ^{\beta_p}\mathbf{v}_{\mathbf{S_{ki}}} \tag{5.24}$$

$$^{\beta_p}\varepsilon^2 = \sum_{i=1}^{N} \left\| \,^{\beta_p}\mathbf{x}^{\mathbf{t+1}}_{\mathbf{S_{ki}}} - \,^{\beta_p}\mathbf{R}\,^{\beta_p}\mathbf{x}^{\mathbf{t}}_{\mathbf{S_{ki}}} - \,^{\beta_p}\mathbf{t} \right\|^2 \tag{5.25}$$

$^{\beta_p}\mathbf{R}$ represents a standard 3x3 rotation matrix, $^{\beta_p}\mathbf{t}$ stands for a 3D translation vector and $^{\beta_p}\mathbf{v_i}$ is a noise vector. The optimal transformation $^{\beta_p}\mathrm{R}$ and $^{\beta_p}\mathrm{t}$ that maps the set $\{^{\beta_p}\mathbf{x}^{\mathbf{t}}_{\mathbf{S_{ki}}}\}$ on to $\{^{\beta_p}\mathbf{x}^{\mathbf{t+1}}_{\mathbf{S_{ki}}}\}$ can be obtained through the minimization of equation (5.25) using a least square criterion.

The least squares solution is the optimal transformation only if a correct correspondence between 3D point sets is guaranteed. The singular value decomposition (SVD) of a matrix is used to minimize Eq. (5.25) and obtain the rotation (standard orthonormal 3x3 matrix) and the translation (3D vector) [38][97][138]. In order to calculate rotation first, the least square solution requires that $\{^{\beta_p}\mathbf{x}^{\mathbf{t}}_{\mathbf{i}}\}$ and $\{^{\beta_p}\mathbf{x}^{\mathbf{t+1}}_{\mathbf{i}}\}$ point sets share a common centroid. With this constraint, a new of equation can be written using the following definitions:

$$\overline{^{\beta_p}\mathbf{x}^{\mathbf{t}}_{\mathbf{i}}} = \frac{1}{N}\sum_{i=0}^{n} {}^{\beta_p}\mathbf{x}^{\mathbf{t}}_{\mathbf{i}} \qquad \overline{^{\beta_p}\mathbf{x}^{\mathbf{t+1}}_{\mathbf{i}}} = \frac{1}{N}\sum_{i=0}^{n} {}^{\beta_p}\mathbf{x}^{\mathbf{t+1}}_{\mathbf{i}} \tag{5.26}$$

$$^{\beta_p}\mathbf{x}^{\mathbf{t}}_{\mathbf{ci}} = {}^{\beta_p}\mathbf{x}^{\mathbf{t}}_{\mathbf{i}} - \overline{^{\beta_p}\mathbf{x}^{\mathbf{t}}_{\mathbf{i}}} \qquad ^{\beta_p}\mathbf{x}^{\mathbf{t+1}}_{\mathbf{ci}} = {}^{\beta_p}\mathbf{x}^{\mathbf{t+1}}_{\mathbf{i}} - \overline{^{\beta_p}\mathbf{x}^{\mathbf{t+1}}_{\mathbf{i}}} \tag{5.27}$$

$$\varepsilon^2 = \sum_{i=1}^{N} \left\| \,^{\beta_p}\mathbf{x}^{\mathbf{t+1}}_{\mathbf{ci}} - \,^{\beta_p}\mathbf{R}\,^{\beta_p}\mathbf{x}^{\mathbf{t}}_{\mathbf{ci}} \right\|^2 \tag{5.28}$$

Maximizing $Trace(^{\beta_p}\mathbf{RH})$ enables us to minimize the generated equation (5.28), with **H** being a 3x3 correlation matrix defined by $\mathbf{H} = ^{\beta_p}\mathbf{x}^{\mathbf{t+1}}_{\mathbf{ci}}(^{\beta_p}\mathbf{x}^{\mathbf{t}}_{\mathbf{ci}})^{\mathbf{T}}$. Considering that the singular value decomposition of **H** results on **H=UDV**^{T}, then the optimal rotation matrix, $^{\beta_p}$**R**, that maximizes the referred trace is $^{\beta_p}$**R**= **U** diag(1; 1; det(**UV**^{T})) **V**^{T} [38][97][138] is given by equation (5.29) and the optimal translation that aligns $\{^{\beta_p}\mathbf{x}^{\mathbf{t+1}}_{\mathbf{i}}\}$ centroid with the rotated $\{^{\beta_p}\mathbf{x}^{\mathbf{t}}_{\mathbf{i}}\}$ centroid is given by equation (5.30).

$$^{\beta_P}\mathbf{R} = \mathbf{U}\mathbf{V}^{\mathbf{T}} \qquad (5.29) \qquad ^{\beta_P}\mathbf{t} = \overline{^{\beta_P}\mathbf{x}_\mathbf{i}^{\mathbf{t+1}}} - ^{\beta_P}\mathbf{R}\overline{^{\beta_P}\mathbf{x}_\mathbf{i}^{\mathbf{t}}} \qquad (5.30)$$

5.2.3 Model Mapping

Suppose that the mapping from the world coordinates to one of the scans of the sequence is known (ex: scan 0), and it is represented by the transformation $^{0}\mathbf{T}_\mathbf{w}$. As described before, for any consecutive pair of scans (t, t+1) from tracked points, it is possible to estimate rotation and translation and combine them into a single homogeneous matrix 4x4, $^{t+1}\mathbf{T}_\mathbf{t}$, $\mathbf{T} = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}$.

Therefore it is possible to compute equation (5.31):

$$^{i}\mathbf{T}_\mathbf{0} = {}^{i}\mathbf{T}_{\mathbf{i-1}}{}^{\mathbf{i-1}}\mathbf{T}_{\mathbf{i-2}} \ldots .{}^{1}\mathbf{T}_\mathbf{0} \qquad and \qquad {}^{i}\mathbf{T}_\mathbf{w} = {}^{i}\mathbf{T}_\mathbf{0}{}^{0}\mathbf{T}_\mathbf{w} \qquad (5.31)$$

To update the reconstructed model, each acquired 3D point set is transformed to the world coordinate system using $^{i}\mathbf{T}_\mathbf{w}$. This alignment step adds a new scan to the dense 3D model. Alignment between successive frames enables to track the body position over small displacements.

5.2.4 Tracking and Registration refining

SURF features are detected and matched over consecutive undistorted images. These features are invariant to affine transformations, so they allow the detection of the feature points from different angles and range. Although SURF provides good distinctive descriptors, undesirable matches related to static background areas and image body boundaries can occur. To overcome this situation, it is possible to define a working reconstruction space for the body and a mask for the SURF algorithm. After finding the set of matched image features, a correspondence between 2D and 3D is set up. These annotated 3D point pairs are then used to estimate the motion between two time consecutive point clouds. Assuming the identification problem is solved, we must compute the rigid transformation (rotation and translation) that aligns the two successive 3D scans. The solution should consider that the data are typically affected by noise: correspondences may be false, and some key data patches may be partially occluded.

Registration refining using RANSAC: False correspondent point pairs that wrongly bias the rigid body transformation estimation are removed using the RANSAC method [158]. The approach randomly samples three 3D points correspondent pairs from consecutive scans and iteratively estimates the rigid body transformation [38] until it finds enough consensus or reaches a maximum number of iterations based on the probability of outliers. The procedure uses a small initial data set and enlarges the number of samples consistent with the model. K iterations are performed while the eligible solution, with the highest number of inliers based on the sum of the distances between pairs of correspondent points, is selected

as the best transformation model. The K iterations number follows equation $K = \frac{log(1-p)}{log(1-(n_{inliers}/N_{pts})^S)}$, where p stands for the desired probability of finding at least one model transformation without outliers within K iteration [158], $n_{inliers}$ is the number of eligibles pairs of points that fit the current estimation, N_{pts} represents the total number of pairs of points and S is the minimum number of eligible samples to fit the transformation model. The registration refining method is described in algorithm 3.

Algorithm 3 Registration refining algorithm - outliers removal

1: $\mathbf{Input}$:X_p, X_q
{assumed correspondent 3D point pairs}
2: $\mathbf{Output}$:$[R, t]$
{rigid body transformation estimation}
3: **while** $(i < MAXITER)$ **do**
4: randomly select 3 pairs of points
5: $[R_i, t_i] \leftarrow$ estimate 6DOF rigid body transformation for these 3 pairs
6: $X'_q = R_i * X_q + t_i$
7: $inliers_i = |(X'_q - X_p) < \tau|, number\_of\_inliers_i$
8: **if** $(sizeof(inliers_i) > T_{threshold})$ **then**
9: $[R, t] \leftarrow$ re-estimate the transformation model using all $inliers_i$
10: $EXIT$
11: **end if**
12: **if** $(number\_of\_inliers_i > bestscore)$ **then**
13: $bestscore \leftarrow number\_of\_nliers_i$
14: $best\_inliers \leftarrow inliers_i$
15: update $MAXITER$ {using eq.of K iteration}
16: **end if**
17: $i = i + 1$
18: **end while**
19: $[\mathbf{R}, \mathbf{t}] \leftarrow$ re-estimate the transformation model using all points from $best\_inliers$

5.2.5 Global model reconstruction algorithm

The global model reconstruction algorithm can be described as follows algorithm 4:

The research aims to integrate newly acquired mesh into a reconstructed mesh model using range images. A moving object is presented to a fixed range's image data sensor providing depth images measured on the referential sensor. Different views of the object must be aligned in order to reconstruct a consistent surface that describes the underlying global object structure.

3D modeling consists mainly on four main phases:

1. Scan object surface from different view
2. Register the views

Algorithm 4 Model reconstruction algorithm

Input: rgb_images, depth_images
Output: 3D_mesh_model
initialize $[R_g, t_g]$, I^0, D^0, cams_calibration
$t = 1$
while (1) **do**
  $I^t =$ get rgb_image, $D^t =$ get depth_image
  $\mathbf{x}^t \leftarrow$ compute 3D points using $f(i, j, D^t(i, j))$
  $ID^t \leftarrow$ map rgbcolor to depth image using x^t and I^t
  $S_k^{'t} \leftarrow$ detect key point features on ID^t
  $S_l^{'t-1} \leftrightarrow S_k^{'t}$, consecutive frame feature match
  $\{\mathbf{x}_{S_l}^{t-1}, \mathbf{x}_{S_k}^{t}\} \leftarrow$ correspondence2D3D$(S_l^{'t-1} \leftrightarrow S_k^{'t})$
  $[R, T] \leftarrow$ solve $\{\mathbf{x}_{S_l}^{t-1} = R\mathbf{x}_{S_k}^{t} + T + V_{S_k}\}$
  $[R_g, T_g] \leftarrow$ update global transformation with $[R, T]$
  $\mathbf{x}_{S_l}^{t-1} \leftarrow \mathbf{x}_{S_k}^{t}$ keep data for future use
  project $\mathbf{x^t}$ points to base coordinates using $[R_g, T_g]$
  $M^t \leftarrow$ Crust mesh model generation of $\mathbf{x}^t$
  add partial mesh model to global model: $M = M + M^t$
  $t = t + 1$
end while

3. Integrate the views
4. Render the integrated data

5.2.6 Mesh Integration

Once registered two range image data aiming at a single surface, two situations can arise: *non-overlapped* region that contains new information for the 3D model or and *overlapped region* that might contain redundant data or confident data useful for the model refining. The acquired data must be evaluated, and in this case, the uncertainty of the range sensor is analyzed. Sensor accuracy measures depend on the incident angle between the measuring ray, and the surface [327]. Moreover, the confidence of the measured depth value is inversely proportional to the distance from where the sensor acquires the data and the angle of the line-of-sight and the surface normal. By associating the points in polygons, it is possible to attribute a confidence level to each triangle, for example. Later, this property enables one to select which triangles should be included in the global mesh model and which should be discarded. The segmentation process and the consequent triangle removal procedure lead to holes that must be reconnected to represent the 3D surface smoothly.

Overlapping segmentation, front face checking and matching

The overlapping region is determined by projecting the pre-built mesh vertices into the newly scanned sensor 2D plane before registration transformation and

by detecting overlapping triangles on the previous scanned range data image and the newly scanned range. We could re-triangulate all the points on the overlapping region, but misalignment errors can result in a bumpy surface. We take an intelligent approach where the update triangulations only happen where the associated entropy is high. The entropy associated with each triangle relates to the confidence measure for each 3D point (triangle vertices) [21]. The distance from where the sensor acquires the data and the angle that it stands from the surface is inversely proportional to the confidence:

$$C_i = |\frac{1}{L\theta}| \tag{5.32}$$

where L is the distance from the 3D point to the optical center to the corresponding range sensor (sensor pose when that point was acquired).

The angle θ between is given by

$$\theta = arcos(\overrightarrow{n_i}, \overrightarrow{r_i}) \tag{5.33}$$

where $\overrightarrow{n_i}$ is the normal of a triangle and $\overrightarrow{r_i}$ is the normalized measurement ray from the sensor's optical center to the point. The measures of confidence capture the fact that point close to the sensor and surfaces close to a front-parallel orientation is typically captured more accurately by range sensors. A point's normal vector consists of averaging a normal vector of triangles formed with pairs of neighbours. Figure 5.11 illustrates, through a 2D example, the confidence measure principle of a range sensor composed of 3-ray measure beams while scanning an object from different positions. In this case, the range sensor acquires data from 4 different points of view, S_0, S_1, S_2, S_3. For example, due to overlapping data measures, between S_0, S_1, S_3 we can incrementally update the global model with the more confident edges (ex: P_{30}, P_{31}, P_{32}).

The requirement for further 3D model refining is determined using a cost function based on entropy. For each new scanned 3D frame, a 3D list of triangles (faces) can be updated with entropy information related to 3D point positions.

Considering that the j^{th} face of a given 3D mesh is associated with the confidence value, $C_i = |\frac{1}{L\theta}|$, that is inversely proportional to the distance L and angle θ of data acquisition. The total confidence of the surface S can be calculated by the formula: $S = \sum_{j=0}^{m} C_j$.

If the confidence probability $(\frac{C_i}{S})$ of the j^{th} face is considered as its *information probability*, the *viewpoint entropy* E that estimates the quality [486] for a given $3D$ mesh model can be defined as:

$$E = -\sum_{j=0}^{m} \frac{C_j}{S} log_2 \frac{C_j}{S} \tag{5.34}$$

This definition is based on Shannon entropy [447] of a discrete variable X with

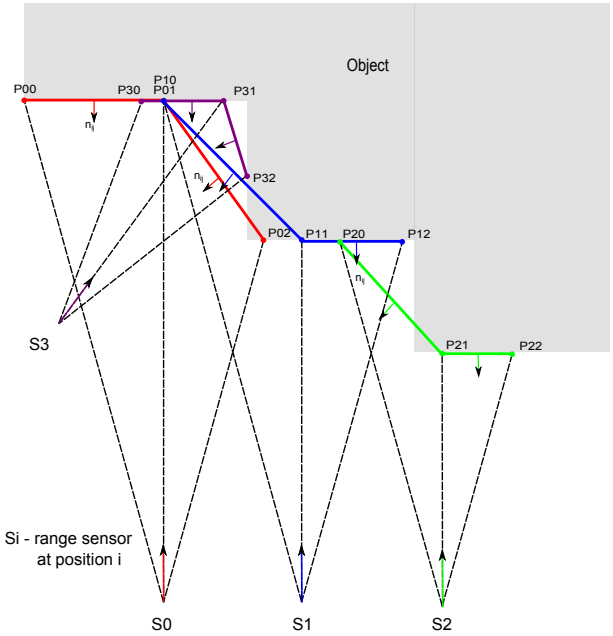


Figure 5.11: A range sensor, composed by 3 ray measure beams, scans an object from different positions (2D example)

values in the set $\{a_1, a_2, \ldots, a_n\}$ defined as $H(x) = -\sum_{i=i}^{n} p_i log(p_i)$, where $p_i = P_r[X = a_i]$ and the logarithm are in base 2, that is in "bits" of information. $-log(p_i)$ represents the information associated with the result a_i, and the entropy provides the average information or the uncertainty of a random variable.

The maximum entropy is obtained when all the faces have the same confidence probability. Since the maximum of formula (5.34) is $log_2(m+1)$, the normalized version results from dividing the formula by the maximum value:

$$E = -\frac{1}{log_2(m+1)} \sum_{j=0}^{m} \frac{C_j}{S} log_2 \frac{C_j}{S} \tag{5.35}$$

In summary, a newly acquired face triangle only integrates or replaces an existing one from the current model if it contributes to lowering the global model entropy.

3D Modeling

Initial triangulation: recent planar sensors range scanners, like Kinect or Swiss ranger, store depth measurements as a 2D grayscale image, being possible to recover the 3D coordinates of a real point knowing the calibration parameters. An initial triangulation is built considering four neighbouring points and six possible connections by exploiting the sensor's grid structure. Triangles are formed by connecting four neighbouring points only if they differ by a depth threshold. Fig. 5.12 presents the six possible combinations to generate triangles. It is important to preserve scene depth discontinuities. The diagonal edge results from the connections of opposite points with the smallest depth difference value. If the depth difference is too big, no connection is established, and the point is invalid.

Two or more invalid neighbouring points do not allow triangle creation. Due to the grid point organization, this mesh approximation method is faster than the Delaunay algorithm.

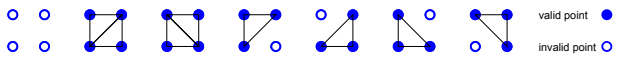


Figure 5.12: Initial mesh triangles: four point with six possible connections

Mesh simplification

A polygonal simplification algorithm is also used to reduce the number of faces used to represent a surface while preserving its structure, topology, shape, volume and boundaries. The approach described in Figure 5.13 enables the reduction of data information used to represent the surface, removing redundant vertices that do not contribute to defining the structure (ex: reducing the number of triangles used to describe large planar surfaces).

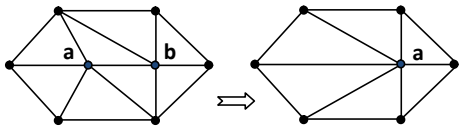


Figure 5.13: Edge collapse

The edge collapse algorithm cost function (5.36) measures the angle between two adjacent triangle faces:

$$cost(u,v) = \|a-b\| * \max_{f \in Ma}\{\min_{n \in Mab}\{\frac{1 - f.normal \bullet n.normal}{2}\}\} \tag{5.36}$$

where Ma is the mesh of triangles that contains vertice a, Mab is the mesh of triangles that have both a and b vertices, and $f.normal \bullet n.normal$ is the dot product between the normals of adjacent triangles that share edge $\overline{ab}$. The operation implements the steps described in algorithm5:

Algorithm 5 Edge collapse mesh simplification algorithm

1: **Input:** 3D_triangle_mesh_model
2: **Output:** Simplified 3D_triangle_mesh_model
3: **repeat**
4: evaluate edges cost
5: delete any triangle on the edge $\overline{ab}$
6: update remaining triangles that used to use a as vertice to use b instead
7: delete vertice a
8: **until** reach a target number of triangles for the model

5.2.7 Implementation and Results

The 3D body model reconstruction algorithm was experimentally tested, registering several 3D point clouds while a person was rotating in front of the Kinect device.

Calibration:

Such device combines a regular RGB camera and a 3D scanner, consisting of an infrared (IR) projector and an IR camera as shown in figure 5.14a). Due to manufacture differences, a calibration step [26] is performed to undistort the RGB and IR images and to map depth pixels with colour pixels (figures 5.14 a,b,c). The maximal range of the Kinect raw depth is 2^{11}, and it is possible to convert the raw depth to metric depth using a linear approximation after a previous depth calibration $d_m(x_{ir}, y_{ir}) = f(rawdepth(x_{ir}, y_{ir}))$.

From the metric depth, the 3D metric position (X_{ir}, Y_{ir}, Z_{ir}) of the pixel, with the respect to the IR camera, can be computed using the equation (5.58),

$$\begin{pmatrix} X_{ir} \\ Y_{ir} \\ Z_{ir} \end{pmatrix} = \begin{pmatrix} \frac{(x_{ir}-c_{xir})*d_m(x_{ir},y_{ir})}{f_{xir}} \\ \frac{(y_{ir}-c_{yir})*d_m(x_{ir},y_{ir})}{f_{yir}} \\ d_m(x_{ir}, y_{ir}) \end{pmatrix} \tag{5.37}$$

where x_{ir}, y_{ir} are the coordinates of the depth pixel in image, f_{xir},f_{yir} are the lengths in effective horizontal and vertical pixel size units (IR camera focal length), c_{xir}, c_{yir} are the coordinates of the image center of IR camera, and d_m is depth in meters.

A small baseline separates the IR and RGB cameras. And using chessboard target data and stereo calibration algorithms, it is possible to determine the 6 DOF transform between them. Knowing the rotation $\mathbf{R}$ and translation $\mathbf{T}$ between the RGB and IR camera, we can re-project each 3D point on the colour image and get its colour. The mapping between the colour image and depth image can be expressed by equations (5.38):

$$\begin{pmatrix} X_{rgb} \\ Y_{rgb} \\ Z_{rgb} \end{pmatrix} = \mathbf{R} \begin{pmatrix} X_{ir} \\ Y_{ir} \\ Z_{ir} \end{pmatrix} + \mathbf{T} \qquad \begin{aligned} x_{rgb} &= \frac{(X_{rgb}*f_{xrgb})}{Z_{rgb}} + c_{xrgb} \\ y_{rgb} &= \frac{(Y_{rgb}*f_{yrgb})}{Z_{rgb}} + c_{yrgb} \end{aligned} \tag{5.38}$$

where x_{rgb}, y_{rgb} are the coordinates of the rgb pixel in image, f_{xir}, f_{yir} are the lengths in effective horizontal and vertical pixel size units (RGB camera focal length), c_{xrgb}, c_{yrgb} are the coordinates of the image center of RGB camera, and d_m is depth in meters.

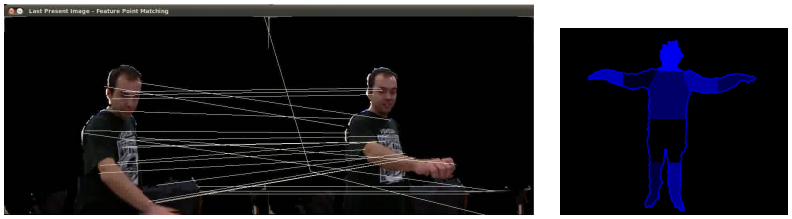
Matching and body segmentation

In figure 5.15(a), we present an example of correspondence between consecutive image features using the SURF method (white lines indicate correspondent point).

(a) (b) (c)

Figure 5.14: (a) undistorted RGB image (b) undistorted depth Image, the body white pixels have unknown depth value, due occlusions or reflective surface material (c) Map between undistorted RGB image and depth image.

Some matches are undesirable and are related to static background areas. Our solution is to confine the reconstruction space with precise limits or develop a movement segmentation filter. Erroneous match bias is minimized by the number of good matches, using the described minimization method with outliers removal to obtain the transformation. Due to the articulated nature of the human body, several parts suffer different motion transformations. We perform a body segmentation using the depth image and the OpenNI skeleton, applying various rigid body transformations during the movement. Each set of 3D body points is labelled with a tag, represented in figure 5.15(b) by levels of blue.



(a) (b)

Figure 5.15: (a) SURF features matched on consecutive time frames. (b) Body segmentations approach to address the articulate characteristic during motion.

Outliers removal: The registration refining improvement described on algorithm 3 was analyzed by measuring the mean euclidean distance between several consecutive registrations with and with outliers removed after applying the transformation to X_q scan to map it into X_p reference frame ($X'_q = R_i * X_q + t_i$). The red balls line (without outliers), in Fig. 5.16, presents a much lower error than considering all SURF-matched points into rigid body transformation. Fig. 5.17, presents for each consecutive rigid body transformation estimation the total number of SURF matched points (blue bars) and the number of inliers for that take (red bars).

Experimental results show that considering many inliers (not all SURF point features) makes the transformation estimation more robust and increases the alignment accuracy. Figure 5.18 shows pairs of consecutive images and their points' correspondence.

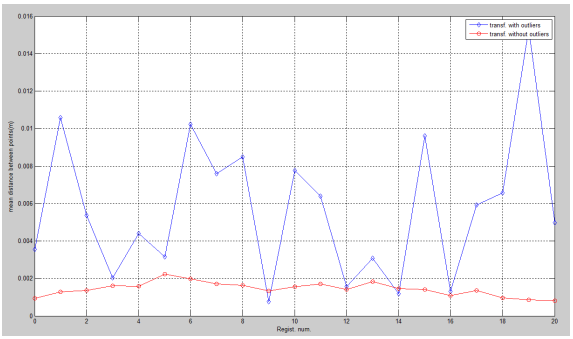


Figure 5.16: Mean euclidean distance between pair of corresponding points on each alignment take with and without outliers removed (in red and in blue respectively).

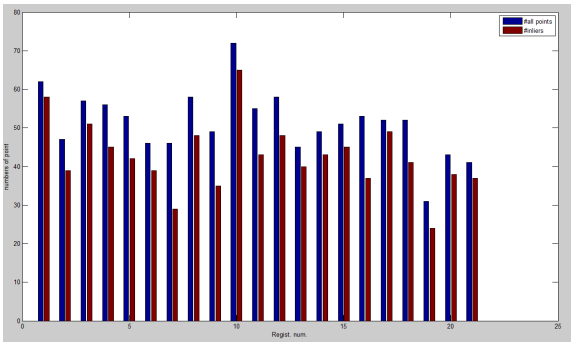


Figure 5.17: Number of points number (blue bars) vs Number of inlier's (red bars) on each registration.

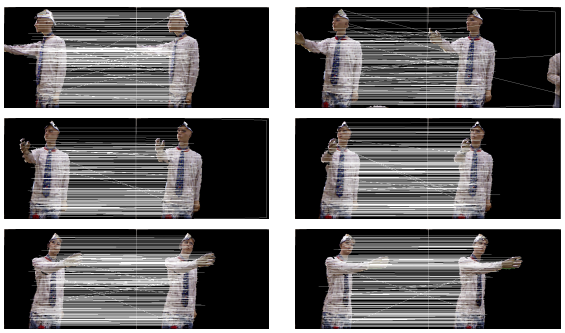


Figure 5.18: Pair of consecutive images displaying correspondent SURF point features later annotated with their 3D position and used to create a global 3D model

3D Modeling

Figure 5.21 shows a sequence of mesh models to be integrated into the model. Based on the depth data sensor grid structure, the initial triangulation is generated in real-time, using GPU resources to speed up the process.

Figure 5.19 depicts a sequence of scans that creates a 3D person model. They result from several 3D point clouds fused in real time after applying successive 3D rigid body transformations.

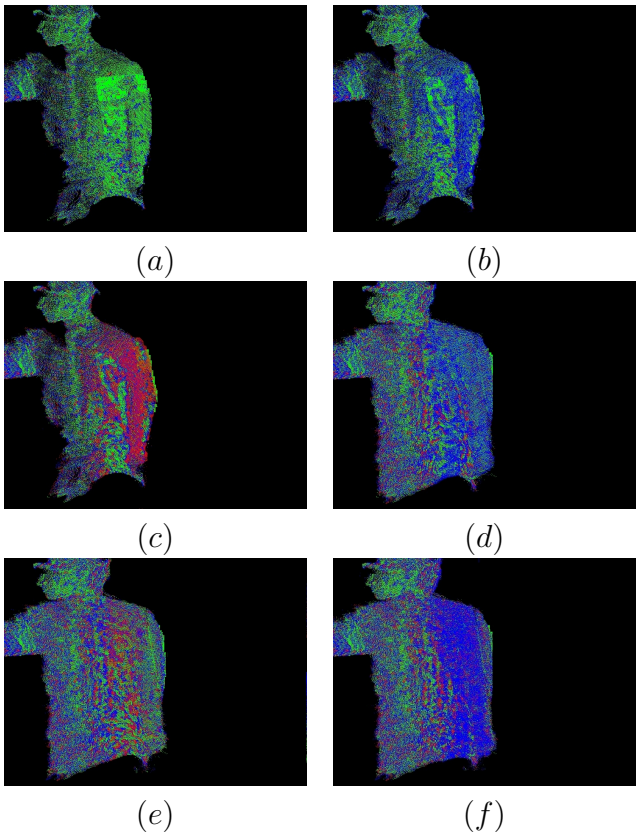
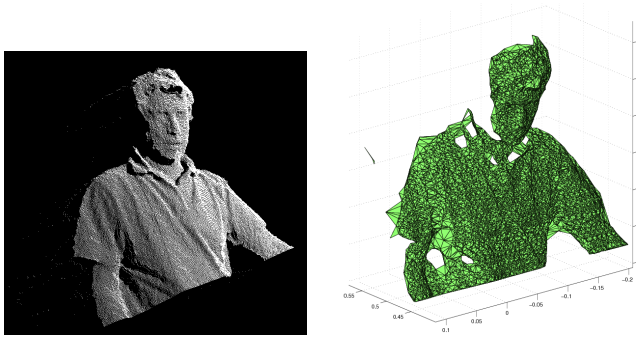


Figure 5.19: 3D Model, real time sequence of point clouds (a) .. (f) being registered on the same referential, each color represents time sequential scans

An example of off-line mesh generation, using unorganized Kinect 3D points, is provided in Figure 5.20(a). Delaunay triangulation computation results on 99334 vertices, and 1223930 faces and an example of mesh generation using the proposed incremental adaptation of Crust algorithm is provided in Figure 5.20(b) with 27864 vertices and 31810 faces (took 9617msec to process).

Processing time measurements: The system performance is about 2 HZ (C++ implementation on a Core 2 Duo CPU E8200). The time-consuming stage is related to the surf feature extraction, which takes an average of 300 ms. It depends on the number of detected good features of the image. GPU used in this step will improve the speed. The involved number of points also influences the transformation time calculus. Table 5.1 presents specific time measures involving



(a) (b)

Figure 5.20: a) Mesh model using Delaunay triangulation results on 1223930 faces and 99334 vertices (b) Mesh model with 27864 vertices and 31810 faces using the proposed incremental adaptation of Crust algorithm

Table 5.1: Processing time measurements

Algorithm Steps	**(ms)**
Acquisition	1.55
Undistort Images	10.61
DepthRGB Map and last frame update	36.13
SURF feature extraction	314.853
Matching and transformation calculus	78.0282
Alignment, display and interaction	30.377
Total	471.56 (f=2.12 Hz)

algorithm steps.

Figures 5.21 and 5.22 depicts the sequence of mesh models to be integrated, in which mesh triangulation was based on depth data sensor grid structure.

Figure 5.23 illustrates synthesized views of a on-line 3D reconstructed model dependent of observer point of view.

5.2.8 Summary

A free viewpoint system framework is proposed to generate view dependent synthesis based on scene 3D mesh model. Our approach explores virtual view synthesis through motion body estimation and hybrid sensors composed by video cameras and a low cost depth camera based on structured-light. The solution addresses the geometry reconstruction challenge from traditional video cameras array, that is, the lack of accuracy in low-texture or repeated pattern regions. We present a full 3D body reconstruction system that combines visual features and shape-based alignment. Although SIFT has better accuracy as key feature descriptor, we have chosen SURF method to achieve the real-time characteristic. Experimental results shows that considering a high number of inliers (not all SURF point features) increases the alignment accuracy. Modeling is based on meshes computed from

dense depth maps in order lower the data to be processed and create a 3D mesh representation that is independent of view-point. Research contributions includes a new incremental version of Crust algorithm that efficiently adds new vertices to an already existing surface without having to recompute previous generated meshes, an entropy topological incremental reconstruction approach based on confidence measures that avoids redundant data information computation. This work presents an on-line incremental 3D reconstruction framework that can be used on low cost telepresence applications or human robot interaction applications. Nevertheless results showed that the RGB-D sensors are noisy, suggesting a deeper study regarding sensor's error modeling. This study is presented in following section.

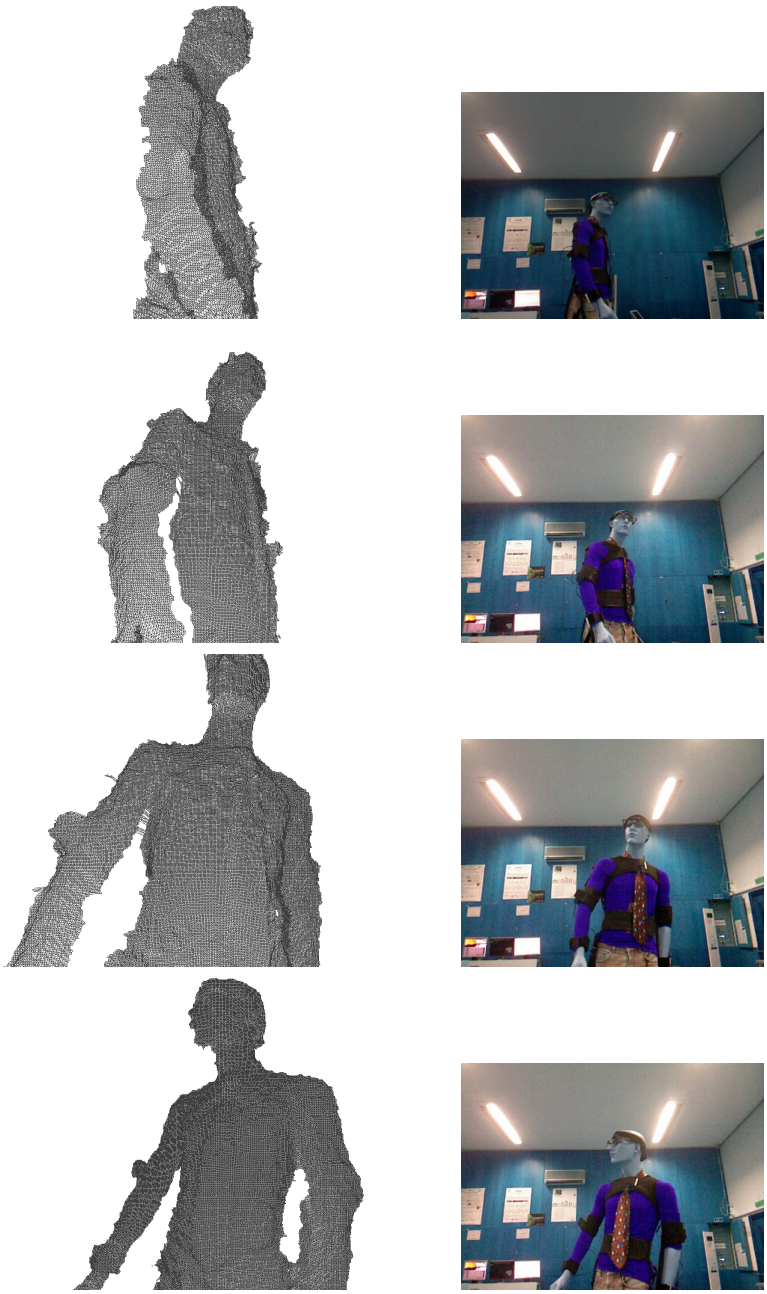


Figure 5.21: Sequence of mesh models to be integrated, triangulation based on depth data sensor grid structure

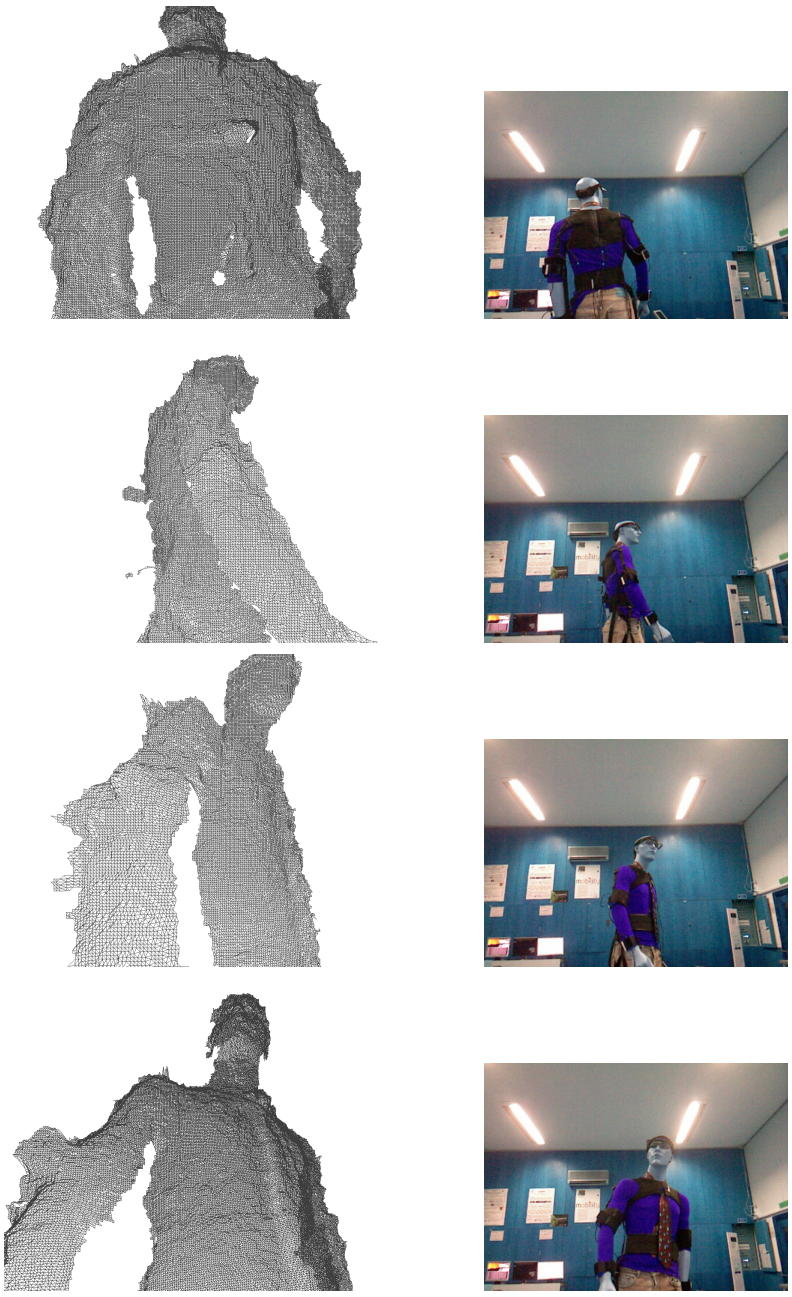


Figure 5.22: Sequence of mesh models to be integrated, triangulation based on depth data sensor grid structure

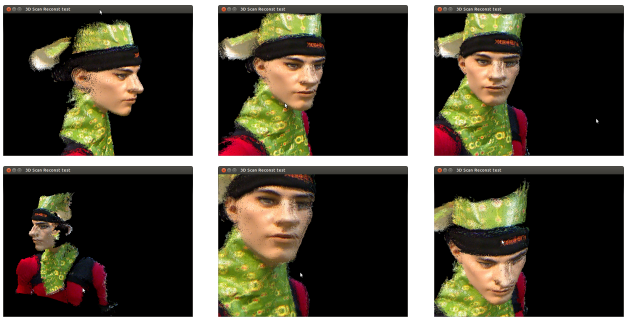


Figure 5.23: Synthesized views of a on-line 3D reconstructed model dependent of observer point of view.

5.3 Kinect accuracy and error analysis

Kinect depth maps are noisily unstable and may compromise the accuracy and precision of applications that depend on them. This chapter aims to characterize the Kinect depth map noise, identify its causes, and propose methodologies to minimize such noise through filtering or calibration techniques selectively. Structural analysis shows that the depth accuracy and the depth resolution decrease quadratically as the distance from the sensor increases. Depth maps may contain holes or inconsistent data in the depth of image object boundaries. The main problem, however, is that the depth measurement for a particular pixel often varies along the time for a static scene, while a neighbouring pixel can exhibit a different vibrating behaviour. Such limitations have motivated our research for fast and efficient algorithms that fills small depth data gaps and smooths the depth maps without introducing new values on large inexistent surfaces. Enhancing the input data quality is crucial for 3D reconstructions.

The Xbox 360® Kinect™ Sensor is a motion-sensing device that enables users to interact with the game console through gestures in a contactless way. The sensor enables RGB, infrared (IR), depth maps, skeleton and audio streams at a low cost. Due to its functionalities, it has been widely used for research applications in computer vision, 3D reconstruction, robotics, human-computer interaction (HCI) and augmented reality (AR).

The Xbox 360® Kinect™ Sensor combines an RGB camera and a structured light 3D scanner, consisting of an infrared camera and an infrared (IR) laser projector. The depth measurement principle is based on a triangulation process [165]. A single IR laser beam is split into thousands of beams by a diffraction grating, projecting a constant pattern of speckles onto the objects. The reflected pattern is re-captured by the IR camera, and it is correlated against a built-in reference pattern. Stored in the device memory, this factory reference pattern results from acquiring a plane at a specific distance. Speckles projected onto the objects, further or near the reference plane, will shift their infrared image position along the baseline direction. The baseline is the line between the infrared camera perspective center and the laser projector center. An image correlation procedure is applied to all speckles to measure the re-projected shifts leading to a disparity image. The Kinect sensor can acquire RGB and depth data up to 30 fps, using an internal processor to compute the disparity image and encoding the depth in 11 bits (0 - 2047 values), see Figure 5.24. The default RGB video and depth stream are 640 x 480 pixels at 30 fps. However, the RGB camera sensor array and the IR monochrome camera sensor array have resolutions of 1280 x 1024 pixels. USB communications constraints have limited RGB-D data transmissions. RGB camera can operate at lower rates with higher resolutions (1280 x 1024 pixels at 10 fps). The sensor optics has an angular field of view of 57° horizontally and 43° vertically.

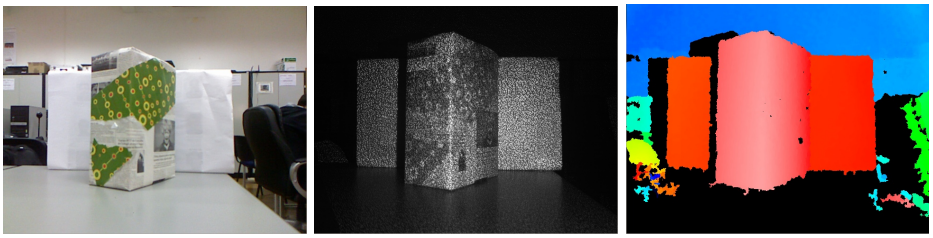


Figure 5.24: (a) RGB image, (b) IR monochromatic image with speckles pattern projected onto a scene, (c) Depth map with distances associated to colors

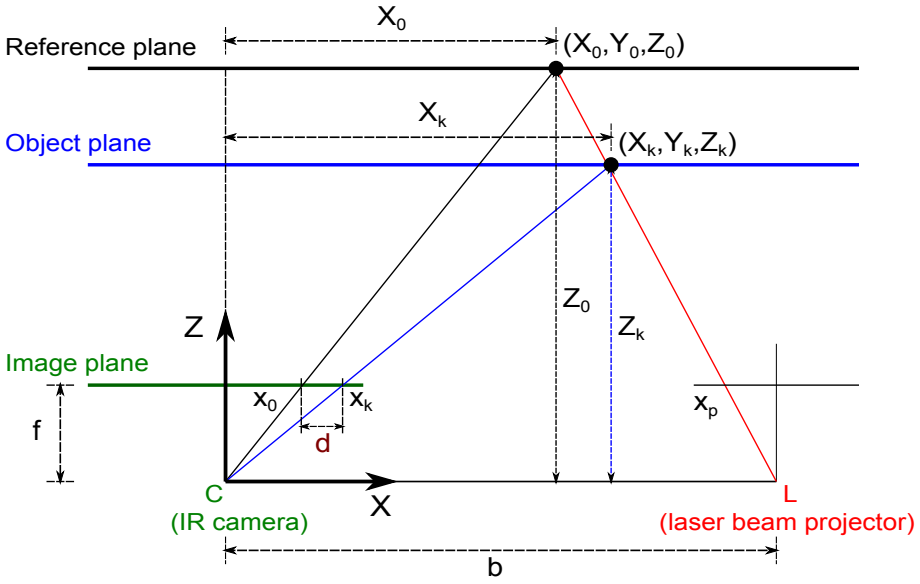


Figure 5.25: Kinect geometry that relates relative depth with disparity

5.3.1 Geometric disparity depth model

Figure 5.25 depicts the mathematical model geometry that relates the distance of an object point (X_k, Y_k, Z_k) to the kinect sensor, with the disparity d and a reference plane at depth Z_0 (top view model). Let us consider the 3D system referential with its origin at the projection center of the IR camera, C. The Z axis is coincident with the IR camera optical axis, which towards to the scene and is perpendicular to the image plane. The X axis is perpendicular to the Z axis and aligned with the baseline, b. The baseline connects the IR camera perspective center C, with the laser projector perspective center, L. The Y axis is orthogonal to Z and X axes and toward up from the draw.

How the system determines the position in space (X_k, Y_k, Z_k) is triangulation, that is, intersecting the ray defined by the IR camera center of projection C and the point image x_k, with the ray (laser beam) defined by the laser center of projection L and its virtual point image x_p (assumed projector optics similar to a camera, with both optics axes aligned and front-parallel). Consider that this object point has image coordinates (x'_k, y'_k) and (x'_p, y'_p) in the camera and projector image planes, respectively. Let f be the focal length of the camera and projector, the perpendicular distance between the lens center of projection and the image plane. Depth Z_k, is the distance between the 3D object point and the baseline.

Considering the pin-hole camera model and the similarity of triangles, we obtain the following expressions:

$$\frac{x'_k}{f} = \frac{X_k}{Z_k} \tag{5.39}$$

$$\frac{x'_p}{f} = \frac{X_k - b}{Z_k} \tag{5.40}$$

$$\frac{y'_k}{f} = \frac{y'_p}{f} = \frac{Y_k}{Z_k} \tag{5.41}$$

Solving the equation system and expressing (X_k, Y_k, Z_k) independently gives:

$$(X_k, Y_k, Z_k) = (\frac{x'_k b}{(x'_k - x'_p)}, \frac{y'_k b}{(x'_k - x'_p)}, \frac{bf}{(x'_k - x'_p)}) \tag{5.42}$$

where the $(x'_k - x'_p)$ quantity is the horizontal disparity.

Consider an object point (X_0, Y_0, Z_0) in three-dimensional coordinates on the reference plane at depth Z_0, and its speckle projected into the image plane of the IR camera, x_0. For the same laser beam, when the object gets closer to the sensor, say (X_k, Y_k, Z_k), its projected speckle x_k will be shifted in the direction of X axis. From (5.42), any K object point has its depth given by Z_k=$\frac{bf}{(x'_k - x'_p)}$, while a point on the reference plane has a depth given by Z_0=$\frac{bf}{(x'_0 - x'_p)}$. A depth shift relative to the reference plane can be described by the difference:

$$Z_k - Z_0 = \frac{bf}{(x'_k - x'_p)} - \frac{bf}{(x'_0 - x'_p)} \tag{5.43}$$

$$Z_k = \frac{Z_0}{1 + \frac{Z_0}{bf}(x'_k - x'_0)} \tag{5.44}$$

Expressing Z_k in terms of the remaining variables and having as disparity, $d = (x'_k - x'_0)$, enables equation (5.45) to establish the mathematical model that relates depth with the measured disparity:

$$Z_k = Z(d) = \frac{Z_0}{1 + \frac{Z_0}{bf}(d)} = \frac{1}{\frac{1}{Z_0} + \frac{1}{bf}(d)} \tag{5.45}$$

depth for a specific pixel is inversely scaled to the measured disparity, where the parameters $(\frac{1}{Z_0})$ and $(\frac{1}{bf})$ can be inferred through a depth camera calibration process while using a least square fitting approach.

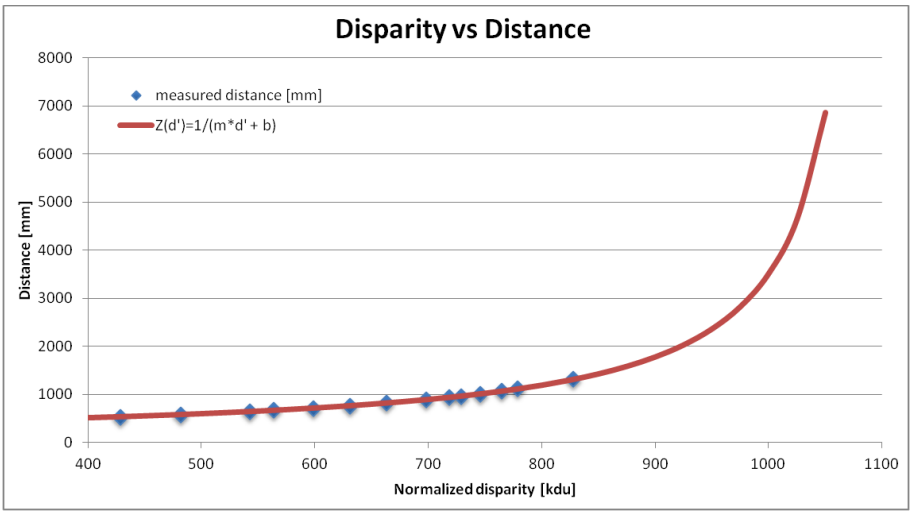


Figure 5.26: Relation between normalized disparity and the real depth distance (blue square markers), mathematical depth model (eq. 5.46) relating the normalized disparity with the depth measured data (red line)

Standard stereo camera calibration methods can be used to determine intrinsic and extrinsic parameters from RGB camera, IR camera and the transformation between both, i.e. focal length, coordinates of the image center sensors image, baseline (b) and depth of the reference plane (Z_0). Focal lengths and image centers are intrinsic parameters easily determined; however, the baseline b and the reference plane (Z_0) depth are harder to compute separately. Kinect raw disparity image output is normalized and quantized, ranging integers from 0 to 2047, using 11 bits to encode such values [250].

Due to this fact, the d variable from equation (5.45), should be expressed by $md'+n$, where d' stands for normalized disparity and m,n are the linear normalization coefficients.

$$Z(d') = \frac{1}{\frac{1}{Z_0} + \frac{1}{bf}(md' + n)} = \frac{1}{(\frac{m}{bf})d' + (\frac{1}{Z_0} + \frac{n}{bf})} \tag{5.46}$$

Once again, $(\frac{m}{bf})$ and $(\frac{1}{Z_0} + \frac{n}{bf})$ can be estimated through a depth calibration process, where normalized disparity d' is measured for several known depth positions of a plane. A least square fitting approach for $Z_k^{-1} = (\frac{m}{bf})d' + ({Z_0}^{-1} + \frac{n}{bf})$ enables to determine those normalization parameters, but Z_0 and b cannot be isolated.

The disparity depth mathematical model (eq. (5.46) is depicted in figure 5.26 (red line). The real depth distances (metric ruler measured) were plotted against the normalized disparity measured by the sensor (blue square markers), while the red line shows the fitting results of the proposed depth calibration function.

The parameters of disparity vs depth model (5.46) were obtained through a calibration process with a plane perpendicular to Kinect Z axis, was positioned at

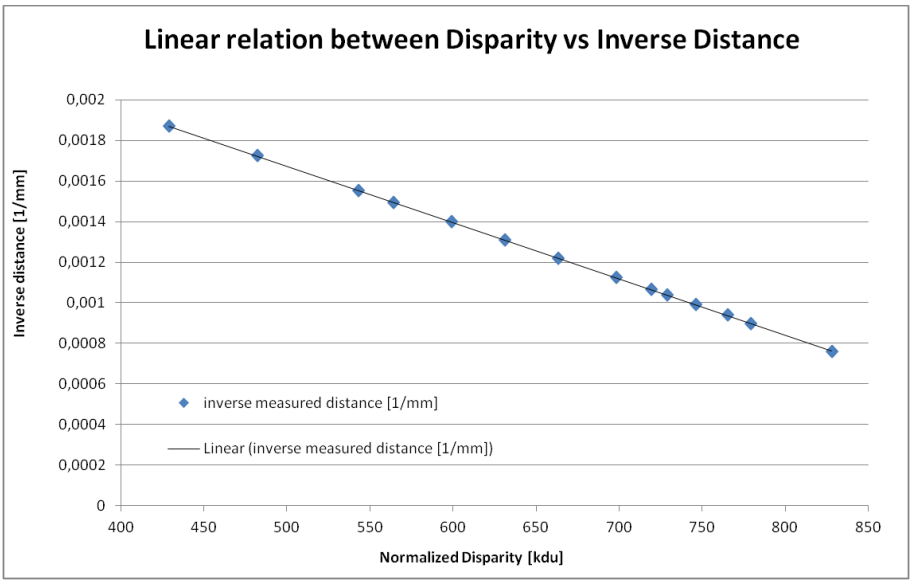


Figure 5.27: Linear relation of normalized disparity with inverse depth distance

fourteen different distances, and the respective real inverse depth distances were manually recorded against the sensor normalized disparity readings (figure 5.27). A simple least square line regression enabled to determine $(\frac{m}{bf})$ =-2.77527E-06 as the slope and $(\frac{1}{Z_0} + \frac{n}{bf})$ =0.003059 as the YY axis intersection with the line approximation.

5.3.2 Modeling Sensor Errors

Khoshelham and Elberink [250] refer that errors and artefacts in Kinect depth data may arise from three primary sources: *the sensor*, *measurement setup*, and *reflectance of the object surface*.

The *sensor errors* are mainly related to improper calibration and incorrect disparity measurements. Improper calibrations and erroneous estimation of calibration parameters cause systematic errors in point object coordinate determination. Adequate and precise calibrations are then required. Inexact disparity measurement due to the correlation algorithm and disparity quantization errors also negatively affect the accuracy of a point object coordinates.

The *measurement setup errors* are often related to the scene illumination conditions and the system imaging geometry. Strong incident light reduces the infrared image contrast generated by the projected laser speckles and negatively influences the correlation procedures and disparity estimation. It can cause gaps in the resultant point clouds or even outliers. The system imaging geometry introduces structural errors that are a function of the distance to the object and the sensor orientations relative to the object's surface. As demonstrated in the next section, the random error of depth measurement increases as the object distance increases. According to the official specs [340], the Kinect for Windows sensor has two depth operating ranges: the *default range* mode and the *close range* mode, while the Kinect for Xbox 360 sensor has just the default mode. Default mode

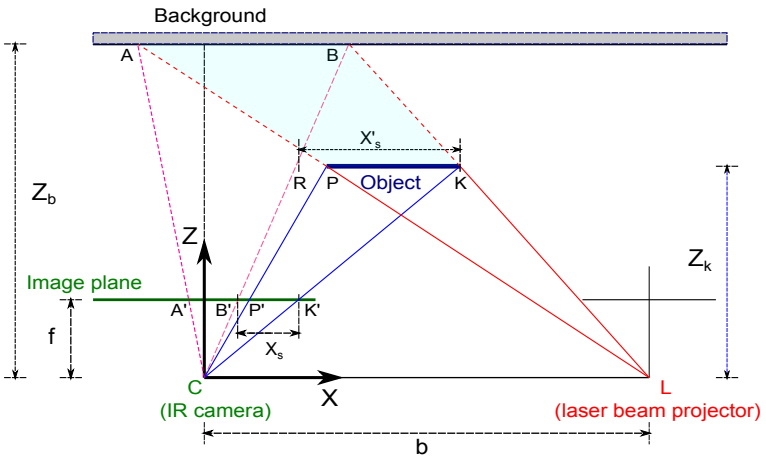


Figure 5.28: Kinect shadow model

ranges from 0.8m to 4.0m and near mode ranges from 0.4m to 3.0m. In practice, tracking operations are possible through an extended range of approximately 0.5m to 6.0m.

The system imaging geometry can also contribute to scene part occlusions and shadows. The system cannot estimate the depth when an object obstructs the projected IR speckles. As illustrated in figure 5.28, the laser speckles projector L irradiates onto the scene. As the region AB, behind the object PK, do not get any IR speckle, its respective image $A'B'$ will not be infrared illuminated, thus resulting in a shadow. Based on pin-hole camera model (5.39) and (5.41) and triangles similarity, ${X'}_s/Z_k = X_s/f$ and ${X'}_s/b = (Z_b - Z_k)/Z_b$. The shadow offset X_s yields:

$$X_s = bf(\frac{1}{Z_k} - \frac{1}{Z_b}) \tag{5.47}$$

Depending on the object's position, the shadow can be more or less noticeable. When the object is further away from the sensor, the beginning of the image shadow B' may become hidden by the object, reducing the shadowing area. However, the shadow image A' limit will be present.

The *reflectance of the object surface* also influences the determination of disparity measurements and point coordinates. A shiny, smooth surface that looks over-exposed in the IR image does not enable correlations and disparity estimations, originating gaps in the resultant point clouds.

Until recently, most of the Kinect sensor analyses were carried out by independent researchers, mainly through re-engineering, due the device's proprietary technology. Zhengyou Zhang [535], known for his work on camera calibrations and a principal researcher for Microsoft Research, refer that depth values produced by the Kinect sensor are sometimes "inaccurate because the calibration between the IR projector and the IR camera becomes invalid". The cause can be related to the heat, the vibration during transportation, or drift in the IR laser [536]. To solve this problem, they developed a simple recalibration technique based on a card with circles that are shipped with the Kinect sensor. For a high-precision calibration

process, a bundle adjustment of depth and colour camera is required, as the one proposed by Herrera and Heikkila [207][412][104][531].

Several concept definitions are used for the sensor error understanding: accuracy, resolution and precision.

Depth accuracy indicates how close the measurement's real depth value is.

Depth resolution is the minimum difference between two depth values related to the depth quantization step size.

Depth precision indicates how close the depth measurements are between them for independent observations of a depth value.

Depth accuracy

When derivate, the theoretical model that relates depth Z with disparity d, that is, equation (5.45), enables determining how a variation in the computed disparity affects the depth.

$$\frac{\partial}{\partial d}\left(\frac{1}{\frac{1}{Z_0}+\frac{1}{bf}d}\right)=\frac{-\frac{1}{bf}}{\left(\frac{1}{Z_0}+\frac{1}{bf}d\right)^2}=(-\frac{1}{bf})Z^2 \tag{5.48}$$

or applying the derivate to equation (5.46)

$$\frac{\partial}{\partial d'}\left(\frac{1}{(\frac{1}{Z_0}+\frac{n}{bf})+(\frac{m}{bf})d'}\right)=\frac{-\frac{m}{bf}}{\left((\frac{1}{Z_0}+\frac{n}{bf})+(\frac{m}{bf})d'\right)^2} \tag{5.49}$$

The depth error increases with the squared distance: $(\frac{m}{bf})Z^2$

Error propagation The uncertainty of the measured disparity affects the calculated depth. Considering the normalized disparity d' as a random variable that follows a normal distribution and all other calibration parameters as constants, the variance of the normalized disparity variable $\sigma_{d'}^2$ propagates to the variance of depth variable σ_Z^2, according $\sigma_Z^2=(\frac{\partial Z}{\partial d})^2\sigma_{d'}^2$. Expressing the standard deviation of computed depth σ_Z, with respect to the standard deviation of the normalized disparity $\sigma_{d'}$, results in:

$$\sigma_Z=(\frac{m}{bf})Z^2\sigma_{d'} \tag{5.50}$$

According the pin-hole camera model (5.39) and (5.41), $X_k = \frac{x'_k}{f} Z_k$ and $Y_k = \frac{y'_k}{f} Z_k$, meaning that depth error Z propagates to the X and Y coordinates calculus.

Considering that intrinsic calibration parameters are constant, determined accurately and neglecting the random errors of image coordinates x_k and y_k, results in the following expressions for the random error of X and Y coordinates.

$$\sigma_x = (\frac{mx'_k}{bf^2}) Z^2 \sigma_{d'} \qquad \sigma_y = (\frac{my'_k}{bf^2}) Z^2 \sigma_{d'} \tag{5.51}$$

Depth resolution

The minimum measurable depth difference between two successive levels of disparity determines the depth resolution. Kinect encodes the disparity measurement value using 11 bits to express integers from 0 to 2047. From equation (5.46), $Z(d')$ is a function that relates depth with the normalized disparity d', while the depth difference between two successive disparities levels is given by:

$$\Delta_Z(d') = Z(d') - Z(d'-1) \approx (\frac{\partial Z}{\partial d}) = (\frac{m}{bf}) Z^2 \tag{5.52}$$

To sum up, the depth accuracy and resolution decrease quadratically as the distance from the sensor increases.

Figure 5.29 depicts the depth resolution (blue curve) and theoretical random error (red curve). The depth resolution curve (blue) results from equation (5.52) plotting. The theoretical random error results from equation (5.50) evaluation, assuming a maximum disparity measurement error of half a pixel ($\sigma_{d'} = \frac{1}{2}$).

Spatial X/Y Resolution

for a given depth, the X/Y image resolution refers to the length along the X or Y axis covered per pixel (mm/pixel). Kinect for Xbox 360 throughputs a depth image containing 640 pixels along X axis and 480 pixels along Y axis, with a horizontal angle of field of view of 57° and a vertical angle of 43° [341][408].

At a given distance Z, the spatial $X_{resolution}(Z) = (2 * Z * tan(\frac{57^\circ}{2}))/640$ (mm/pixel) and the spatial $Y_{resolution}(Z) = (2 * Z * tan(\frac{43^\circ}{2}))/480$ (mm/pixel). For example, at 1-meter distance the $X_{resolution}(1000) = 1.70$ mm/pixel and the $Y_{resolution}(1000) = 1.64$ mm/pixel.

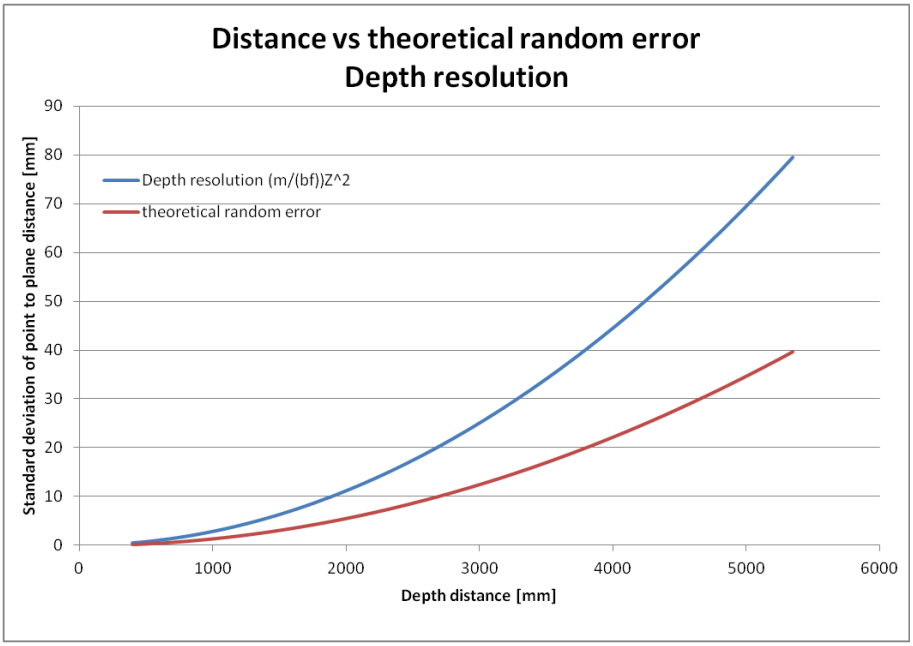


Figure 5.29: Depth resolution (blue) and theoretical random error (red)

5.3.3 Error Analysis Statistics

Nguyen and Izadi [362] have derived an empirical model for the Kinect sensor noise to extend the successful real-time 3D reconstruction and tracking application, KinectFusion [361]. They systematically measured axial and lateral noise distribution as a function of sensor distance and observation angle to a planar target. They have proposed a 3D noise model distribution for the depth measurements in terms of *axial noise* (z-direction) and *lateral noise* (z-perpendicular directions), as depicted in Figure 5.30.

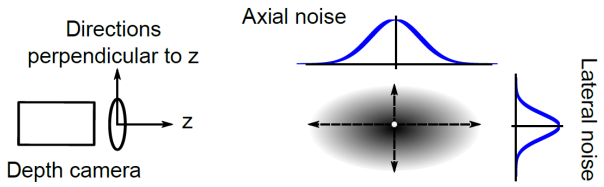


Figure 5.30: 3D noise model distribution for the depth measurements in terms of *axial noise* (z-direction) and *lateral noise* (z-perpendicular directions)

The authors found little changes regarding the lateral noise as a function of Z-distance.

The lateral noise model was fitted by a linear plus hyperbolic curve following the next equations, in which equation (5.53) gives the standard deviation distribution in pixels, and equation (5.54) converts this to real-world units (meters):

$$\sigma_L(\theta)[px] = 0.8 + 0.035 * \frac{\theta}{\pi/2 - \theta} \tag{5.53}$$

$$\sigma_L(\theta)[m] = \sigma_L(\theta)[px] * z * p_x/f_x \tag{5.54}$$

The axial noise model was fitted assuming constant Z-axial noise for angles between 10^o and 60^0

$$\sigma_L(z,\theta)) = 0.0012 + 0.0019 * (z - 0.4)^2, 10^o \leq \theta \leq 60^o \tag{5.55}$$

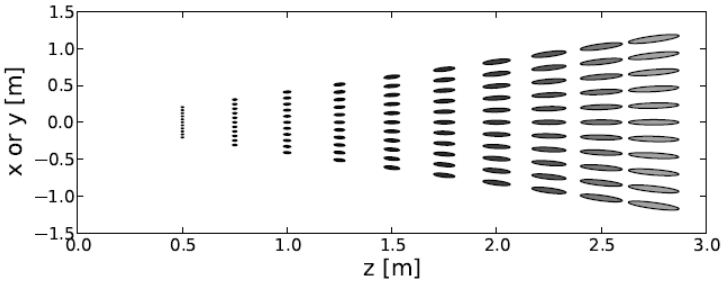


Figure 5.31: Top view visualization of the PDF contours of Kinect sensor noise distributions in 3D space [362]. Each ellipse represents the noise distribution with σ_Z and σ_L scaled up by a factor of 20.

5.3.4 Calibrations

Due to manufacture differences, calibration is usually required to undistort the RGB and IR images and to map depth pixels with colour pixels. Alternatively, such mapping can be performed by the Kinect driver [382][381][340], making use of a built-in lookup table that maps depth in each RGB image pixel in millimetres. The maximal range of the Kinect raw depth is 2^{11}, and it is possible to convert the raw depth to metric depth using a linear approximation after a previous depth calibration using equation (5.45) as depth expressed in terms of disparity.

Color camera intrinsics

The projection of a point from color camera coordinates $x_c = [x_c; y_c; z_c]^T$ to color image coordinates $p_c = [u_c; v_c]^T$ is obtained through the following equations. The point is first normalized by $x_n = [x_n; y_n]^T = [x_c/z_c; y_c/z_c]$. Distortion is then performed:

$$x_g = \begin{pmatrix} 2k_3 x_n y_n + k_4(r^2 + 2{x_n}^2) \\ k_3(r2 + 2y2n) + 2k_4 x_n y_n \end{pmatrix} \tag{5.56}$$

$$x_k = (1 + k_1 r^2 + k_2 r^4 + k_5 r^6)x_n + x_g \tag{5.57}$$

where $r_2 = {x_n}^2 + {y_n}^2$ and $k_c = [k_1, ..., k_5]$ is a vector containing the distortion coefficients.

Figure 5.32 illustrates the reference frames and transformations [207].

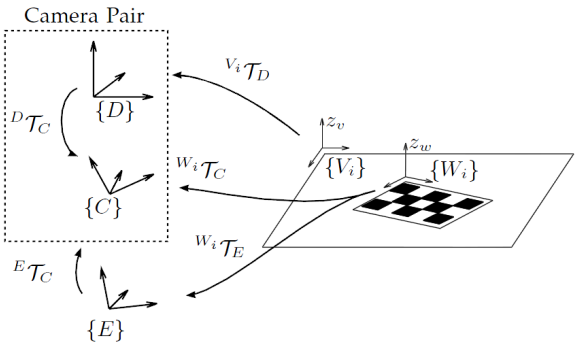


Figure 5.32: Reference frames and transformations. $\{D\}$, $\{C\}$, and $\{E\}$ are the depth, color, and external cameras. For image i, $\{V_i\}$ is attached to the calibration plane and $\{W_i\}$ is the calibration pattern.

From the metric depth, the 3D metric position (X_{ir}, Y_{ir}, Z_{ir}) of the pixel, with respect to the IR camera, can be computed using the equation (5.58),

$$\begin{pmatrix} X_{ir} \\ Y_{ir} \\ Z_{ir} \end{pmatrix} = \begin{pmatrix} \frac{(x_{ir} - c_{xir}) * Z_k(x_{ir}, y_{ir})}{f_{xir}} \\ \frac{(y_{ir} - c_{yir}) * Z_k(x_{ir}, y_{ir})}{f_{yir}} \\ Z_k(x_{ir}, y_{ir}) \end{pmatrix} \tag{5.58}$$

where x_{ir}, y_{ir} are the coordinates of the depth pixel in the image, f_{xir},f_{yir} are the lengths in effective horizontal and vertical pixel size units (IR camera focal length), c_{xir}, c_{yir} are the coordinates of the image center of IR camera, and $Z_k(x_{ir}, y_{ir})$ is depth in millimetres.

The mapping between color image and depth image is expressed by equations (5.38).

5.3.5 Filtering Methods

Depth maps containing holes, inconsistent data in the depth image object boundaries and vibrating behavior at the depth pixel level should be addressed to improve 3D reconstructions. Temporal filtering methods based on time data averaging improve the depth map's quality, although they are impractical in real-time applications or where moving objects exist.

Several noise removal methods are discussed to enhance the Kinect depth maps quality: *median filter*, *bilateral filter*, *joint bilateral filter*, *non-local means filter*, *moving square fiiting*

Median filter

The median approach is a non-linear noise reduction filter widely used in image processing due to its characteristic of suppressing impulsive noise while preserving discontinuities (edges)[480]. A median filter replaces a given depth pixel $D(i,j)$ with the median values found in a local neighbourhood of (i,j). The larger the neighbourhood, the smoother the result is. If ω is a window centered at the spatial (i,j), $D_X(i,j)$ the depth input signal and $D_Y(i,j)$ the respective output signal, then:

$$D_Y(i,j) = m(i,j), \tag{5.59}$$

$$m(i,j) = median\{D_X(i+h, j+k)|h,k \in \omega\} \tag{5.60}$$

The standard median filter usually fulfils 3D reconstruction requirements involving measurement preservation and edge preservation, although additional tuning may be required to avoid depth/texture pixel shifts. Related works [143] [310] have implemented the median filter based on three thresholds to decide when to apply the filter: the number of valid depth values inside the window, the number of depth accurate data on window edges, and the window depth values range.

Bilateral filter

The bilateral filter is a non-linear filter based on Gaussian distribution, which reduces the noise smoothing the signal while preserving the edges. Introduced by Tomasi [495] result from a normalized convolution where the weighting for each pixel s depends on its spatial distance from the center pixel p and its relative difference in intensity. Each neighbour's weight decreases with the distance in the image plane (the spatial domain S) and the distance on the intensity axis (the range domain R).

The technique has been successfully applied in image processing and computer graphics, although the original algorithm has a nominal $O(r^2)$ cost per pixel. The approach efficiency was later improved by Durand [135] and speed up by Paris [388].

For an input depth signal $D()$, an output depth $BD()$, a window ω centered at the spatial p, and a spatial pixel in the kernel s, the depth bilateral filter is defined as follow:

$$BD(p) = \frac{\sum_{s \in \omega} D(s) G_{\sigma_S}(||p-s||) G_{\sigma_r}(|D(p)-D(s)|)}{\sum_{s \in \omega} G_{\sigma_S}(||p-s||) G_{\sigma_r}(|D(p)-D(s)|)} \tag{5.61}$$

where G_{σ_s} and G_{σ_r} are Gaussian functions that determine the weighted spatial geometric distances and the range weighting (similarity), respectively.

Joint bilateral filter

The Joint Bilateral Filter (JBF), (a.k.a cross-bilateral filter) is a bilateral filter variant that makes use of more reliable information, like the RGB image $I()$ (e.g. edges), to guide the depth filtering process. The joint depth bilateral filter is defined as:

$$BD(p) = \frac{\sum_{s \in \omega} D(s) G_{\sigma_S}(||p-s||) G_{\sigma_r}(|I(p)-I(s)|)}{\sum_{s \in \omega} G_{\sigma_S}(||p-s||) G_{\sigma_r}(|I(p)-I(s)|)} \tag{5.62}$$

The formulation is similar to (5.61) except that I replaces D in the range weight domain G_{σ_r}

Related works

Error Analysis related work based on a theoretical mathematical model are presented in [250][465][465][362][389] and calibrations work are described in [207][531][412][104][143]

Reconstruction Applications

Recent RGB-D reconstruction-related works use alignment and integration approaches based on sparse SLAM methods. Henry et al. [205] combine visual feature matching with ICP-based pose estimation to build a pose-graph which they optimize to create a globally consistent map. Newcombe et al. [361] presented an improved accurate solution known as KinectFusion, which uses a new algorithm for real-time dense 3D mapping. KinectFusion integrates depth maps from the Kinect into a truncated signed distance formula (TSDF) representation. The pose estimation required to fuse the depth maps bases on a GPU fast iterative closest point (ICP) implementation.

5.3.6 Experiments and Results Analysis

A crucial step in reconstruction is registration. With precise 3D information from the scene, it is possible to align the point clouds correctly. Experimental results show that for a static scene, the depth measures obtained by the Kinect for a fixed (u, v) pixel can change along time.

Figure 5.33 illustrates the depth variation of 6 specific points on a plane for three different distances using Kinect manufacturer built-in calibration parameters (integer depth values in millimetres). 400 measures were acquired and recorded for the pixels $(230, 310)$, $(230, 320)$, $(230, 330)$, $(250, 310)$, $(250, 320)$, $(250, 330)$, with a plane positioned at 500 mm, 720 mm and 800 mm respectively. For each considered depth plane, measures were acquired in millimetres for each point while maintaining the light conditions as stable as possible and the chosen points far from plane edges.

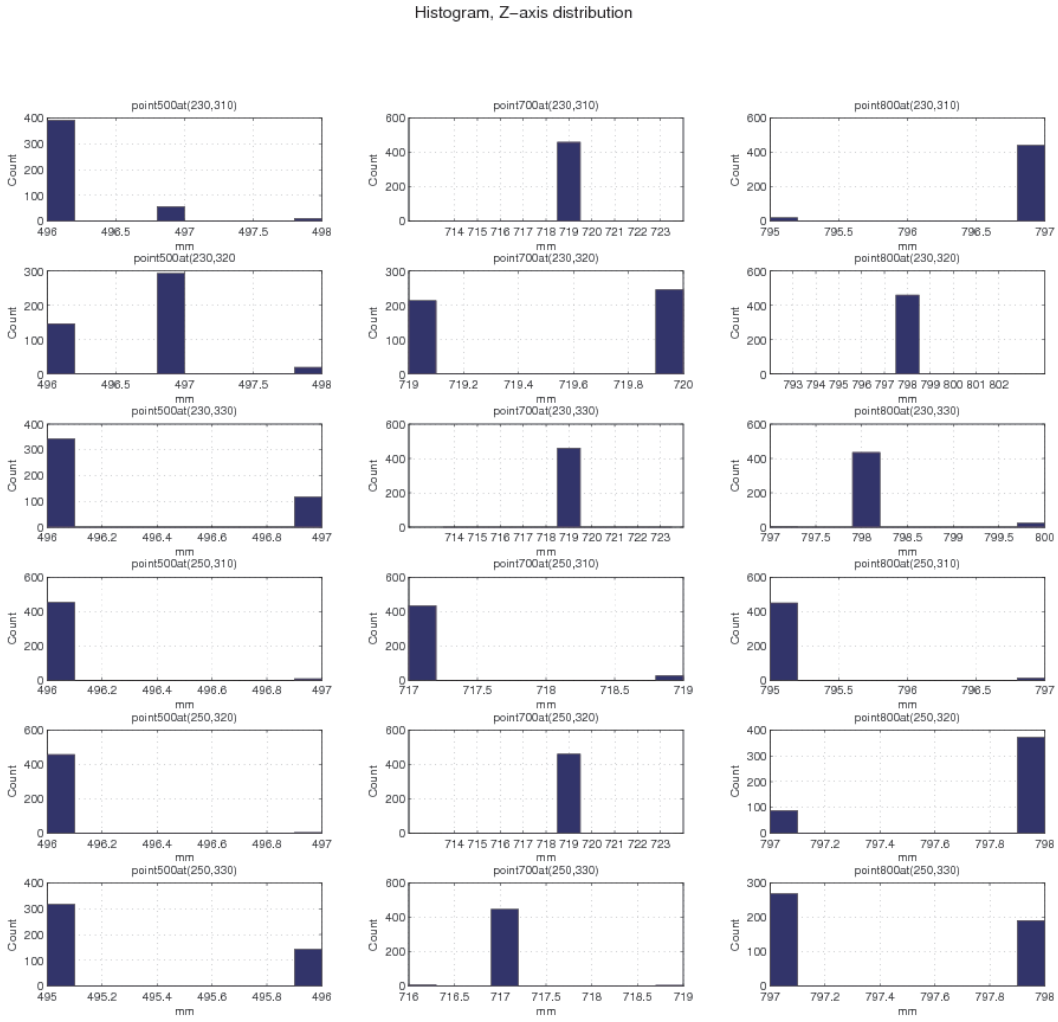


Figure 5.33: Depth variation at six specific points of a plane when positioned at 500mm, 720mm and 800mm, respectively using Kinect manufacturer built-in calibration parameters. Observed depth variations of 3 mm for a static object point in 400 consecutive measures

Results show that although some depth measures are exclusively centered in one same value, like the case of *point700at(230, 310)*, *point700at(230, 330)* or *point700at(250, 320)*, centered on the 719 mm, there are other measures, like point *700at(230,320)* that disperse its measures equally between two consecutive values, 719 mm and 720 mm. The uncertainty is higher in *point500(230,320)*, where for the same $(230, 320)$ pixel, we observe depth variations of 2 mm. At pixel *point700at(250,330)*, the variation reaches 3 mm, although the measure is mainly centered on the value 717 mm.

5.4 Multi-sensor 3D Volumetric Reconstruction Using CUDA

This chapter presents a full-body volumetric reconstruction of a person in a scene using a sensor network, where some of them can be mobile. The sensor network is comprised of couples of camera and inertial sensor (IS). Taking advantage of IS, the 3D reconstruction is performed using no planar ground assumption. Moreover, IS in each couple is used to define a virtual camera whose image plane is horizontal and aligned with the earth cardinal directions. The IS is furthermore used to define a set of inertial planes in the scene. The image plane of each virtual camera is projected onto this set of parallel-horizontal inertial-planes, using some adapted homography functions. A parallel processing architecture is proposed in order to perform human real-time volumetric reconstruction. The real-time characteristic is obtained by implementing the reconstruction algorithm on a graphics processing unit (GPU) using Compute Unified Device Architecture (CUDA). In order to show the effectiveness of the proposed algorithm, a variety of the gestures of a person acting in the scene is reconstructed and demonstrated. Some analyses have been carried out to measure the performance of the algorithm in terms of processing time. The proposed framework has potential to be used by different applications such as smart-room, human behaviour analysis and 3D tele-conference.

5.4.1 Introduction

Performing 3D volumetric reconstruction of people is one of the major research topics in the computer vision area. In this paper we present a volumetric 3D reconstruction method using no planar ground assumption. In order to observe the scene, a sensor network is employed. Each node of the network is comprised of a couple of Inertial Sensor (IS) and camera. In each couple, the IS is used to define a virtual camera whose plane is horizontal and its axes are aligned to the earth cardinal directions. Moreover, a set of inertial-planes, which are parallel to each other and horizontal, is defined in the scene for the purpose of 3D data registration. The image planes of virtual cameras are projected onto these inertial-planes using a geometric method through the concept of homography.

This work is an extension of our previous work [16]. In this paper, after presenting a comprehensive description of the framework, we have provided a parallel processing architecture for the algorithm which allows us to have a real-time implementation of the proposed approach. An effective implementation using GPU-CUDA has been carried out. 3D reconstruction of a person acting in a scene is provided while he is performing different gestures. To analyse the system's performance, task time measurement were executed while changing some system parameters and are available on this text.

There have been many works in the area of 3D reconstruction. Khan in [249] proposed a homographic framework for the fusion of multi-view silhouettes. A marker-less 3D human motion capturing approach is introduced in [338] using multiple views. Zhang in [534] introduced an algorithm for 3D projective recon-

struction based on infinite homography. Homography-based mapping is used to implement a 3D reconstruction algorithm by Zhang and Hanson in [537]. Sorman et al. in [469] presented a multi-view reconstruction method based on volumetric graph-cuts. Lai and Yilmaz in [270] used images from uncalibrated cameras for performing projective reconstruction of buildings based on Shape From Silhouette (SFS) approach where buildings structure is used to compute vanishing points. Feldmann et al. [146] utilized the volumetric 3D reconstruction for the aim of online body motion tracking system. A multi-resolution volumetric 3D object reconstruction has been proposed by Guerchouche et al. in [191]. 3D object reconstruction of an object using uncalibrated images taken by a single camera is proposed by Azevedo et al. in [45]. Lee et al. in [277] applied a 3D reconstruction method using photo consistency in images taken from uncalibrated multiple cameras. A dynamic calibration and 3D reconstruction using homography transformation is proposed by Zhang and Li in [530]. In [290] SFS is combined with stereo imaging for the sake of 3D reconstruction by Lin. Michoud in [339] proposed a method to eliminate appearing ghost object in SFS-based 3D reconstructions. Aliakbarpour and Dias in [15] proposed a method to SFS-based 3D reconstruction. 3D reconstruction of a dynamic scene is investigated in [92] by Calibi. Franco in [163] used a Bayesian occupancy grid to represent the silhouette cues of objects.

In order to have a real-time processing time many researchers have already started to use GPU-based (GP-GPU and CUDA) parallelization of their algorithms.

Joao Filipe et al. in [151] proposed a real-time implementation of Bayesian models for perception through multi-modal sensors by using CUDA.

Almeida et al. implemented the stereo vision head vergence Using GPU-based cepstral filtering [19].

A GPU-based background segmentation algorithm is proposed in [189] by Griesser et al. Ziegler in [540] proposed a GPU data structure for graphic and vision. Real-time space carving using CUDA is investigated in [367] by Nitschke et al. In [474] CUDA is used to accelerate advanced MRI reconstructions. GPU-based method is used in [508] by Waizenegger for the purpose of high resoulution and real-time reconstruction using visual hulls. A GPU-based shape from silhouette (SFS) algorithm is implemented in [527] by Yous et al. An approach for volumetric visual hull reconstruction, using a voxel grid that focuses on the moving target object, is proposed by Knoblauch et al. [262]. A real-time 3D reconstruction system is presented in [268] by Ladikos et al. to achieve real-time performance. Yguel et al. in [525] implemented a GPU-based construction of occupancy grids using several laser range-finders.

As mentioned in [17, 241], recently the use of MEMS[1]-IS has been continuously increasing and its price is decreasing. Nowadays one can see many smart phones equipped with this sensor as well as equipped with camera. There have been many authors showing the advantages of coupling an IS with camera for different purposes such as facilitating the camera network calibration process or increasing the robustness of the calibration result [17, 241, 295].IS is error-prone in sensing the heading direction (rotation in vertical axis) but there have been many methods

[1] Microelectromechanical systems

to overcome such a weakness as can be seen in the literature [241]. Therefore similar to what is mentioned in [241], we do not enter in the area of justifying the benefits of such a coupling and we just assume that each cameras in the network is coupled with an IS.

The rest of this paper is arranged as following: Three dimensional data registration using inertial-planes is investigated in Sec. 5.4.2 where the geometric relations among the virtual-planes are also explored. Moreover, The parallelizing architecture and its implementation using GPU-CUDA is also proposed in Sec 5.4.2. Sec. 5.4.5 presents some experiments where an acting person in the scene is reconstructed in different cases. Moreover some analysis related to the processing is provided in the same section. Eventually Sec. **??** is dedicated to the conclusion part.

5.4.2 Three dimensional data registration using inertial planes

A framework to register the 3D data of the scene is proposed and explained in this section. In this paper, we use the following convention for mathematical symbols: Vectors and matrices are all in bold, except for rotation matrices, camera calibration matrices and homography transforms which appear in normal capital. 3D points appear in capital bold and 2D points in small bold. Superscript in a variable indicates the reference frame in which the variable is expressed. For transformation matrices, subscript indicate the origin system and superscript means the destination system.

5.4.3 Overall 3D reconstruction scheme

An overall scheme of the proposed volumetric reconstruction approach is depicted in Fig. 5.34. Two types of sensors are used: camera, for image grabbing and IS, for obtaining 3D orientation. Each camera is rigidly coupled to an IS. The outputs of each couple are fused using the concept of infinite homography and leads to have a downward-looking virtual camera whose axes are aligned to the earth cardinal direction (North-East-Down). Moreover, the 3D orientation of IS is used to define a set of inertial planes that are all virtual and parallel. The image planes of virtual cameras are projected onto this set of inertial-planes and the 3D volumetric reconstruction of the person (or generally an object) is obtained.

Regarding camera model, the pinhole camera model has been used [198]. In the pinhole camera model, a 3D point $\mathbf{X} = [\ X \quad Y \quad Z\ ]^T$ from the scene is projected onto the image plane of the camera as a 2D point, $\mathbf{x} = [\ x \quad y \quad 1\ ]^T$, using the following model:

$$\mathbf{x} = K\,(R\,\mathbf{X} + \mathbf{t}) \tag{5.63}$$

where K is the *camera calibration matrix,* R and $\mathbf{t}$ are respectively rotation matrix and translation vector between the world and camera coordinate systems[198].

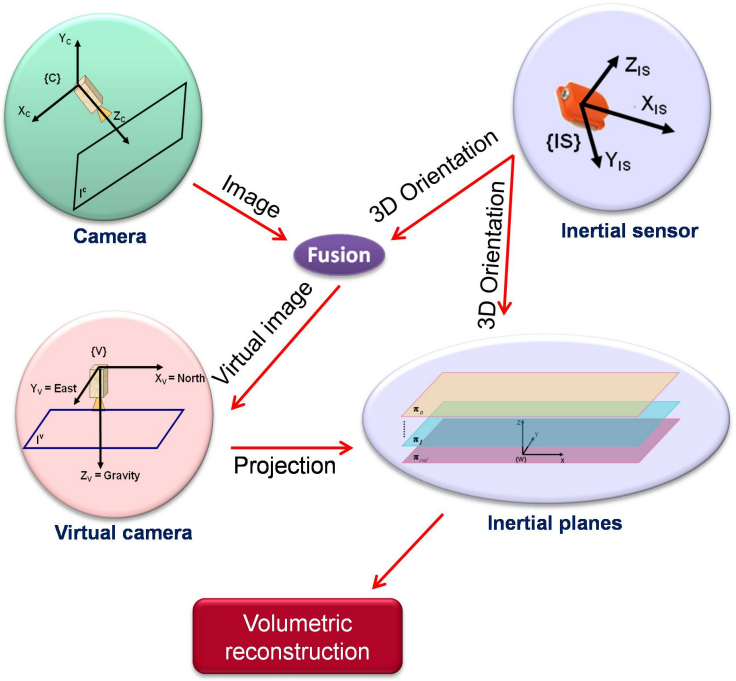


Figure 5.34: Overall scheme of the proposed 3D volumetric reconstruction: 3D orientation from IS and image from camera are fused (using the concept of infinite homography) to define a downward-looking virtual camera whose axes are aligned to the earth cardinal direction (North-East-Down). 3D orientation from IS is as well as used to define a set of inertial-planes in the scene. The 3D reconstruction can be obtained by projecting the virtual images onto this set of parallel inertial planes.

The camera matrix K, which is also referred as *intrinsic parameter matrix,* is defined by:

$$K = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.64}$$

in which f_x and f_y represent the focal length of the camera in the directions of x and y. u_0 and v_0 are the elements of the principal point vector, P. In order to map points from one plane to another plane (with preserving the collinearity) the concept of homography [198] is used. Suppose a 3D plane is observed by two cameras. Moreover, assume that $\mathbf{x_1}$ and $\mathbf{x_2}$ are the image points of a 3D point $\mathbf{X}$ on the 3D plane. Then $\mathbf{x_1}$ and $\mathbf{x_2}$ are called a pair of corresponding points and the relation between them can be expressed as $\mathbf{x_2} = H\,\mathbf{x_1}$ in which H is a 3×3 matrix called *planar homography* induced by the 3D plane [526] and is equal to (up to scale)

$$H = K'\,(R + \frac{1}{d}\,\mathbf{t}\,\mathbf{n}^T)\,K^{-1} \tag{5.65}$$

where R and $\mathbf{t}$ are respectively rotation matrix and translation vector between the two cameras centers, $\mathbf{n}$ is Normal of the 3D plane, d is the orthogonal distance between the 3D plane and the camera center, eventually K and K' are intrinsic parameters of the two cameras (the first camera coordinate system is assumed as the world reference).

Fig. 5.35 shows a sensor network setup with a number of cameras. π_{ref} is an inertial plane[2], defined by the 3D orientation of IS, and is common for all cameras. Here $\{W\}$ is the world reference frame (a detailed specification of this reference frame shall be introduced in Sec. 5.4.3). In this setup, as mentioned before, each camera is rigidly coupled with an IS. The intention is to register a 3D point $\mathbf{X}$, observed by camera C, onto the reference plane $\pi_{\mathbf{ref}}$ as ${}^{\pi}\mathbf{x}$ (2D), by the concept of homography and using inertial data. A virtual image plane is considered for each camera. Such a virtual image plane is defined (using inertial data) as a horizontal image plane at a distance f below the camera sensor, f being the focal length[346]. In other words, it can be thought that beside of each real camera C in the setup, a virtual camera V exists whose center, $\{V\}$, coincides to the center of the real camera $\{C\}$ (see Fig. 5.37). So that the transformation matrix between $\{C\}$ and $\{V\}$ just has a rotation part and the translation part is a zero vector.

In order to register a 3D point $\mathbf{X}$ onto the π_{ref} as ${}^{\pi}\mathbf{x}$, three steps can be taken:

- First, the 3D point $\mathbf{X}$ is projected on the camera image plane by ${}^{\mathbf{c}}\mathbf{x} = P\,\mathbf{X}$ (P is the projection matrix of the camera C).
- Second, ${}^{\mathbf{c}}\mathbf{x}$ (the imaged point on the camera image plane) is projected to its corresponding point on the virtual camera's image plane as ${}^{\mathbf{v}}\mathbf{x}$. Since this

[2]It might appear just as π in the equations

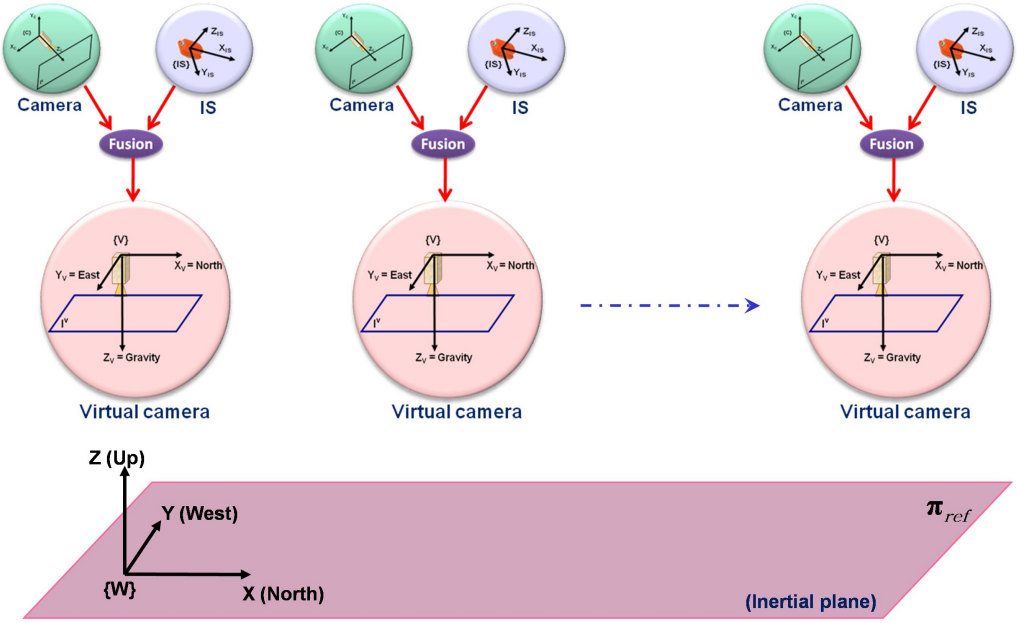


Figure 5.35: A network of sensors observes a scene. The sensor network is comprised of a quantity of IS-camera couples. The inertial and visual information in each couple are fused using the concept of *infinite homography* which leads to define a virtual camera.π_{ref} is a virtual reference plane which is defined by using 3D orientation of IS and is common for all virtual cameras.

operation is plane to plane so it can be done by using ${}^{\mathbf{v}}\mathbf{x} = {}^{v}H_{c}\ {}^{\mathbf{c}}\mathbf{x}$) in which ${}^{v}H_{c}$ is a homography matrix[198].

- Third, the projected point on the virtual image plane, ${}^{\mathbf{v}}\mathbf{x}$, is reprojected to the world virtual plane, $\pi_{\mathbf{ref}}$, by having a suitable homography matrix, called ${}^{\pi}H_{v}$ (this operation is also plane to plane).

The first step it done by the camera based on the pinhole camera model (previously introduced). The second and third steps are described in the following two sub-sections. Assuming to already have ${}^{v}H_{c}$ and ${}^{\pi}H_{v}$, the final equation for registering a 3D point $\mathbf{X}$ onto the reference plane $\pi_{\mathbf{ref}}$ will be (see Fig. 5.38):

$$ {}^{\pi}\mathbf{x} = {}^{\pi}H_{v}\ {}^{v}H_{c}\ K\,(R\,\mathbf{X} + \mathbf{t}) \tag{5.66} $$

The way of obtaining ${}^{v}H_{c}$ (homography matrix between the real camera image plane and virtual camera image plane) and ${}^{\pi}H_{v}$ (homography matrix between the virtual camera image plane and the world 3D plane $\pi_{\mathbf{ref}}$) is discussed in the next sub-sections by starting to describe the conventional coordinate systems.

Image plane of virtual camera

The definition of virtual camera is introduced in this sub-section. We start by presenting the coordinate systems. As seen in Fig. 5.36, there are five coordinate systems involved in this approach to be explained here:

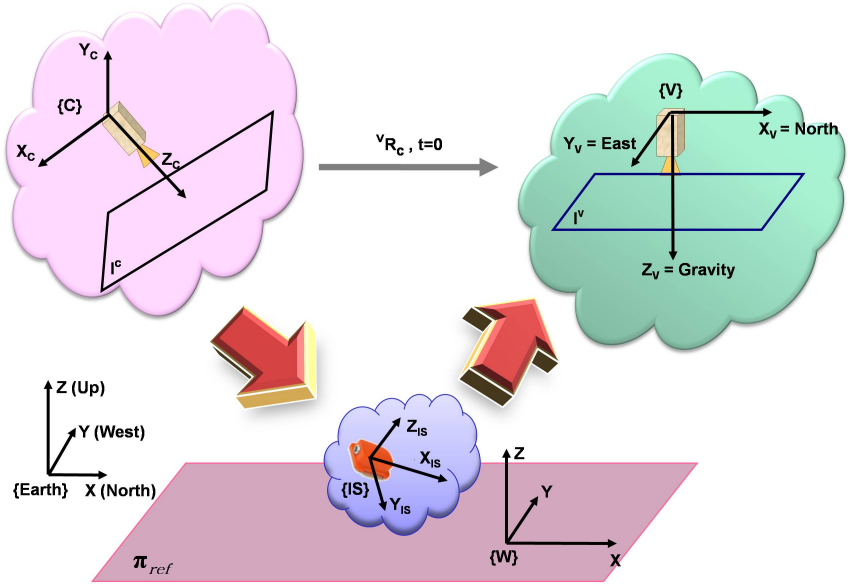


Figure 5.36: Involved coordinate references in the definition of virtual camera; $\{Earth\}$: Earth cardinal coordinate system, $\{IS\}$: Inertial reference frame expressed in $\{Earth\}$, $\{W\}$: world reference frame of the framework, $\{C\}$: camera reference frame, $\{V\}$: reference frame of the virtual camera corresponding to $\{C\}$.

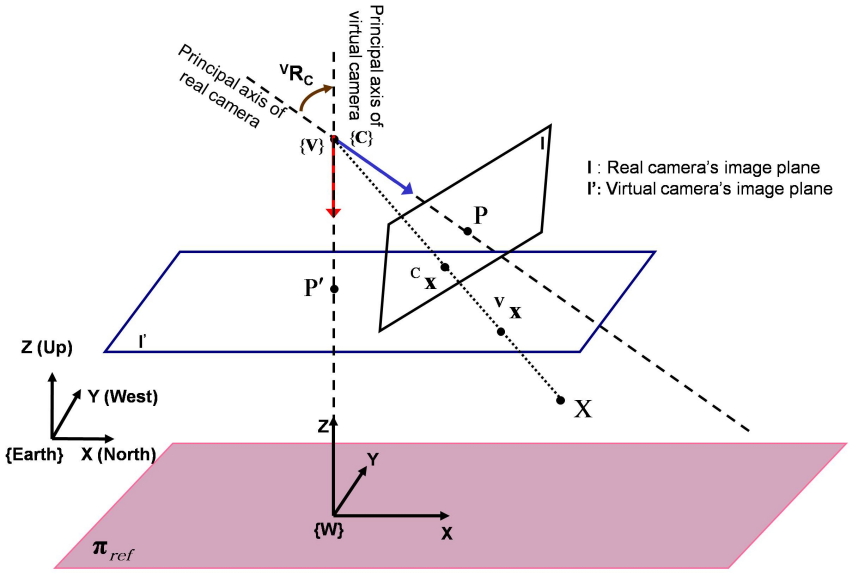


Figure 5.37: Geometrical view of the virtual camera: The concept of infinite homography is used to fuse inertial-visual information and define an earth cardinal aligned virtual camera. Moreover using the inertial information, π_{ref} is defined as a virtual world plane which is horizontal and parallel to the image plane of virtual camera.

- *Real camera reference frame*$\{C\}$: The local coordinate system of a camera C is expressed as $\{C\}$.
- *Earth reference frame* $\{E\}$: Which is an earth fixed reference frame having its X axis in the direction of *North*, Y in the direction of *West* and Z upward.
- *Inertial Measuring Unit reference frame* $\{IS\}$: This is the local reference frame of the IS sensor which is defined w.r.t. to the earth reference frame $\{E\}$.
- *Virtual camera reference frame* $\{V\}$: As explained, for each real camera C, a virtual camera V, is considered by the aid of a rigidly coupled IS to that. $\{V\}$ indicates the reference frame of such a virtual camera. The centers of $\{C\}$ and $\{V\}$ coincide and therefore there is just a rotation between these two references.

The idea is to use the 3D orientation provided by IS to register image data on the reference plane $\pi_{\mathbf{ref}}$ defined in $\{W\}$ (the world reference frame of this approach). The reference 3D plane $\pi_{\mathbf{ref}}$ is defined such a way that it spans the X and Y axis of $\{W\}$ and it has a normal parallel to the Z (See Fig. 5.36). In this proposed method the idea is to not using any real 3D plane inside the scene for estimating homography. Hence we assume there is no a real 3D plane available in the scene so our $\{W\}$ becomes a virtual reference frame and consequently $\pi_{\mathbf{ref}}$ is a horizontal virtual plane on the fly. Although $\{W\}$ is a virtual reference frame however it needs to be somehow specified and fixed in the 3D space. Therefore here we start to define $\{W\}$ and as a result $\pi_{\mathbf{ref}}$. With no loss of generality we place $\mathbf{O_W}$, the center of $\{W\}$, in the 3D space such a way that $\mathbf{O_W}$ has a height d w.r.t the first virtual camera, V_0. Again with no loss of generality we specify its orientation the same as $\{E\}$ (earth fixed reference). Then as a result we can describe the reference frame of a virtual camera $\{V\}$ w.r.t $\{W\}$ via the following homogeneous transformation matrix

$$ {}^{W}T_{V} = \begin{bmatrix} {}^{W}R_{V} & \mathbf{t} \\ \mathbf{0_{1\times 3}} & 1 \end{bmatrix} \tag{5.67} $$

where ${}^{W}R_{V}$ is a rotation matrix defined as (see Fig. 5.36):

$$ {}^{W}R_{V} = \begin{bmatrix} \hat{\mathbf{i}} & -\hat{\mathbf{j}} & -\hat{\mathbf{k}} \end{bmatrix} \tag{5.68} $$

$\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$ and $\hat{\mathbf{k}}$ being the unit vectors of the X, Y and Z axis, respectively.

and $\mathbf{t}$ is a translation vector of the V 's center w.r.t $\{W\}$. Obviously using the preceding definitions and conventions, for the first virtual camera we have $\mathbf{t} = \begin{bmatrix} 0 & 0 & d \end{bmatrix}^T$.

Here we continue the discussion to obtain a 3×3 homography matrix ${}^{v}H_{c}$ which transforms a point ${}^{\mathbf{c}}\mathbf{x}$ on the real camera image plane I to the point ${}^{\mathbf{v}}\mathbf{x}$ on the virtual camera image plane $I^{'}$ as ${}^{\mathbf{v}}\mathbf{x} = {}^{v}H_{c}\,{}^{\mathbf{c}}\mathbf{x}$ (see Fig. 5.37). As described, the real camera C and virtual camera V have their centers coincided to each other, so

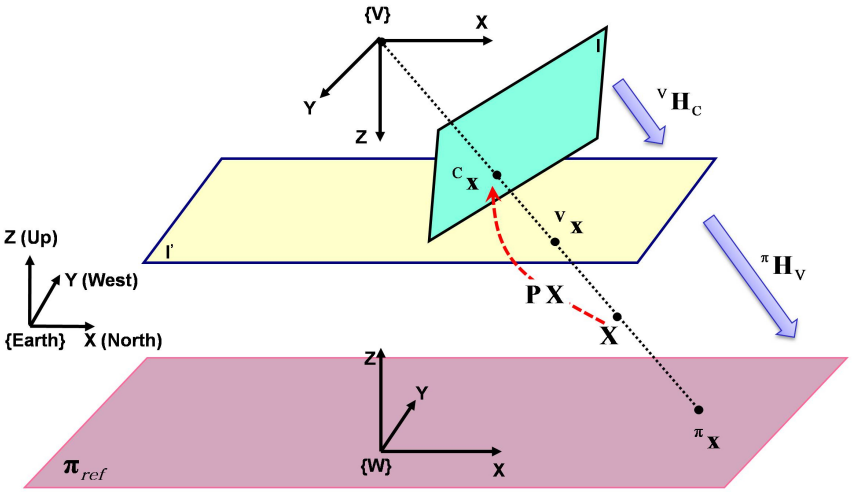


Figure 5.38: One projection and two consecutive homographies are needed to register a 3D point $\mathbf{X}$ from the scene onto a world virtual plane $\pi_{\mathbf{ref}}$ through using IS. ${}^{V}H_{C}$: Homography from real camera image plane to the virtual one, ${}^{\pi}H_{V}$: Homography from the image plane of virtual camera to the reference inertial-plane $\pi_{\mathbf{ref}}$.

the transformation between these two cameras can be expressed just by a rotation matrix. In this case ${}^{v}H_{c}$ is called *infinite homography* since there is just a pure rotation between real camera and virtual camera centers [198].Such an infinite homography can be obtained using a limiting process on Eq. (5.65) by considering $d \rightarrow \infty$ (as described in [198, 526].

$$ {}^{V}H_{C} = \lim_{d\to\infty} K\,({}^{V}R_{C} + \frac{1}{d}\,\mathbf{t}\,\mathbf{n}^{T})\,K^{-1} = K\,{}^{V}R_{C}\,K^{-1} \tag{5.69} $$

where K is the camera matrix ${}^{V}R_{C}$ is the rotation matrix between $\{C\}$ and $\{V\}$. ${}^{V}R_{C}$ can be obtained through three consecutive rotations which is mentioned in Eq. (5.70) (see the reference frames in Fig. 5.36). The first one is to transform from real camera reference $\{C\}$ to the IS local coordinate $\{IS\}$, the second one transforms from the $\{IS\}$ to the earth fixed reference $\{E\}$ and the last one is to transform from $\{E\}$ to virtual camera reference frame $\{V\}$:

$$ {}^{V}R_{C} = {}^{V}R_{E}\ {}^{E}R_{IS}\ {}^{IS}R_{C} \tag{5.70} $$

${}^{IS}R_{C}$ can be obtained through a IS-camera calibration procedure. Here, *Camera Inertial Calibration Toolbox* [296] is used in order to calibrate a rigid couple of a IS and camera. Rotation from IS to earth, or ${}^{E}R_{IS}$, is given by the IS sensor w.r.t $\{E\}$. Since the $\{E\}$ has the Z upward but the virtual camera is defined to be downward-looking (with a downward Z) then the following rotation is applied to reach to the virtual camera reference frame:

$$ {}^{V}R_{E} = \left[\ \hat{\mathbf{i}}\quad -\hat{\mathbf{j}}\quad -\hat{\mathbf{k}}\ \right] \tag{5.71} $$

Projection of 3D data onto a world inertial plane

In this section we describe a method to find a homography matrix that transforms points from a virtual image plane $I^{'}$ (the image plane of virtual camera V) to the common world 3D plane $\pi_{\mathbf{ref}}$ (recalling that these two planes are defined to be parallel. See Fig. 5.38). A 3D point $\mathbf{X}$ on π_{ref} is expressed in $\{W\}$ as $\mathbf{X} = [\ X \quad Y \quad 0 \quad 1\ ]^T$ in its homogeneous form (recalling that XY-plane of $\{W\}$ corresponds to π_{ref} and therefore any points on this plane has $Z = 0$). For a general case (pinhole camera), $\mathbf{X}$ is projected on the image plane as following:

$$\mathbf{x} = K\,[\ \mathbf{r}1 \quad \mathbf{r}2 \quad \mathbf{r}3 \quad \mathbf{t}\ ] \begin{bmatrix} X \\ Y \\ 0 \\ 1 \end{bmatrix} = K\,[\ \mathbf{r}1 \quad \mathbf{r}2 \quad \mathbf{t}\ ] \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \tag{5.72}$$

where $\mathbf{r}1$, $\mathbf{r}2$ and $\mathbf{r}3$ are the columns of the 3×3 rotation matrix, K is the camera calibration matrix (defined in Eq. 5.64) and $\mathbf{t}$ is the translation vector between $\pi_{\mathbf{ref}}$ and camera center [198]. As can be seen Eq. (5.72) indicates a plane to plane projective transformation and therefore can be expressed like a planar homography:

$$\mathbf{x} = {}^{V}H_{\pi}\,{}^{\pi}\mathbf{x} \tag{5.73}$$

where

$${}^{V}H_{\pi} = K\,[\ \mathbf{r}1 \quad \mathbf{r}2 \quad \mathbf{t}\ ] \tag{5.74}$$

, ${}^{\pi}H_{V}$ denoting a 3×3 homography matrix and ${}^{\pi}\mathbf{x} = [\ X \quad Y \quad 1\ ]^T$. Here we recall that for each camera within the network a virtual camera is defined (using inertial data). All such virtual cameras have the same rotation w.r.t world reference frame $\{W\}$. In other words it can be thought there is no rotation among the virtual cameras. ${}^{W}R_{V}$ or the rotation matrix between a virtual camera and $\{W\}$ was described through Eq. (5.68). Then considering ${}^{W}R_{V}$ from Eq. (5.68) and $\mathbf{t} = [\ t_1 \quad t_2 \quad t_3\ ]^T$ as the translation vector, Eq. (5.74) can be formulated as :

$${}^{\pi}H_{V}^{-1} = K\,[\ \hat{\mathbf{i}} \quad -\hat{\mathbf{j}} \quad \mathbf{t}\ ] = \begin{bmatrix} f_x & 0 & f_x\,t_1 + u_0\,t_3 \\ 0 & -f_y & f_y\,t_2 + v_0 t_3 \\ 0 & 0 & t_3 \end{bmatrix} \tag{5.75}$$

It should be mentioned that the translation vector t can be obtained from different approaches. A method to estimate the translation vector among two cameras using two 3D points is proposed in work[17]. In case of possibility of using GPS sensor (e.g. in outdoor scenarios), the translation can also be obtained from this device.

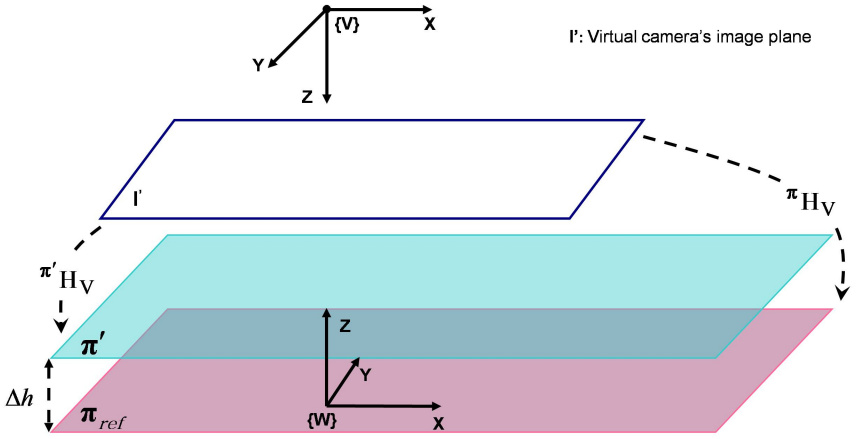


Figure 5.39: Extending homography for planes parallel to π_{ref}. ${}^{\pi}H_V$ is the available homography matrix among virtual image plane $I^{'}$ and the first inertial-based virtual plane π_{ref}. $\pi^{'}$ is another inertial-based virtual plane, parallel to π_{ref}. Δh is the distance among π and $\pi^{'}$. The idea is to obtain ${}^{\pi'}H_V$, the homography between the image plane and $\pi^{'}$, having the ${}^{\pi}H_V$ and Δh (see Eq. (5.76)).

Extension for homographies among virtual camera and inertial-planes

In the preceding section the homography matrix from the image plane of a virtual camera V to the world 3D plane $\pi_{\mathbf{ref}}$ was obtained as ${}^{\pi}H_V$ (see Eq. (5.75)). It is also desired to obtain the homography matrix from a virtual image to another world 3D plane parallel to $\pi_{\mathbf{ref}}$ once we already have ${}^{\pi}H_V$. Lets consider $\pi^{'}$ as a 3D plane which is parallel to $\pi_{\mathbf{ref}}$ and has a height Δh w.r.t it (see Fig. 5.39). ${}^{\pi^{'}}H_V$ denotes the homography transformation which maps the points of the image plane of V onto $\pi^{'}$. By substituting t_3 in the equation (5.75) with $t_3 + \Delta h$, ${}^{\pi^{'}}H_V$ can be expressed as a function of ${}^{\pi}H_V$ and Δh as follows:

$$\boxed{{}^{\pi^{'}}H_V^{-1}(\Delta h) = {}^{\pi}H_V^{-1} + \Delta h\, P\, \hat{\mathbf{k}}^T} \tag{5.76}$$

where $P = [\ u_0 \quad v_0 \quad 1]^T$ is the principal point of the camera V and $\hat{\mathbf{k}}$ is the unit vector of the Z axis.

Algorithm 6 Algorithm of 3D data registration using inertial-planes: First, the image plane of each virtual camera is obtained. Note that the background-subtracted images are binary. Then the image of each virtual camera is projected onto a set of inertial-planes. N_c and N_π indicate the number of cameras and number of inertial-planes, respectively. I_{c_i} and I_{v_i} respectively are the image planes of camera C_i and its corresponding virtual camera V_i. Δh is the euclidean distance among the inertial-planes which also can be interpreted as the vertical resolution of the algorithm. (The labels 'Gpu_Warping', 'Gpu_Project2VirtualPlane' and 'Gpu_Plane_Intersection' correspond to the flowchart in Fig. 5.42)

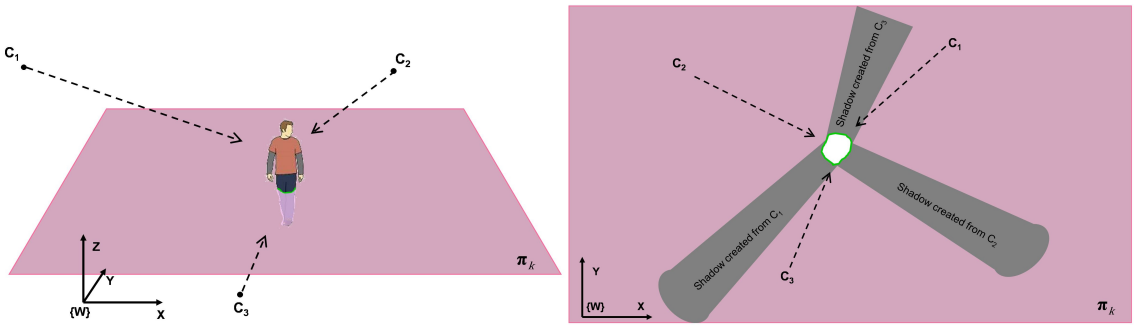


Figure 5.40: Illustration of the registration using homography concept. Left: A scene including a human is depicted. π_k is one inertial-based virtual world plane. The cameras are observing the scene. Right: The registration layer (top view of the plane π_k of left figure). Each camera can be interpreted as a light source.

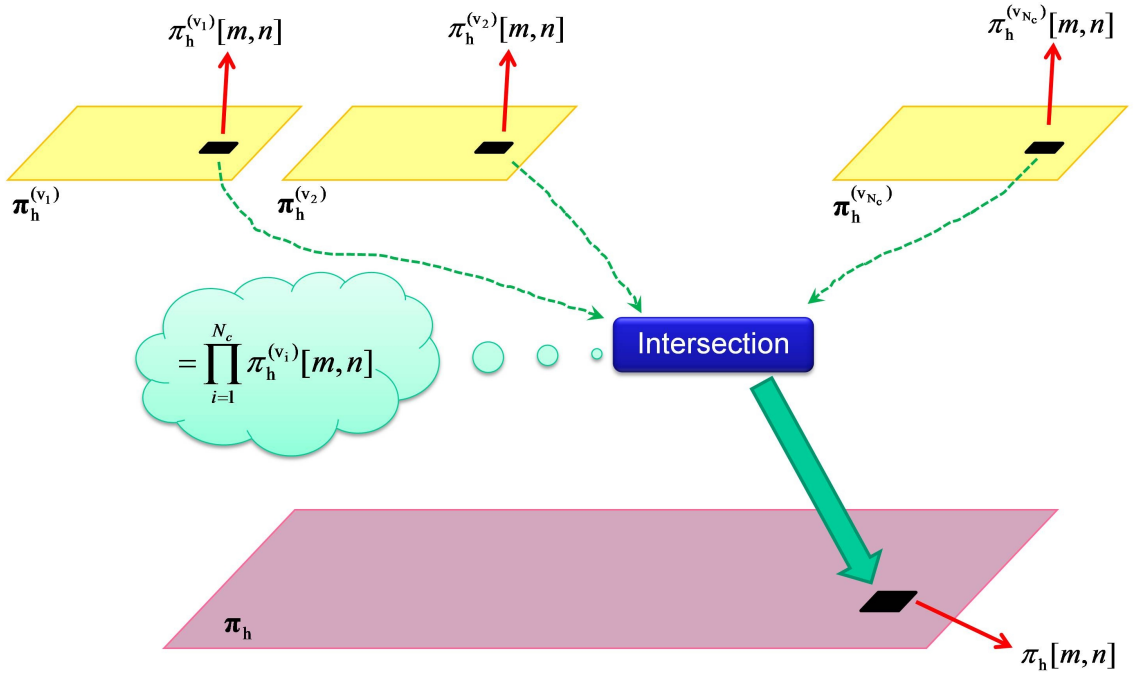


Figure 5.41: Cell-wise intersection of the projections of the virtual images onto an exemplary inertial-plane π_h: Firstly the images of all virtual cameras get projected onto a temporary inertial plane. $\pi_h^{(v_i)}$ indicates the temporary inertial-plane corresponding to the virtual camera V_i. Then the corresponding cells of all temporary inertial-planes are fused using an AND operator in order to provide the final registration on the inertial-plane π_h. (m and n indicate the indices of a cell). Note that the images are considered as binary.

5.4.4 3D Reconstruction using GPU-based parallel processing

Normally the volumetric reconstruction of person is time consuming due to the huge number data to be processed. In order to have a real-time processing (which is necessary for many applications) we propose a parallelizing of the 3D reconstruction algorithm.

GPU Hardware Architecture: in CUDA terminology, the GPU is called the device and the CPU is called the host (see Fig. 5.44). A CUDA device consists of a set of multi-core processors. Each multicore processor is simply referred to as a multiprocessor. Cores of a multiprocessor work in a single instruction, multiple data (SIMD) fashion. All multiprocessors have access to three common memory spaces (globally referred to as device memory but with different access time). The CUDA program is organized into a host program, consisting of one sequential thread running on the host CPU, and several parallel kernels executed on the parallel processing device (GPU). A kernel executes a scalar sequential program on a set of parallel threads. The program organizes these threads into a grid of thread blocks.

The geometric models for projecting 3D data onto a set of virtual horizontal planes based on the concept of homography was previously introduced. Indeed here the homography transformation can be basically interpreted as shadow on each inertial-based virtual plane created by a light source located at the camera position. Considering several cameras (remembering light sources) which are observing the object then different shadows will appear on the inertial planes. For each inertial-plane, the intersection of all shadows on that gives the cross-section of that particular inertial-plane with the object. This interpretation is illustrated in the Fig. 5.40. The left figure shows an exemplary scene where a person is being observed by three cameras. In this figure an inertial-plane π_k is considered. The right figure shows a top view of the same scene which the shadows created by the three cameras. As can be seen the cross-section of the person with the inertial plane is obtained by intersecting all three shadows (the contour shown in white color). By considering a set of inertial-planes in different heights, obtaining the cross-section of each one with the object and stacking them over together the 3D volumetric reconstruction of the object will be obtained.

The proposed reconstruction approach is encapsulated and described as an algorithm in Alg. 6. First, the image plane of each virtual camera is obtained. Then the image of each virtual camera is projected onto a set of inertial-planes. N_c and N_π indicate the number of cameras and number of inertial-planes, respectively. I_{c_i} and I_{v_i} respectively are the image plane of camera C_i and its corresponding virtual camera V_i. Δh is the euclidean distance among the inertial-planes which also can be interpreted as the vertical resolution of the algorithm. The labels 'Gpu_Warping', 'Gpu_Project2VirtualPlane' and 'Gpu_Plane_Intersection' correspond to the fellow-chart in Fig. 5.42.

The images of all virtual cameras initially get projected onto a temporary inertial plane. This part of the algorithm (labelled as 'Gpu_Plane_Intersection') is illustrated in Fig. 5.41 to demonstrate the intersection for an exemplary inertial-plane π_h among the inertial planes. $\pi_h^{(v_i)}$ indicates the temporary inertial-plane corre-

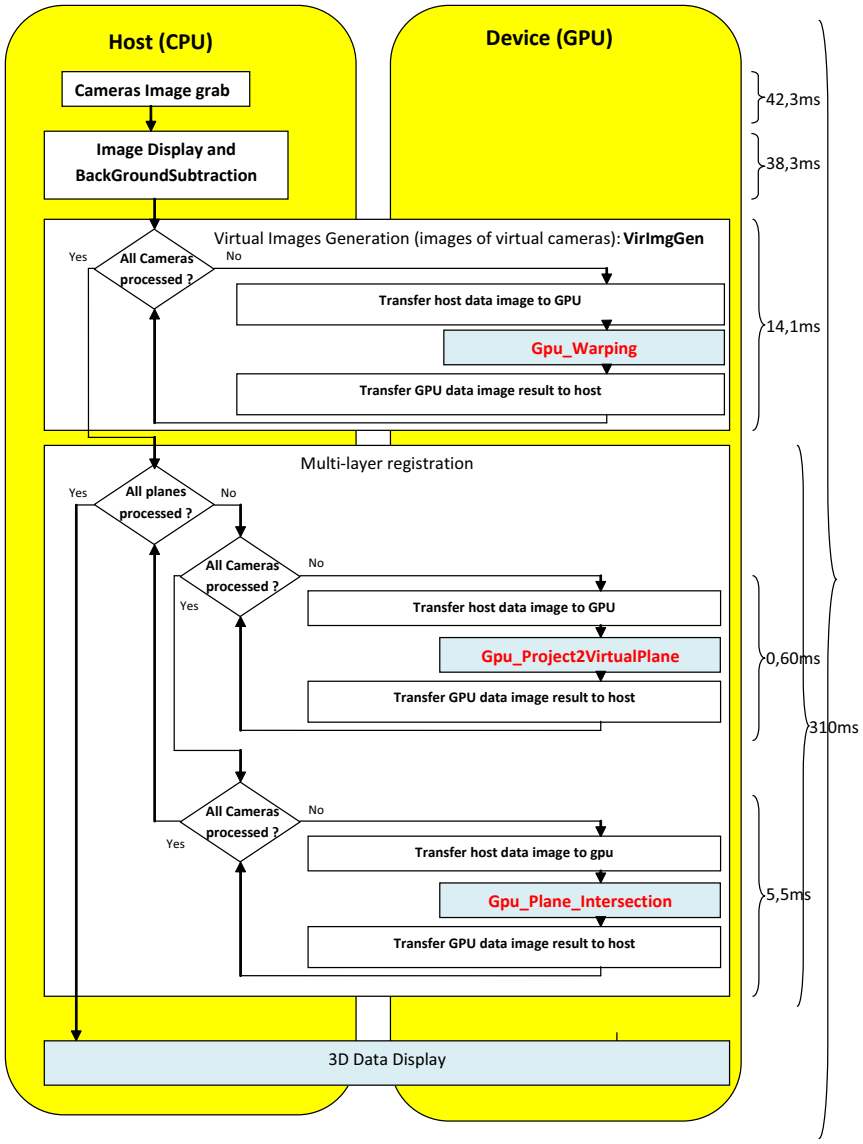


Figure 5.42: Flowchart of CUDA implementation of the proposed inertial-based 3D reconstruction. In the beginning the images are grabbed and then the silhouettes are extracted. After that the silhouettes are loaded on the GPU memory. The loaded images on GPU memory are warped to generate the images of virtual cameras (*VirImgGen*). This part for each camera is done using parallel implementation. After having the images of the virtual cameras generated, the images are projected on the different inertial-planes in order to register the 3D data on them (*GPU_Project2VirtualPlane*). Once images of all cameras get projected onto the inertial-planes, a pixel-wise AND operator is applied to them in order to obtain the intersections. In this point the 3D volumetric reconstruction has been obtained. Eventually the registered data are passed to a visualizer to display the result.

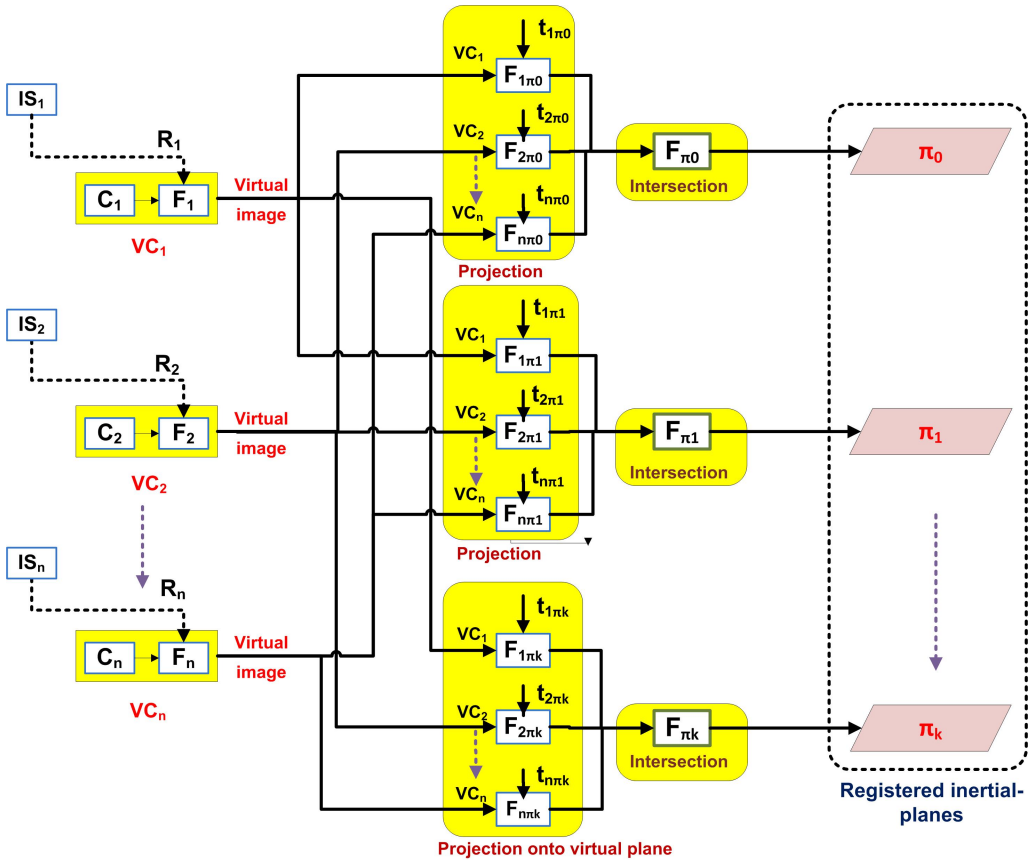


Figure 5.43: The architecture corresponding to the proposed algorithm. The parts coloured in yellow are implemented on CUDA.

sponding to the virtual camera V_i. Then the corresponding cells of all temporary inertial-planes are fused using an AND operator in order to provide the final registration on the inertial-plane π_h. (m and n indicate the indices of a cell). Note that the images are considered as binary.

Fig. 5.43 depicts the another view of the algorithm. The parts colored in yellow are implemented on CUDA. Fig. 5.42 demonstrates the flowchart of the parallel implementation using CUDA. In the beginning the images are grabbed and then the silhouettes are extracted. After that the silhouettes are loaded on the GPU memory in order to be processed by CUDA. The loaded images on GPU memory are warped to generate the images of virtual cameras (labelled as *VirImgGen*). After having the images of the virtual cameras generated, the images are projected on the different inertial-planes in order to register the 3D data on them (labelled as *GPU_Project2VirtualPlane*). Once images of all cameras get projected onto the inertial-planes, a pixel-wise AND operator is applied to them in order to obtain the intersections (labelled as *Gpu_Plane_Intersection*). In this point the 3D volumetric reconstruction has been obtained. Eventually the registered data are passed to a visualizer to show the result. The processes labelled by *VirImgGen*, *GPU_Project2VirtualPlane* and *Gpu_Plane_Intersection* are the part which are done on CUDA using a parallel implementation.

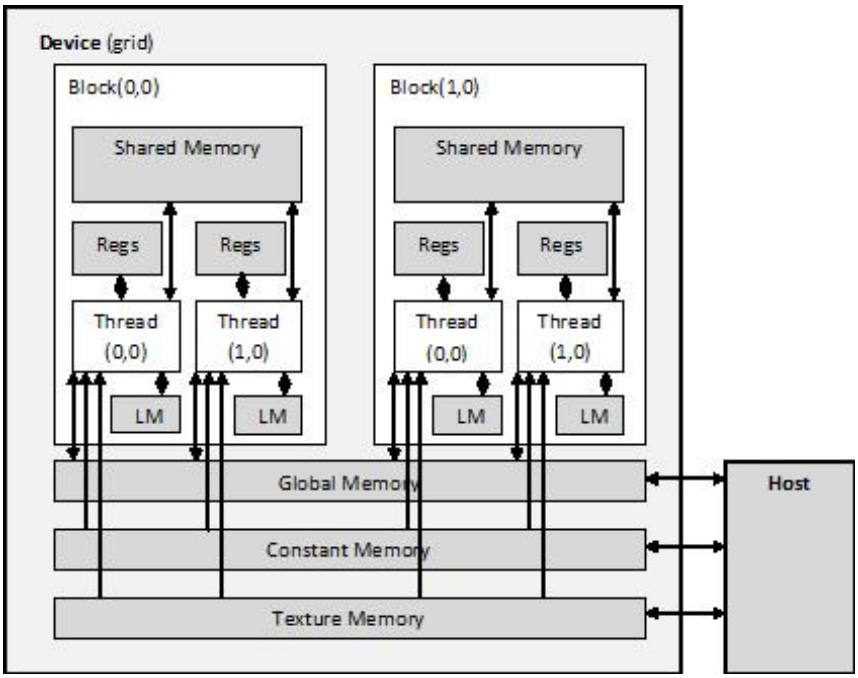


Figure 5.44: CUDA architecture.

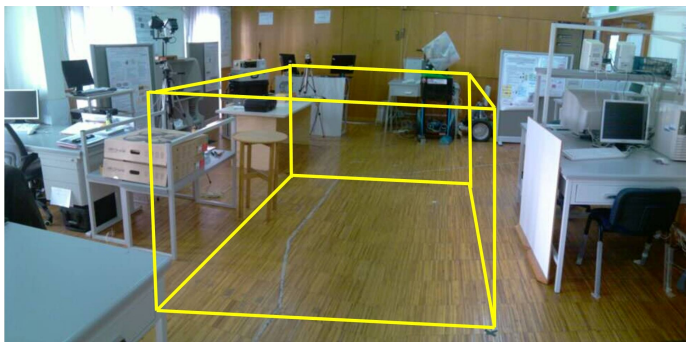


Figure 5.45: The scene used in the 3D reconstruction experiments. The superimposed area indicates where all cameras have overlap in their field of view.

5.4.5 Experiments

Infrastructure

Fig. 5.45 shows the smart-room of the laboratory of mobile robotic in the University of Coimbra [34], used in our experiments. The superimposed area in this figure is observed by a camera network. The cameras are *AVT Prosilica GC650C GigE Color* , synchronized by hardware. Each camera is rigidly coupled with an IS (we used Xsens MTx [522]). The purpose of using IS is to have 3D orientation with respect to earth, obtain virtual camera and define virtual horizontal planes. First the intrinsic parameters of the cameras are estimated using *Bouguet Camera Calibration Toolbox* [85] and then *Camera Inertial Calibration Toolbox* [296] is used for the sake of extrinsic calibration between the camera and IS (to estimate ${}^{C}R_{IS}$). For extrinsic calibration of cameras, a method proposed in our previous work [17] is used. After acquiring image from each camera, a color-based background

subtraction step is performed. To east the background subtraction, the person is dressed in red and his silhouette is separated from the background through color segmentation using the HSV (hue, saturation, value) model. This model is less sensible to illumination changing conditions.A 1-D Hue histogram is sampled from the human area and stored for future use. During frame acquisition, the stored color histogram is used as a model, or look-up table, to convert incoming video pixels to a corresponding probability of body image. Using this method, probabilities range in discrete steps from zero (0.0) to the maximum probability pixel (1.0). Later it is multiplied by a binary mask.

The reconstruction algorithm was developed using the C++ language, OpenCV library [380] and NVIDIA's CUDA software [373] for Ubuntu Linux v10.10. The processing unit responsible for all the sensory and vision algorithm (including CUDA processing) is composed by a PC (Intel Core2 Quad processor Q9400, 6 MB Cache, 4 GB RAM, 1333 MHz and a PCI-Express NVIDIA GeForce 9800 GTX+).

Reconstruction results

A set of experiments have been carried out using the proposed inertial-based 3D reconstruction method by a GPU-based implementation. 24 samples are demonstrated in Fig. 5.47 and 5.48 where an acting person is reconstructed in 3D. One of the samples is seperatly shown in Fig. 5.46 in order to have a more detailed view. In these examples, 48 inertial-planes are used for the purpose of 3D data registration. The interval distance among two consecutive inertial-plane is $5\,cm$. Although the area of the scene in these experiments is small however in the computation the area is considered as $384 \times 384\,cm^2$ which is relatively large. Using a parallel implementation of the algorithm (using GPU), we managed to have a frequency reconstruction close to $2.5\,Hz$ for the mentioned area (using the hardware stated in sub-section 5.4.5). The number of layers and their intervals can be adjusted depending to the application and available hardware.

Statistical analysis on the processing times

Some statistics are carried out in order to show the time which each part of the algorithm takes to run. In Fig. 5.42, processing time for each part of the algorithm is imprinted. The times refer to the case where 48 inertial-planes, each one having a size of $384 \times 384\,cm^2$, have been used. The infrastructure and hardware are as stated in sub-section 5.4.5. The total processing time for a complete 3D reconstruction is $405\,ms$ which leads to have a frequency close to $2.5Hz$.

Fig. 5.49 depicts the average processing time in ms for different size of inertial-planes (the scale is $10^4\,cm^2$). The number of inertial-planes in this experiments is a constant equal to 48. The blue line demonstrates the processing time for generating the images of virtual cameras. Since the number of cameras are fixed in all tests, the execution time for that is almost constant. The red line indicates a

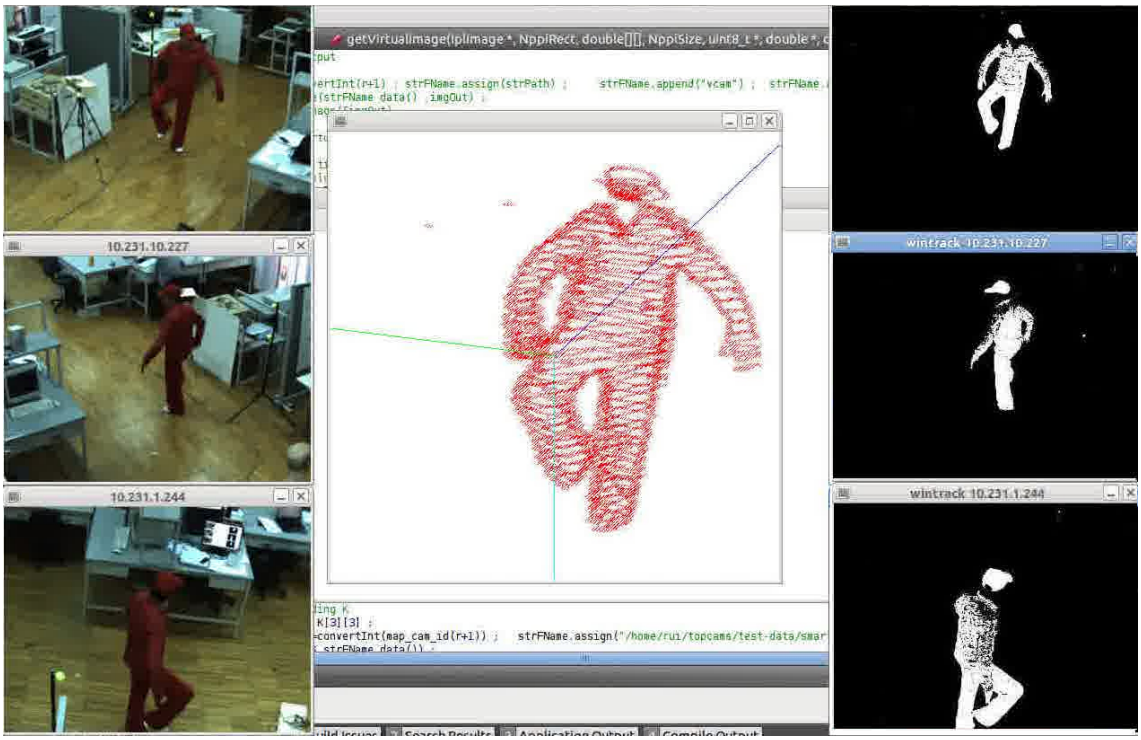
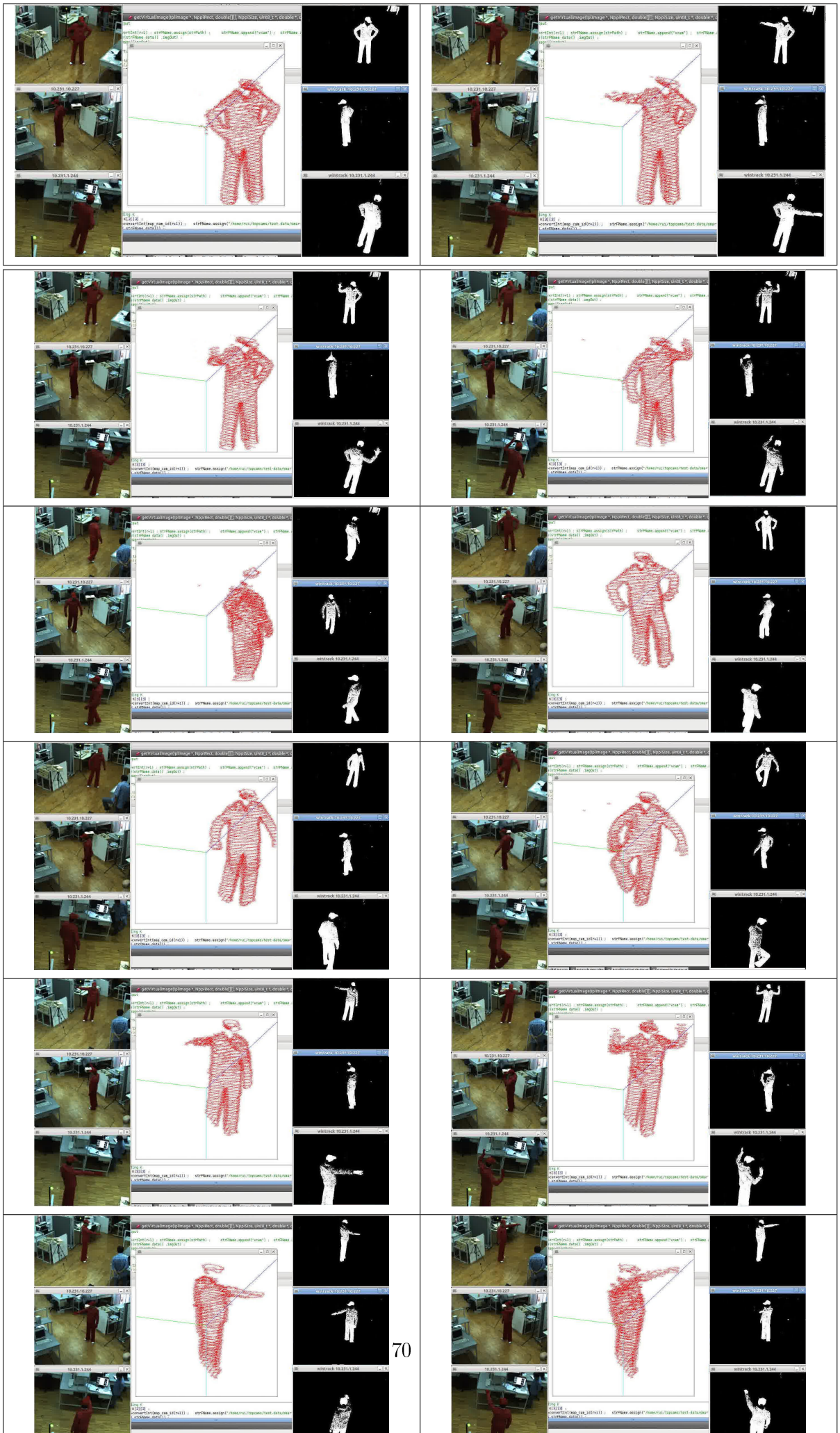
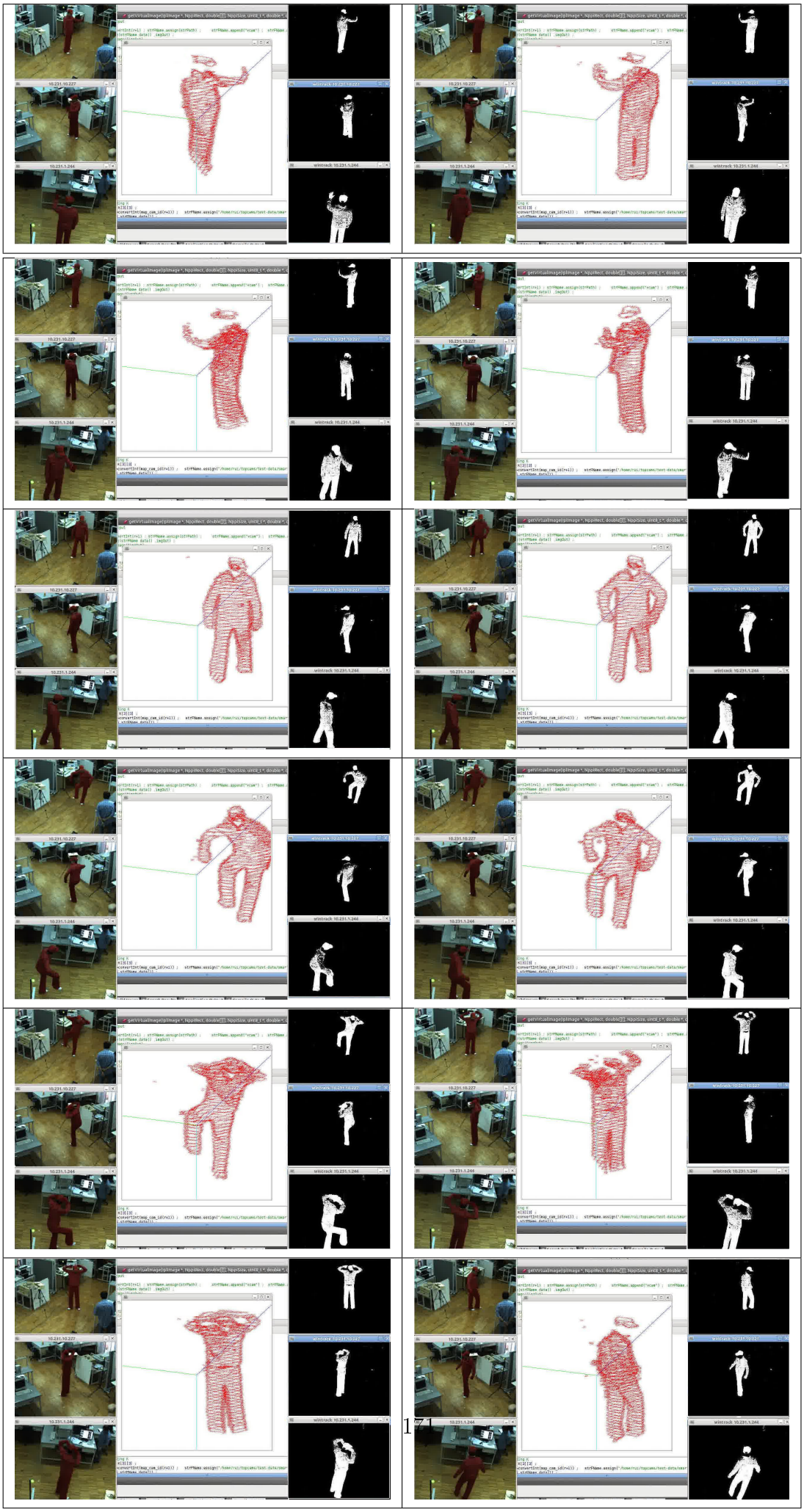


Figure 5.46: Results of 3D volumetric reconstruction using the proposed framework: The camera images before and after background subtraction (silhouette) are respectively shown in the left and right columns. The result of volumetric reconstruction using the silhouette is illustrated in the middle. A network of IS-camera is used to observe the scene. 48 inertial-planes are used to register 3D data from the scene. The interval distance among two consecutive inertial-plane is $5\,cm$.





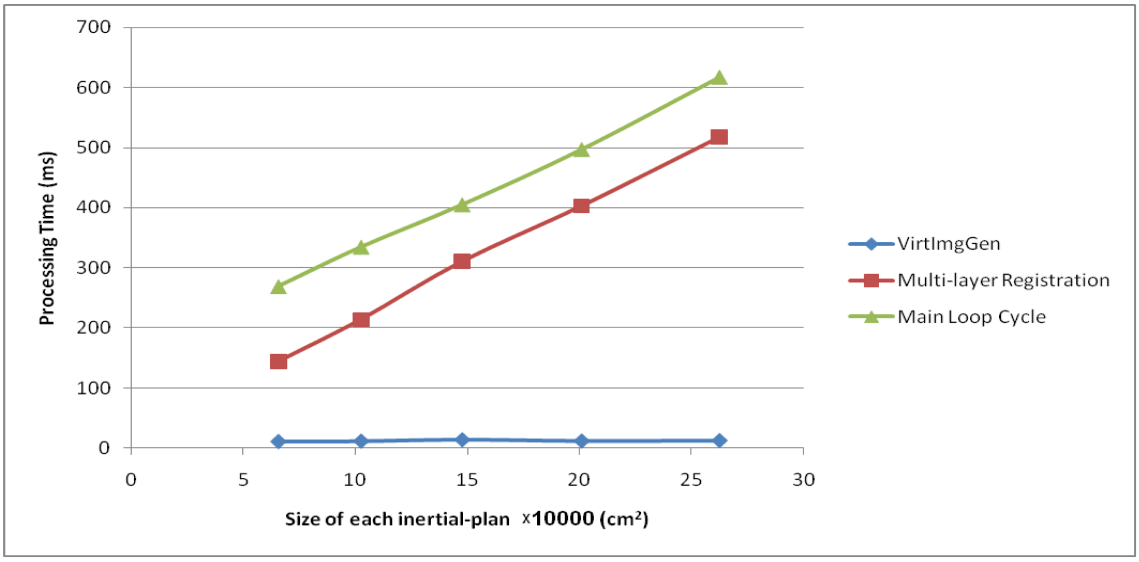


Figure 5.49: Average processing times in ms for different size of inertial-planes. The notations are related to the flowchart shown in Fig. 5.42. Number of 2D inertial-planes used in this statistic is 48.

part where images of all virtual cameras get projected onto a set of inertial-planes. Eventually the total processing time is shown in green color. As it is visible in the diagram, the processing time has a linear proportion related to size (area) of inertial-planes.

Another diagram showing the processing time versus time of inertial-planes is shown in Fig. 5.50. The size of inertial-planes (they are equal in the sizes) is considered as a constant equal to $384 \times 384\, cm^2$. Similar to Fig. 5.49, the colors blue, red and green respectively indicate the processing time of virtual images generation, projection of generated virtual imaged onto a set of inertial-planes and the total algorithm cycle. Also in this diagram the processing time has a linear proportion related to number of allocated inertial-planes.

Extension for mobile sensor

The previously shown experiments were carried out by using static sensors. In some scenarios, it would be very useful to have a mobile sensor which could move inside the scene and collect data from an arbitrary point of view. The data provided by it can be used as a regular node of the sensor network. Such a mobile sensor has two main advantages: Firstly, always it is not possible to have many cameras (specially in large areas) to have all details of the different parts of the scene. Secondly, in some cases one of the main nodes (IS-camera couples) could be occluded or in any reason stop to work. In such situations, a mobile sensor could approach to an appropriate position in the scene, gather and transmit close-view information to the infrastructure. The proposed framework has the ability to integrate the data coming from a mobile sensor. The localization and navigation of a mobile sensor are the two old topics in the area of robotics and computer vision and there can be found many papers in the literatures which

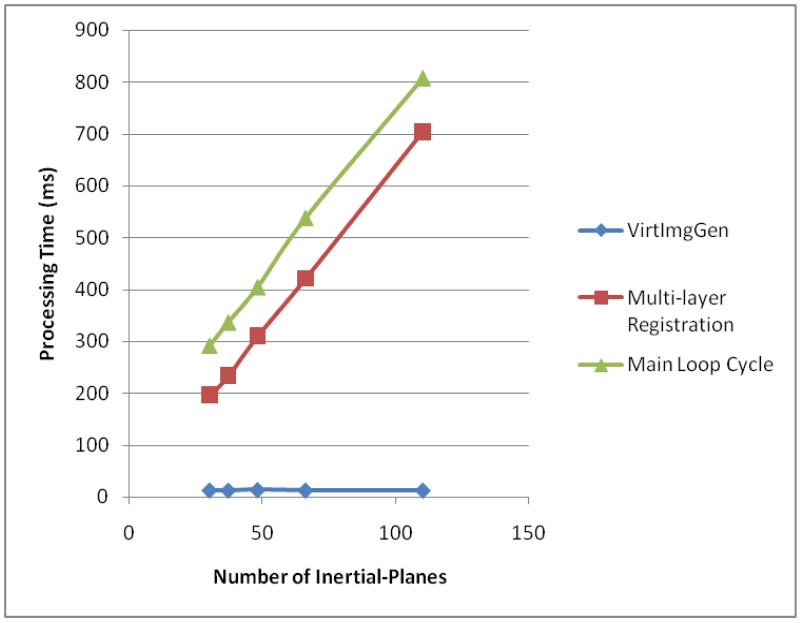


Figure 5.50: Average processing times in ms for different number of inertial-planes. The notations are related to the flowchart shown in Fig. 5.42. The size of each 2D inertial-planes used in this statistic is $384 \times 384\, cm^2$.

proposed different solutions for these problems. Therefore we do not enter in these areas and just assume that we have these techniques already available. In following, an experiment is provided to show the advantage of using a mobile sensor. In order to localize the mobile sensor, the method proposed in [17] is used. Fig. 5.51 shows a case where just two cameras from the infrastructure is used for the 3D reconstruction of a manikin (we intentionally blinded the other cameras). As can be seen, in such situations that there is not enough views to see the scene, the result of 3D reconstruction is not good enough. As seen, there is no enough detail about the reconstructed person and moreover a ghost object has appeared as noise. In order to have more details of the scene, a mobile sensor is navigated close to the manikin. Then after localizing the mobile sensor, its view is integrated as a new node in the network. The results of the 3D reconstruction by using two fixed IS-camera couples and a new added couple is demonstrated in fig. 5.52. This figure shows the advantage of having a mobile sensor which could cooperate with the infrastructure.

Discussion

A set of experiments to demonstrate the applicability and effectiveness of the proposed reconstruction method has been provided. The provided analytical diagrams show acceptable processing times for performing the fully reconstruction of human body within a scene using different parameters. In our experiments although the ground is planar however this ground is not used for estimation of homography matrices. Instead of using planar ground assumption we used the inertial planes to define a virtual ground plane. Many of the introduced papers in the state-of-the-art assumed to have a planar ground such a way that it has to be

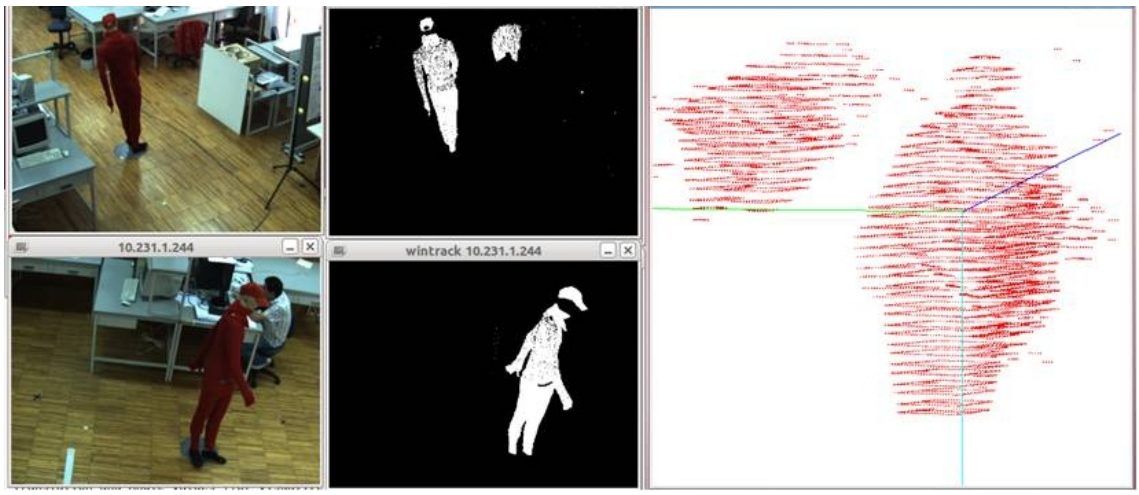


Figure 5.51: Mobile sensor experiment: Result of 3D reconstruction when just two IS-camera couples are used. The other cameras are intentionally blinded. The result is shown in the right column. Because of lack of views, the details are not clear and moreover a ghost object has appeared.

Figure 5.52: Result of 3D reconstruction when a mobile sensor is augmented to the network (corresponding to Fig. 5.51); In order to have more details of the scene, a mobile sensor is navigated close to the manikin and its view is integrated as a new node in the network. The left two columns are the images corresponding to the two fixed cameras and the third column from left is the image corresponding to the mobile camera. The results of the 3D reconstruction by using two fixed IS-camera couples and a new augmented couple is demonstrated in the right column.

possible to mark some points on the ground in order to estimate the homography matrix among the image plane and ground [45][249][507][537][277]. This is not always possible for two main reasons. Firstly in some outdoor scenarios it is not possible to have a flat ground plane. Secondly in some textured grounds it is not possible to use (or mark) a set of points from the ground with known geometry relation among them. Some other authors assumed to have a set of vertical parallel lines in the scene and their images in order to estimate the vertical vanishing point [270][269]. This assumption can not be satisfied as well in some scenarios because of not availability of enough vertical lines in the scene, specially for not man-made scenes.

5.5 Conclusion

A free viewpoint system framework is proposed to generate view dependent synthesis based on scene 3D mesh model. Research contributions includes a new incremental version of Crust algorithm that efficiently adds new vertices to an already existing surface without having to recompute previous generated meshes, an entropy topological incremental reconstruction approach based on confidence measures that avoid redundant data information computation and the introduction of eye/head scan path entropy measures as a potential objective cue for *sense of presence* in conference contexts.. Our approach explores virtual view synthesis through motion body estimation and hybrid sensors composed by video cameras and a low cost depth camera based on structured-light. The solution addresses the geometry reconstruction challenge from traditional video cameras array, that is, the lack of accuracy in low-texture or repeated pattern region. We present a full 3D body reconstruction system that combines visual features and shape-based alignment. Although SIFT has better accuracy as key feature descriptor, we have chosen SURF method in order to achieve the real-time characteristic. Experimental results shows that considering a high number of inliers (not all SURF point features) increases the alignment accuracy. Modeling is based on meshes computed from dense depth maps in order lower the data to be processed and create a 3D mesh representation that is independent of view-point. Research contributions includes a new incremental version of Crust algorithm that efficiently adds new vertices to an already existing surface without having to recompute previous generated meshes, an entropy topological incremental reconstruction approach based on confidence measures that avoids redundant data information computation. This work presents an on-line incremental 3D reconstruction framework that can be used on low cost telepresence applications or human robot interaction applications.

A depth accuracy and error analysis was performed on the Kinect 1 RGD-D sensor. Additionally, an human volumetric reconstruction was developed to support applications such as human-behaviour understanding, smart-room, health-care, surveillance etc. Nowadays, camera network is frequently deployed for public or even private observations for different purposes depending to the application. Recently, IS is becoming much cheaper and more available. Even many smart phones can be found equipped in both IS and camera. Taking advantage of

this, we used a network of IS-camera couples to observe the scene and then a method for 3D reconstruction of a person using inertial data and with no planar ground assumption was proposed. In order to achieve a real-time execution, a parallel processing architecture was proposed and implemented on CUDA. The 3D reconstructions of a person acting in the scene in different gestures are quite promising. The experiments demonstrated the applicability and effectiveness of the proposed approach for many applications.

Chapter 6

Social Robotics towards Telepresence / Co-Presence

Telepresence robots are becoming popular in social interactions involving health care, elderly assistance, guidance, or office meetings. There are two types of human psychological experiences to consider in robot-mediated interactions: (1) telepresence, in which a user develops a sense of being present near the remote interlocutor, and (2) co-presence, in which a user perceives the other person as being present locally with him or her. This work presents a literature review on developments supporting robotic social interactions, contributing to improving the sense of presence and co-presence via robot mediation. This survey aims to define social presence, co-presence, identify autonomous "user-adaptive systems" for social robots, and propose a taxonomy for "co-presence" mechanisms. It presents an overview of social robotics systems, applications areas, and technical methods and provides directions for telepresence and co-presence robot design given the actual and future challenges. Finally, we suggest evaluation guidelines for these systems, having as reference face-to-face interaction.

6.1 Introduction

Telepresence robots are becoming popular in the context of social interactions. Typically, these systems enable people to look at a distant place via teleoperating a robot and interacting with another person at a remote location using the built-in communication devices. Some relevant applications include health care, elderly assistance, autism therapy, guidance, and office meetings [11, 41, 221, 228, 368, 451, 496].

This literature review aims to gather knowledge to help roboticists design improved user- and environment-adaptive systems and technical methods that contribute to enhancing the sense of presence or co-presence via social robot mediation. Reviews have addressed user-adaptive systems [321, 368] and environment-adaptive systems [202] for social robotics (in which the robot is generally an autonomous agent serving the bystander user). However, we further explore

telepresence social robotics, with an emphasis placed on the relationship between the robot's operator and the bystander user.

Within social telepresence robots interactions, two types of human psychological experiences can be considered (see Figure 6.1). The first one involves the *remote user*, in which he or she should sense being in the *local environment* (i.e., telepresence) [222, 345], and the second type involves the *local user*, in which ultimately he or she should sense that the remote user is with him or her in the local environment (i.e., *co-presence*) [69, 453]. This research will focus on this last type of interaction, or how to enhance the sense of *co-presence* via robot mediation. To clarify the role of each agent in the interaction, the following terminology is adopted:

1. Mobile robotic telepresence (MRP) system: remotely controllable mobile platform with video conferencing equipment that allows remote users to navigate within a local environment and socially interact with other persons. These systems can incorporate semi-autonomous functionalities to mitigate operation loads such as navigation aids, points to follow, and obstacle avoidance.

2. Robotic telepresence (RP) system: remotely controllable or semi-autonomous robotic device with video conferencing capabilities that enable social interaction with people in the local environment without locomotion means. Remote users can explicitly control parts of the robot (e.g., the head's panning, swinging, tilting, eye gazing, and facial expressions, as well as arm or hand gestures) or enable some semi-autonomous behaviors (e.g., blinking, face tracking, eye saccade, and breathing).

3. Remote user: user that steers the robot from a distant location or simply connects to the robot through a computer interface.

4. Local user: user that shares the physical environment with the robot (bystander).

5. Local environment: environment shared by the local user and robot.

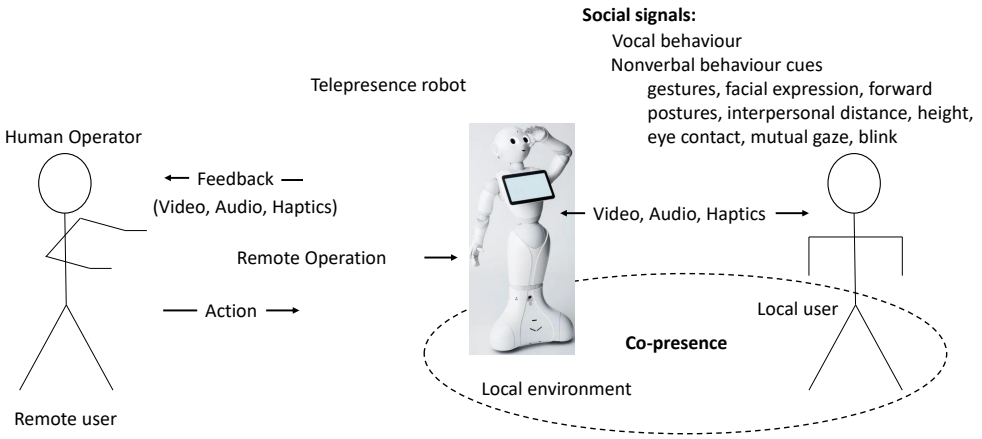


Figure 6.1: Interaction scenario with telepresence and co-presence.

Presence is often defined as the sense of *being there* in a mediated environment [67, 472]. Additionally, Sheridan [450] differentiates presence (virtual) from telepresence (experiential). *Presence* describes the experience of being present within a virtual world, while *telepresence* refers to the sense of being in a mediated remote real environment. *Co-presence* has been used to refer to the *sense of being together* with others in a mediated (either in remote real or virtual) environment [69, 297, 371, 538].

Marvin Minsky introduced the *telepresence* concept in the teleoperation context to describe the phenomenon in which a human operator feels physically present at a remote location through interaction with the human's sensing systems [345] (i.e., "through actions of the user and the corresponding perceptual feedback provided by the teleoperation technology") [222].

Paulos and Canny [395] developed one of the first telepresence robots and referred to it as a personal roving presence (PRoP) device. The goal was to "provide a physical mobile proxy, controllable over the Internet to provide tele-embodiment". The system consisted of a simple controllable mobile platform with a video conference set-up (microphone, speaker, and a video camera with 16x zoom and a 30-cm screen on the top of a plastic pole). Additionally, the robot enabled simple gesturing through a two-DoF pointer. They introduced the concept of *tele-embodiment* in the robotics context to describe the sensation of embodiment of a human in a real distant location [394]. *Tele-embodiment* was defined as *telepresence with a personified perceptible body* [396]. However, they did not address key conditions such as body ownership [461] or agency [74]. Li [284] surveyed and compared 33 experimental works involving people's interactions with virtual agents, telepresence robots, and co-present robots, concluding that robots are more persuasive and positively perceived when they are physically present in the user's environment.

Short [453] introduced the concept of *social presence*, defining it as the degree of salience of the participants involved in an interaction and their interpersonal relationship. He mentioned that *social presence* relied on two concepts: *intimacy* and *immediacy*. *Intimacy* senses the degree of connectedness between the interactants, and *immediacy* refers to the psychological sense of togetherness between the communicators. Taking face to face (FtF) as the reference, both concepts are determined by a set of verbal and nonverbal cues such as vocal cues, gestures, facial expressions, and physical appearance. The capability to deliver such cues differs from communication means, so Short considered social presence as the *quality of the medium itself*. Later, Biocca [69] referred to *social presence* as the effect on one person's behavior caused by the presence of another or caused by knowing that he or she could be observed. *Co-presence*, defined as the "psychological connection to and with another person" [180, 371], has been explored in several works [103, 372, 378, 401].

Cognitive robotics aims to provide robots with intelligent behavior through a processing architecture that involves perception, long- and short-term memory, learning, and reasoning. These approaches try to deal with people's behavior unpredictability and with real-world complexity. Cognitive technologies are a form of hyper-automation that may combine areas such as symbolic representation, automation, prediction, user-adaptive systems, computer vision (CV), machine

learning (ML), deep learning (DL), or artificial intelligence (AI) [145, 202, 368]. Nevertheless, the use of AI methodologies to emulate or interpret human subjective experiences, such as emotions, should be inspired by neurophysiologic-psychological foundations [39].

An inner issue related to teleoperated telepresence robots is the time delay issue (mainly due to the communication channel and less due to the hardware performance). This can affect synchronicity (rate of message exchange between operator and bystander), compromising the social presence [116] (e.g., degradations in audio and video streams, control streams, and haptic feedback). Problems regarding latency, bandwidth limitations, and channel corruptions should be mitigated, and while early solutions involved user interface design and control theory-based models (e.g., supervisory control or passivity-based teleoperation), the approaches evolved to predictive displays and control. Advanced solutions for time delay issues are using time series prediction methods to predict the time delay, a robot's movements, and user intentions. These new adaptive-based control methods make use of nonlinear statistical models and neural network (NN) or machine learning (ML) techniques (e.g., recurrent neural networks, sequence to sequence, long short-term memory, or generative adversarial networks) [145].

The method for this literature review and article selection consisted of retrieving and collecting review studies on social presence, co-presence, and the principles and heuristics of human–robot interaction (HRI), with emphasis on teleoperated telepresence robots. Searches were performed on bibliographic scientific databases, such as ACM's digital library, Google Scholar, MIT Press Direct, Elsevier's ScienceDirect, IEEE's Xplore, PubMed, Scopus, and Springer. Queries included general keywords such as social robots, social robots survey, co-presence taxonomy, copresence or co-presence robots, telepresence robots, adaptive systems, and more specifically, the compositions of these keywords. The selection of papers for in-depth reading was determined by the number of citations, being a recent publication, being a journal (e.g., IEEE transactions, Elsevier's, or Springer), being a book, including user evaluations studies, or being an article in a reputable conference in the field (e.g., ACM HRI conferences, IEEE Robotics and Automation Society, ICRA, or the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)). Citations in these papers directed new readings and paper selections. Figure 6.2a depicts the article's citation distribution per main topic, Figure 6.2b is from the *Co-Presence Taxonomy/Preditors* topic, Figure 6.2c is from the *From Telepresence to Co-Presence Design* topic, and Figure 6.2d shows the article citation distribution per year.

This survey presents an overview of social robotics systems and focuses on how to enhance the sense of co-presence via robot mediation. It reviews the literature to define social presence and co-presence, identifies predictors, and proposes a taxonomy for "co-presence" and "user-adaptive systems" mechanisms. It provides technical methods to support robotic social interactions. The work discussed in this chapter was originally published in the Applied Sciences journal article [25].

The structure of this article is composed of four parts. Section 6.2 identifies potential predictors for social presence, suggesting a taxonomy for "co-presence". Section 6.3 presents several robotic telepresence systems currently available in the

market or used in research. It also reviews autonomous user-adaptive systems for social robots, aiming for a taxonomy, and additionally provides design guidelines for mechanisms that enhance the sense of co-presence in communications through a teleoperated telepresence robot. It includes guidelines for the evaluation of these systems, having as reference the face-to-face interaction. Finally, Section 6.4 presents the conclusions and future work.

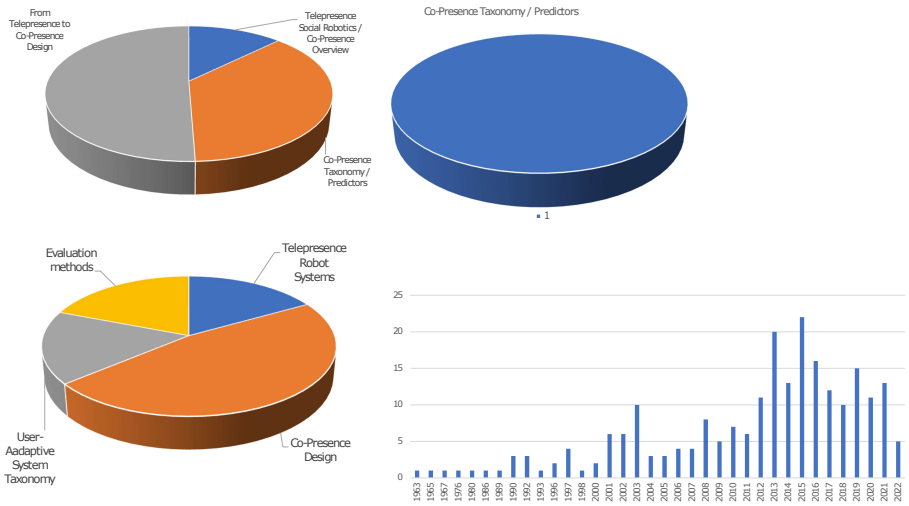


Figure 6.2: (**a**) Article's citation distribution per main topic, (**b**) per *Co-Presence Taxonomy / Preditors* topic, (**c**) per *From Telepresence to Co-Presence Design* topic, and (**d**) the article's citation distribution per year.

6.2 Co-Presence Taxonomy

Social presence has been defined as the *sense of being together with another*, which includes primitive reactions to social cues and automatic creation of simulations or mental models of *"other minds"* [69]. Short et al. [453] defined *social presence* as "the degree of salience of the other person in the interaction and the consequent salience of the interpersonal relationship".

Co-presence is a different concept, introduced by Goffman [180] to describe the active state in which a person perceives his or her interlocutor, and the interlocutor also perceives him or her. *Copresence* refers to a "psychological connection to and with another person", in which "interactants feel they were able to perceive their interaction partner and that their interaction partner actively perceived them" [371]. With co-presence being a subjective concept, it involves different dimensions and interpretations depending on the social science discipline and application area (e.g., sociology or psychology) [378, 401, 538].

Social presence appears in the literature as being related to the *quality of the communication's medium* [453] and the user's perception of the medium. Therefore, preliminary studies have focused on the effect of modality on social presence. They identified potential predictors of social presence by analyzing the technology's capability to reproduce social cues (e.g., visual representation, audio, and

haptic feedback). The findings were biased by the considered concept definitions. Some predictors contribute directly to presence, co-presence, or social presence, while others affect them indirectly by acting on a person's involvement and immersion. Therefore, it is important to distinguish the "immersion" concept and the "presence" concept [67, 82].

Immersion, also known as *sensorimotor immersion*, refers to the extent and fidelity of physical stimulation affecting the human sensory systems and the system's responsiveness to the motor inputs. The immersive level depends on the number and range of sensors and the motor channels connected to a remote agent in a real environment (e.g., a robot) or to a mediated virtual environment. Immersion is determined by the naturalness and coherence between actions (head, body, and gesture movements) and the expected sensory feedback [75, 456, 462, 463].

Presence is the psychological product of technological immersion, defined as *the perceptual illusion of non-mediation* [298] or simply referred to as the sense of *being there* in a mediated virtual environment [67, 82]. Sheridan [450] differentiates presence (virtual) from telepresence (experiential), in which *presence* describes the experience of being present within a virtual world while *telepresence* refers to the sense of being in a mediated remote real environment [297, 429].

Co-presence has been used to refer to the *sense of being together* with others in a mediated environment, either remote real or virtual [68, 538]. As described in the definitions, the use of concepts such as co-presence and social presence should not be confused as they are assessed differently [371].

In the context of social robotics, there are agents with autonomous and semi-autonomous behaviors that are seen by the local person as the "other". Additionally, some agents simply mediate the communications between two persons (the remote and local users). In the former case, the sense of co-presence is assessed between an artificial system and a person, while in the second scenario, co-presence involves two humans. Typically, in robotic telepresence, the representation of the remote real person is shaped by the technology that mediates communication. This affects the perception of thoughts and emotions when compared with actual face-to-face (FtF) interaction. Such representation of remote humans may be supported through text, images, video, 3D avatars, 3D reconstruction, virtual human agents, computers, and robots. Zhao [538], Cummings [116], and Oh, Bailenson, and Welch [378] reviewed the concepts of social presence and co-presence, and their studies suggest a classification for co-presence predictors. This paper adopts some of these literature predictors, framing them in the context of telepresence social robotics.

To unveil a list of technological predictors of social presence, the authors of [116] performed a literature review of empirical studies and grouped them according to similar manipulations. They performed a bottom-up analysis process and identified the following predictors (Table 6.1):

Table 6.1: Social presence predictors.

Predictors of social presence [116] (technologically manipulable)	Behavioral realism. Anthropomorphism. Perceived agency of interactant Level of embodiment Synchronicity Inclusion of imagery Inclusion of imagery (dynamic) Inclusion of voice Inclusion of haptic feedback Others

Initial studies were centered on immersive qualities, but the recent literature also began to address contextual and individual factors, given the subjectivity of the social presence concept [378]. Nevertheless, studies on technological predictors dominate the literature, enlarging the *immersive qualities* class.

The categorization of predictors that affect social presence or co-presence, based on related works, point to (1) immersive qualities, (2) contextual and social properties, and (3) individual traits (see Table 6.2).

Table 6.2: Categorization of predictors.

Co-presence factors	Immersive qualities	Modality, visual representation, interactivity, haptic feedback, audio quality, depth cues, video and display.
	Contextual and social properties	Personality / traits of virtual human, agency, physical proximity, task type, social cues, identity cues.
	Individual traits	Demographic variables, psychological traits.

6.2.1 Immersive Qualities

Modality

The first studies on social presence analyzed the effect of the modality on the levels of presence achieved, given that the immersion degree varies. These studies on general modality identified technological features with an impact on the social presence (e.g., visual representation, interactivity, depth cues, audio quality, and display). However, medium communication comprises multiple features, and it

is a challenge to discriminate the contribution of each affordance. In [119], the *media richness theory* refers to varying the technological qualities of the medium affording distinct levels of social presence. General modality was also identified in [115] as a predictor of telepresence while analyzing the influence of immersion. Initial studies analyzed the influence of modality on social presence by comparing (1) Face-to-face (FtF) real interactions with computer-mediated communication (CMC), (2) text-based CMC with mediums supporting visual and audio modalities, and (3) immersive virtual environments with non-immersive virtual environments.

Face-to-face (FtF) interaction is considered the ground truth for social presence [70], and several works compare face-to-face (FtF) interaction and CMC, evaluating the capability of these mediated communications to elicit social presence. In general, these studies reveal that the sense of social presence is higher in an FtF interaction when compared with CMC conversation. Cortese et al. [110] designed a task in which participants had to discuss a news article for 20 min, either with FtF interaction or through computer-mediated communication (CMC) (chat). Communication apprehension was one of the psychological factors to be assessed (i.e., "the level of anxiety or fear associated with either real or anticipated communication with another person or persons"). They found that the CMC participants experienced a lower level of social presence. Researchers assessing the sociability of a partner and the level of co-location again found higher social presence levels for FtF interaction.

In studies involving decision-making scenarios [70] and in online learning achievement [529], the results privileged FtF interaction. One study [162] involving a series of online seminars for 2 months (the same teacher teaching the same contents online and via in-person, FtF interaction) reported no differences in the levels of social presence between both forms of interactions. One justification might be related to the fact that students had enough time to adapt their communication skills to an online learning platform, the evolution of e-learning technologies, and the fact that the students felt more comfortable not moving to a classroom for 2 months. However, this study did not address the characterization of the subjects' ages in their concluding remarks, which could reveal a tendency.

Video and audio modalities guarantee higher degrees of social presence when compared with text-based CMC. However, this difference is not so clear when comparing video-audio modality against audio-only modality. Studies have shown that the introduction of video modality increases the social presence feeling if participants are required to perform visual tasks [124, 125]. In studies that compared video-audio vs. audio modalities when involving tasks that do not require visual feedback, such as interview tasks or decision-making tasks, the researchers did not report a significant difference in the social presence [62].

These studies suggest that increasing the quality of an immersive component, such as a video feature, may not be proportional to the social presence felt. There seems to exist a threshold from which further enhancements of a given modality may not produce an additional contribution.

Table 6.3 summarizes the relevant aforementioned and next predictor's references, their significant conclusions, and insights on statistics comparisons.

Visual Representation

In communication, the visual representation of interactants is a feature with an impact on social presence. Research has explored to what extent a representation form of the partner can contribute to the sense of social presence. Typically, studies manipulate (1) inclusion or no inclusion of visual representation and (2) the level of realism of the visual representation. The authors of [76] defined *realism* as the extent to which a digital human representation behaves and appears like a real human. The overall concept is referred to as being based on three components: photographic, anthropomorphic, and behavior's communicative realism.

The photographic component assesses the human-like visual appearance in a representation. Most studies report that the existence of a visual representation of the partner leads to higher social presence levels. In [252], participants who spoke with their partners through an avatar while shopping in a virtual mall felt a higher social presence in comparison with those who talked without seeing any partner representation. In [147], in an online support-seeker activity, users reported a higher sense of social presence when a profile picture of the counselor was present, as opposed to to not having a picture. The users also demonstrated a higher willingness to answer questions when the pictures were available.

Anthropomorphism contributes to communicative realism because physical attributes such as the mouth, eyes, arms, and legs are involved in speech to generate facial expressions, gestures, and movements. It addresses the level of interpretation of what is not human or personal in terms of human or personal characteristics. Apart from behavioral realism, this manipulation focuses on the degree to which interactants are presented as human-like on the visual and auditory plane. For instance, users would interact via video as opposed to users who would interact via motion capture-controlled cartoons or by means of anthropomorphic agents or avatars vs. animated forms or emoji.

Communicative realism addresses the degree to which a digital representation of the partner presents physical and social human-like behavior (e.g., breathing, natural blinking, and posture changes). Behavior realism studies manipulate the presence or absence of nonverbal behavior (e.g., animation) or the degree to which the nonverbal behavior of a virtual human resembles a real human (e.g., with or without eye gazing). The effect of communicative (behavioral) realism is more evident when the behavior of an agent or avatar reflects the awareness of the partner's presence (e.g., nodding at the right time, mutual gazing, or blushing). Von der Pütten et al. [506] found that nodding the head of a computer-controlled agent during an interaction contributed to a higher degree of social presence in opposition to no nodding. In another study [387], the participants of an interaction with a virtual agent reported a higher level of social presence when they saw the agent blushing, a consequence of some mistake during a presentation. Study 1 in [62] found that the participants felt higher levels of social presence when the partner's representation (e.g., avatar) was able to maintain a mutual eye gaze in opposition to the absence of eye gazing. However, Study 2 in [62] realized that maintaining a mutual eye gaze for too excessive a time for video and avatars decreased social presence (i.e., an unnatural behavior). Bent's studies carefully tracked participants'

nonverbal behavior using the head's orientation and position sensors, eye gaze trackers, a breath-monitoring chest belt, and data glove-based finger movement trackers. In the avatar's mediation condition, the tracked data were used to animate the avatars in real time (head and body movements, eye movements, and hand and finger movements). Their findings showed a similar activity in terms of visual attention and nonverbal activity either in video or in avatar conditions, contributing both positively and quite similarly to eliciting social presence. This suggests that avatars can be used as a tool to assess social presence, with the advantage that they enable behavior cue segmentation. Another interesting fact is that users tend to direct their heads to their partner's image but their gazes towards the workspace. A justification for this behavior, even knowing that it is a computer representation of the partners, might be related to the human's unconscious social etiquette, that being to keep the face directed to the interaction partner.

Studies show that behavioral realism tends to contribute consistently and positively to social presence [372, 401], while photographic and anthropomorphic realism presents varying effects (positive [244], neutral [50, 62], or even negative contributions [371]). The justification for these discrepancies might be related to several facts reported in the literature: (1) photographic realism is not the main contributor to social presence (i.e., the appearance of the visual representation has a secondary role in comparison with cues from social behavior) [79], (2) manipulations of small features of the visual representation may not be reflected in social presence questionnaires, and (3) the degrees of behavioral realism differ from study to study, making a quantitative comparison with photographic realism vary [77].

In [51, 172], the researchers evaluated the effect of visual and behavioral realism on the perceived quality of communication using avatars. They found a positive effect on social presence when there was consistency between both realism components [172]; that is, although visual representation does not represent a major contribution, the participants felt a greater social presence when the avatar, demonstrating more realistic behaviors (e.g., inferred vs. random eye gaze) which were complemented by a higher level of photographic realism (avatar with a human-like face instead of a dummy face). Bailenson [51] pointed out the consistency between photographic and behavioral realism as a positive social presence predictor.

In [237], the effect of the 3D avatar type (character-like vs. realistically reconstructed) on users' trust and co-presence in a mixed reality-based collaborative teleconference was explored. Visual representation based on realistically reconstructed avatars has been shown to elicit the user's sense of co-presence.

In [401], virtual humans that demonstrated higher responsiveness to events (behavioral realism) contributed positively to the co-presence in mixed reality environments. Experimental conditions involved remote collaboration, in-person collaboration, and communication interactions via mixed reality, augmented reality, virtual reality, video chat, text apps, and virtual assistants.

Interactivity

The definitions of social telepresence, in the context of robot-mediated interaction, rely on the capability to put forward the robot operator's presence to the local person (bystander). Additionally, the extent to which this person is aware that he or she is talking to or interacting with a human being has an impact on social presence [67] and co-presence [206]. Studies on this subject try to understand the effect of the interactivity of the agent on social presence. Such interactivity may refer to a computer agent, a person's avatar, or a telepresence robot, but the focus of this analysis is on the use of a telepresence robot for conversation mediation. Thus, the level of social presence depends on the fidelity of the medium to support the interactivity that characterizes persons' conversations. It includes visual and audio cues, nonvisual sensing (i.e., directional sound and haptics like force feedback and touch) and environmental interactivity (e.g., response rate to user input, reciprocity of the interaction capability between the remote user and local user, and clarity of causal relationships between remote user actions and local user reactions).

Haptic Feedback

To improve the sense of reality, it is important to provide some type of physical feedback to the operator or bystander. Useful contributions include providing tactile cues to let the user recognize the surface texture and materials or support kinesthetic feedback to help the user experience the weight of a virtual object. These kinesthetic and tactile sensations enable haptic perception.

Haptic feedback is a challenge. However, it may improve the degree of presence considerably. Considerable progress has been made in the field of visual and auditory displays, but haptic feedback is in its early stages, gaining much attention nowadays [148, 156, 398]. Touch contact plays an important role in human interactions. From an early age, babies explore their surroundings with their hands and feel physical contact with parents holding them, and at older ages, handshakes, kisses, and embraces trigger emotions and strengthen relationships. Nevertheless, the physical contact of a robot with a person raises safety issues, and that may justify why haptic feedback is not so prevalent.

Touching a robot's part (e.g., hand or body) or sensing that component pulling our hand, operated by a remote party, can improve co-presence. Remote hand-shaking has been explored [405], and examples include the Nao robot hand-shaking the bystander while the robot's operator uses a low-cost haptic device (WiiMote) to feel it [65]. In [358], a robot hand was attached under a videoconferencing terminal's display, and their evaluation demonstrated that mutual touch enhances the feeling of being close. However, the partner's action should not appear in the video. Gregory Welch et al. [352] developed a tactile telepresence system prototype that enables a remote visitor to convey touching patterns on the forehead of an isolated patient through a tablet touch video interface. Regarding human–robot first encounters [41] and greetings, the authors of [96] used Kendon's model [247] to develop an interaction that included six phases (initiation of approach, distance salutation,

head dip, approach, final approach, and close salutation). Human tracked gestures were the inputs for a decision module (based on the hidden Markov model and behavior tree [107]) that initiated a specific phase at the right moment.

Telexistence surrogate anthropomorphic robot (TELESAR) VI can mimic the user's movements and gestures from a mechanically unconstrained full-body master cockpit and provide haptic feedback to the operator [485]. The 10 fingers of the teleoperated robot are equipped with vibration, force, and temperature sensors that can realistically deliver these components of haptic information. Operators can shake the hand of another person through the robot and feel it.

Depth Cues (Stereoscopy and Motion Parallax)

Considering an interaction between two persons through a teleoperated robot, the depth cues become more important for the remote user, since the local user is with the robot, and has natural depth cues. On the other side, if the remote user can perceive the local user in a 3D space, it improves the scene's realism and the co-presence. The use of 3D displays or head-mounted displays (HMDs) by the remote user are common approaches to delivery depth cues. However, this requires 3D sensors in the robot's side (e.g., stereo cameras and RGBD sensors). Additionally, with the inclusion of an autostereoscopic or 3D display in the robot to present the remote user to the local user, it is possible to enhance the closeness [20, 26, 403].

Audio Quality

As mentioned earlier, audio modalities guarantee higher levels of social presence. The audio channel should provide bidirectional communication between the remote and local users to exchange messages. Recognition of a person's voice plays an important role in person identification, contributing to the sense of co-presence [131]. Voice transmission is expected to be fluid without cuts or delays. Telepresence robots quite often make use of an array of microphones to acquire spatial sound, enabling the remote user to identify the direction of the sound source [275] or simply detect the movements of the local user.

Video and Display

The sense of being telepresent is also determined by the fidelity and capability of the medium to present the remote environment, including the visualization of persons (face expressions, gestures, postural behaviors, etc.). To this end, there are mediation technology requirements that include visual display parameters (e.g., latency, frame rate, field of view (FOV), point of view (egocentric vs exocentric), image resolution, color quality, and image clarity) and environment presentation consistency across displays [24]. Display type comparisons reveal a positive effect on co-presence using immersive 3D displays in nonverbal interactions [251]. For example, the Willow Garage Texai robot rely on the principle "reciprocity of vision

(if I see you, you must see me)", while Excite robot designers defend that "The visitor's [user's] environment should be immersive so that the user would have a first-person experience of the destination [remote environment] including full sensory stimulation focusing on immersive vision, audio, and haptics" [496].

6.2.2 Contextual and Social Properties

Early studies on predictors of social presence focused on immersive qualities; however, the research began to address contextual and individual properties. Given the subjectivity associated to social presence experience, and aside from the physical distance and the medium's technological qualities, analyses started to consider a psychological distance between the interactants [35, 378, 410]. These include factors such as Personality/traits of Virtual Human, Agency [371], Physical Proximity [50, 77], Task Type [125, 254, 335], Social Cues [103, 125, 286], or Identity Cues [103].

In [401], a contextual responsiveness predictor is explored that assesses the capability of a virtual human (VH) to detect and respond to events and cues that happen in the shared space of the VH and the user (e.g., to a broom that falls in the user's physical environment or that falls into the virtual VH space). They showed that when the VH detects and directs gazing at the event or orients itself in that direction, the user presents higher levels of co-presence. Studies suggest that users' perception of the physical space affects their co-presence in mixed reality. Ignoring events in the background, such as objects moving or a person walking [255], or the inability to shift attention to an external event does not contribute to co-presence. In [237], the robot that plays a game with the user uses a "cheat" function to trick the user, which affects the user's trust, contributing positively to co-presence.

In [103], users reported a higher level of social presence when communicating simultaneously with several remote interlocutors through a telepresence robot than with a single remote person. In [103], a second study showed that users felt the presence of the remote interlocutor more when the telepresence robot had a low identity than a higher identity (e.g., robot's head LCD with or without a face drawing).

6.2.3 Individual Traits

Gender and age: social studies showed that female subjects tend to experience higher degrees of social presence when compared with males [50, 238], but age is not a relevant factor [289].

Attractiveness: in [77], a human's avatar that looks more attractive in a virtual mirror raised the person's level of self-confidence in the next encounters with other person's avatars and eventually in the real world (distances between avatars are reduced (proxemics)). Such findings provide traits for telepresence and co-presence robot design.

Height: in [77], a human's avatar that looked taller than its interlocutor tended to make that person more persuasive in new interactions with others.

Psychological traits: a person with a higher immersive tendency showed higher degrees of social presence [289]. Additionally, people more prone to human social interactions reported higher levels of social presence in experiments involving social robots [236]. In [110], persons low in communication apprehension (CA) experienced higher levels of social presence than those high in CA. Less sociable people tended to show lower scores on social presence assessments.

Table 6.3 summarizes the relevant aforementioned and next predictor's references, their significant conclusions, and insights on statistics comparisons (e.g., N = number of subjects, μ = mean, σ = standard deviation; subscripts refer to the condition, where superscript $^+$ = significant condition, df = degree of freedom, F = ANOVA statistic F, p = *p*-value, η^2 = eta squared, χ^2 = chi square, β = standardized path coefficient, and r = correlation coefficient).

Table 6.3: Co-presence studies.

Predictor Category	**Predictor**	**Evaluation Process**	**Study**	**Quantitative Comparison (Statistics)**
Immersion	Modality	FtF vs. CMC (NetMetting teleconference)	[70]	$N = 70$ (38 pairs), $\mu_{FtF} = 34.6$, $\mu_{CMC} = 32.1$, $F = 40.2$, $df = 1$, $p = 0.00$
Immersion	Modality	FtF vs. CMC (chat)	[110]	$N = 152$, $\chi^2(8, N = 152) = 6.267$, $p = 0.617$; ($\beta = -0.948$, $p < 0.001$)
Immersion	Modality	FtF vs. CMC	[529]	$N = 257$, $\mu_{FtF} = 3.63$, $\sigma_{FtF} = 0.62$; $\mu_{CMC} = 3.48$, $\sigma_{CMC} = 0.57$; $t(255) = 2.077$, $p = 0.0039$
Immersion	Modality	FtF vs. CMC (online teaching and learning)	[162]	$N = 50$, $\mu_{FtF} = 38.9$, $\sigma_{FtF} = 1.2$; $\mu_{CMC} = 36.91$, $\sigma_{CMC} = 1.36$; $F(1, 48) = 1.194$, $p = 0.28$
Immersion	Modality	Audio vs. Audio + Video	[125]	$N = 34$ (17 pairs), male: $\mu_{audio} \approx 53.5$, $\mu_{audio+video} \approx 71.75$; $F(1, 18) = 9.9$, $p = 0.04$
Immersion	Modality	Text vs. Audio vs. Audio + Video vs. Audio + Avatar	[62]	$N = 150$, Factor scores: $\mu_{text} = -0.48$, $\mu_{audio} = 0.26$, $\mu_{audio+video} = 0.22$, $\mu_{audio+LFavatar} = 0.09$; $\mu_{audio+HFavatar} = 0.10$; $F(4, 137) = 2.59$, $p = 0.04$, $\eta_p^2 = 0.09$
Immersion	Visual Representation	Photographic Realism (Low- vs. High-Fidelity Avatar)	[62]	$N = 150$, Factor scores: $\mu_{audio+LFavatar} = 0.09$; $\mu_{audio+HFavatar} = 0.10$; $F(4, 137) = 2.59$, $p = 0.04$, $\eta_p^2 = 0.09$
Immersion	Visual Representation	Photographic Realism	[252]	$N = 80$, embodiment index: $voice = 1.68$, $voice + avatar = 5.2$, $df = 4$, $p < 0.01$, $\mu_{embodiment} = 3.41$, $\sigma_{embodiment} = 1.94$; $\mu_{copresence} = 5.27$, $\sigma_{copresence} = 1.44$;
Immersion	Visual Representation	Photographic Realism	[50]	$N = 50$, $\mu_{flat\_shaded\_face} \approx \mu_{photographic\_texture\_face}$
Immersion	Visual Representation	Anthropomorphic	[371]	$N = 134$, copresence index: $low_{anthropomorphic\_image}{}^{+}$, $more_{anthropomorphic\_image}$, no_{image}, $R = 0.18$, $F = 4.23$, $p = 0.04$
Immersion	Visual Representation	Anthropomorphic, Behavioral Realism	[372]	Definitions and it uses, digital representations
Immersion	Visual representation	Behavioral realism (mutual gaze)	[50]	$N = 50$, women's social presence score: $\mu_{no\_mutual\_gaze} = -13.25$, $\sigma_{no\_mutual\_gaze} = 18.58$; $\mu_{high\_mutual\_gaze} = 2.5$, $\sigma_{high\_mutual\_gaze} = 15.55$; 5 conditions ($r = 0.30$, $p < 0.03$)
Immersion	Visual Representation	Consistency between Visual and Behavioral Realism	[172]	$N = 48$, $low\_realism$: $\mu_{random\_gaze} = 1.2$, $\sigma_{ramdom\_gaze} = 0.2$; $\mu_{inferred\_gaze} = 0.7$, $\sigma_{inferred\_gaze} = 0.2$; $high\_realism$: $\mu_{random\_gaze} = 0.3$, $\sigma_{ramdom\_gaze} = 0.1$; $\mu_{inferred\_gaze} = 1.1$, $\sigma_{inferred\_gaze} = 0.3$;

Table 6.3: *Cont.*

Predictor Category	Predictor	Evaluation Process	Study	Quantitative Comparison (Statistics)
Immersion	Visual Representation	Consistency between Visual and Behavioral Realism	[51]	$N = 146$, copresence: behavioral realism$^+$, $F(3, 133) = 2.72$, $p < 0.05$, $\eta^2 = 0.06$; visual representation$^+$ $F(6, 133) = 2.18$, $p < 0.05$, $\eta^2 = 0.09$
Immersion	Visual Representation	Avatar Behavioral Realism to Events	[401]	$N = 65$, copresence: $\mu^+_{responsive} = 4.31$, $\sigma_{responsive} = 0.11$; $\mu_{nonresponsive} = 3.96$, $\sigma_{nonresponsive} = 0.12$;, σ; $F(1, 63) = 5.06$, $p = 0.02$
Immersion	Visual Representation	HMD vs. Desktop	[24]	$N = 21$, presence Q5: $\mu_{HMD} = 5.28$, $\sigma_{HMD} = 1.58$; $\mu_{Desktop} = 3.42$, $\sigma_{Desktop} = 1.77$; $F(1, 20) = 26.54$, $p < 0.0001$
Immersion	Visual Representation	HMD vs. Desktop	[174]	$N = 26$, presence Q5: $\mu_{HMD} = 5.88$, $\sigma_{HMD} = 0.52$; $\mu_{Desktop} = 2.48$, $\sigma_{Desktop} = 1.75$; $F(4, 95) = 32.19$, $p < 0.0001$
Immersion	Visual Representation	2D vs. 3D vs. Verbal vs. Nonverbal	[251]	$N = 40$, copresence: $3D^+_{nonverbal}$, $3D_{verbal}$, $t(16.35) = 7.48$, $p < 0.05$; $2D_{nonverbal}$, $2D^+_{verbal}$, $t(17.967) = -8.05$, $p < 0.05$
Immersion	Interactivity	Whole Body Interaction	[23]	$N = 13$, embodiment: *immersive+body intention-based robot control*$^+$, $F(3, 44) = 19.11$, $p < 0.0001$;
Immersion	Haptic Feedback	Present vs. Absent	[279]	$N = 24$, Embodiment score: Haptic feedback$^+$ = 49.8, $F(1, 23) = 29.67$, $p < 0.0001$; Realism score: Haptic feedback = 33.0, $F(1, 23) = 22.97$, $p < 0.0001$;
Immersion	Depth Cues	Stereoscopy (stereo vs. mono)	[7]	$N = 144$, copresence: $\mu^+_{stereo} = 3.85$, $\sigma_{stereo} = 1.34$; $\mu_{mono} = 3.25$, $\sigma_{mono} = 1.48$; $F(1, 140) = 6.97$, $p < 0.01$, $\eta^2_p = 0.05$
Immersion	Audio Quality	Binaural vs. Stereophonic vs. Monophonic	[131]	$N = 82$, presence: $binaural^+$, $mono^-$, $!2 = (4, N = 79) = 10.7$, $p = 0.031$
Immersion	Audio Quality	Attention, Binaural	[149]	active perception (visuo-auditory, vestibular emulation, Bayesian models), $f = 6$–10 Hz.
Immersion	Display	Face-to-Face Point of View	[26]	3D capture, maintain face-directed gaze through robot positioning, $f = 2.12$ Hz
Immersion	Display	Three 55-inch Screens vs. One 55-inch Screen	[7]	$N = 144$, copresence: $\mu^+_{humansize\_display} = 3.94$, $\sigma_{humansize\_display} = 1.46$; $\mu_{smallsize\_display} = 3.17$, $\sigma_{smallsize\_display} = 1.30$; $F(1, 140) = 11.41$, $p < 0.001$, $\eta^2_p = 0.08$
Immersion	Display	Autostereoscopic Telepresence	[309]	3D capture, 3D display, eye/head tracking, frame rates: 34, 48, 74 Hz
Context	Personality or Traits of Virtual Human	Personality manifested by voice and match between content	[278]	$N = 144$, computer voice with a personality (extrovert/introvert) similar to human interlocutor, $F(1, 67) = 11.13$, $p < 0.001$, $\eta^2_p = 0.14$; $voice^+_{extrovert}$, $voice_{introvert}$, $F(1, 71) = 17.91$ $p < 0.001$, $\eta^2_p = 0.20$

Table 6.3: *Cont.*

Predictor Category	Predictor	Evaluation Process	Study	Quantitative Comparison (Statistics)
Context	Agency	Avatar vs. Agent	[371]	$N = 134$, copresence index: $agency_{human\_human\_interaction} \approx agency_{human\_computer\_interaction}$, $R = 0.03$, $F = 0.15$, $p = 0.7$;
Context	Agency	Conscious experience of being someone	[74]	Illusory self-identification
Context	Agency	Avatar vs. Agent	[35]	$N = 90$, $agency^{+}_{human\_human\_interaction}$, $agency_{human\_computer\_interaction}$, $F(1, 90) = 10.870$, $p = 0.001$, $\eta^2 = 0.112$
Context	Physical Proximity	Close vs. Distant (spatial proximity)	[240]	$N = 134$, male social presence: $std_path_est location\_accessibility\_cues^{+} = 0.21$, $std_path_est^{+}_{richer\_medium} = 0.06$
Context	Task Type	Caregiver: Human vs. Robot	[256]	$N = 60$, social presence: $\mu^{+}_{robot\_as\_caregiver} = 5.56$, $\sigma_{robot\_as\_caregiver} = 1.04$, $\mu_{human\_as\_caregiver} = 4.20$, $\sigma_{human\_as\_caregiver} = 0.83$;
Context	Social Cues	Online Buddy: Present vs. Absent Telepresence	[254]	
Context	Identity Cues	Robot, Identity Cues: High vs. Low	[103]	
Individual	Demographic Variables	Gender: Female vs Male	[125]	
Individual	Psychological Traits	Communication Apprehension	[110]	
Individual	Psychological Traits	Belonging Feeling	[256]	

6.3 From Telepresence to Co-Presence Design

Presently, the market [490] offers full solutions for mobile robotic telepresence (MRP) systems [41, 202, 221, 228] (see Table 6.4), and the research presents telepresence robot solutions such as the ones listed in Table 6.5. There are also unmovable robotic telepresence (RP) systems, which are listed in Table 6.6. These robotic telepresence systems are depicted in Figures 6.3 and 6.4.

Table 6.4: Mobile robotic telepresence (MRP) systems: full market solutions.

References	Robotic Telepresence Systems	Application Area	Expression or Manipulation	Navigation Features	Cost
[384]	Giraff	Eldery	Head tilt (screen display/ camera)	No	USD 11,900.00
[133]	Double 2, 3	Office, education, hospital	Motorized height	Accelerometer and gyroscope for balance, kickstands when static	USD 2749.00
[223]	PadBot 2	Office, education, hospital	Tilt head (screen)	Obstacle detection, collision avoidance, anti-falling system	USD 1297.00
[223]	PadBot U1—v2	Office, education, hospital	Tilt head (screen)	Collision-prevention sensors. Edge detection and anti-falling sensors.	USD 797.00
[223]	PadBot T1	Office, home	No	Collision prevention, cliff sensor	USD 185.00
[374]	Beam Pro	Corporate, manufacturing, medical	No	Crash avoidance, assisted driving	USD 14,945.00
[418]	Ava 500	Healthcare, office	No	2D or 3D imaging, sonars and lasers for autonomous navigation, omnidirectional navigation, scheduling capabilities, cliff sensor	USD 32,000.00
[379]	Ohmni SuperCam	Office, home	No	Includes downward-facing camera for full visibility	USD 2195.00
[503]	VGo	Office, education	Tiltable head	Crash avoidance, notification of obstacles locations, cliff sensor	USD 3995.00
[316]	TeleMe 2	Office, education	Laser pointer option	Crash avoidance: infrared sensors detect obstacles and will automatically reduce robot's speed	USD 3995.00
[226]	RP-Vita	Healthcare, FDA clearance	Active patient monitoring	Obstacle avoidance, omnidirectional and autonomous navigation	USD 80,000.00
[40]	Teleporter	Office, factory, hospitals	Laser pointer, secondary webcam	Crash avoidance: infrared, 3D, or sonar sensors	USD 14,995.00

Table 6.5: Mobile robotic telepresence (MRP) systems: research-oriented solutions.

References	Robotic Telepresence Systems	Application Area	Expression or Manipulation	Navigation Features	Cost
[396]	PRoP	Research	Laserpointer, 2 DOF hand and arm	-	-
[169]	FURo-i	Home	No	Bumpers	USD 1800.00
[5]	MeBot	Research	3 DOF neck for screen and 3 DOF arms	Collision prevention, cliff sensor	-
[383]	Origibot	Research	Tiltable head for screen, 1 DOF arm (180°), 2 DOF gripper	No	Low cost
[468]	Nao	Research	Humanoid, 25 DOF, tiltable head, arms, legs, 4 directional microphones and speakers, 2 cameras	No	–
[468]	Pepper	Research	DOF (head: 2, shoulder: 2, elbow: 2, wrist: 1, hands (5 fingers): 1, waist: 2, knees: 1, base: 3 wheels), 2D and 3D cameras and sonars	Autonomous navigation, bumpers	USD 30,000.00
[320, 404]	GrowMeUp	Eldery research	Robot expression, directional microphones and speakers, 2D and 3D cameras, sonars, touchscreen	Obstacle avoidance, autonomous navigation, service, expression, behaviors	-

Table 6.6: Unmovable robotic telepresence (RP) systems.

References	Robotic Telepresence System	Application Area	Expression or Manipulation	Cost
[224]	Kubi	Office, education	Pan 300°, tilt 900°(screen)	USD 675.00
[317]	TableTop TeleMe	Office, education	Pan 360°, tilt head (screen)	USD 3995.00
[445]	SelfieBot	Office, education	Pan 180°, tilt 180° head (screen)	USD 195.00
[386]	Meeting Owl Pro	Office	Static 360° camera (1080p)	USD 999
[326]	Robovie mR2	Research	Expression, arms, gestures, eye blinking (cameras), 18 joints (3 in each eye), 18 servo motors	-

6.3.1 Co-Presence Design

Mechanisms that contribute to enhancing co-presence in telepresence robots should consider the robot-side systems and the remote user (robot's operator) side solutions. Robot-side interfaces support interactions between the robot and the local user (bystander) and between the robot and the remote user (opera-

tor). Human–robot interfaces can be classified into sight, hearing, touch, and body-sensing technologies. Technological advances include robust robot sensory (vision, face and expression recognition, object recognition, activity identification, pressure, touch, temperature, speech understanding, sound localization, etc.), acting (mobility, proxemics, gestures, gazing, facial expressions, speech synthesis, etc.), reasoning (localization, planning, context awareness, grasping, etc.), and appearance (familiar, unfamiliar, human-like, and mechanical) [183, 184, 496].

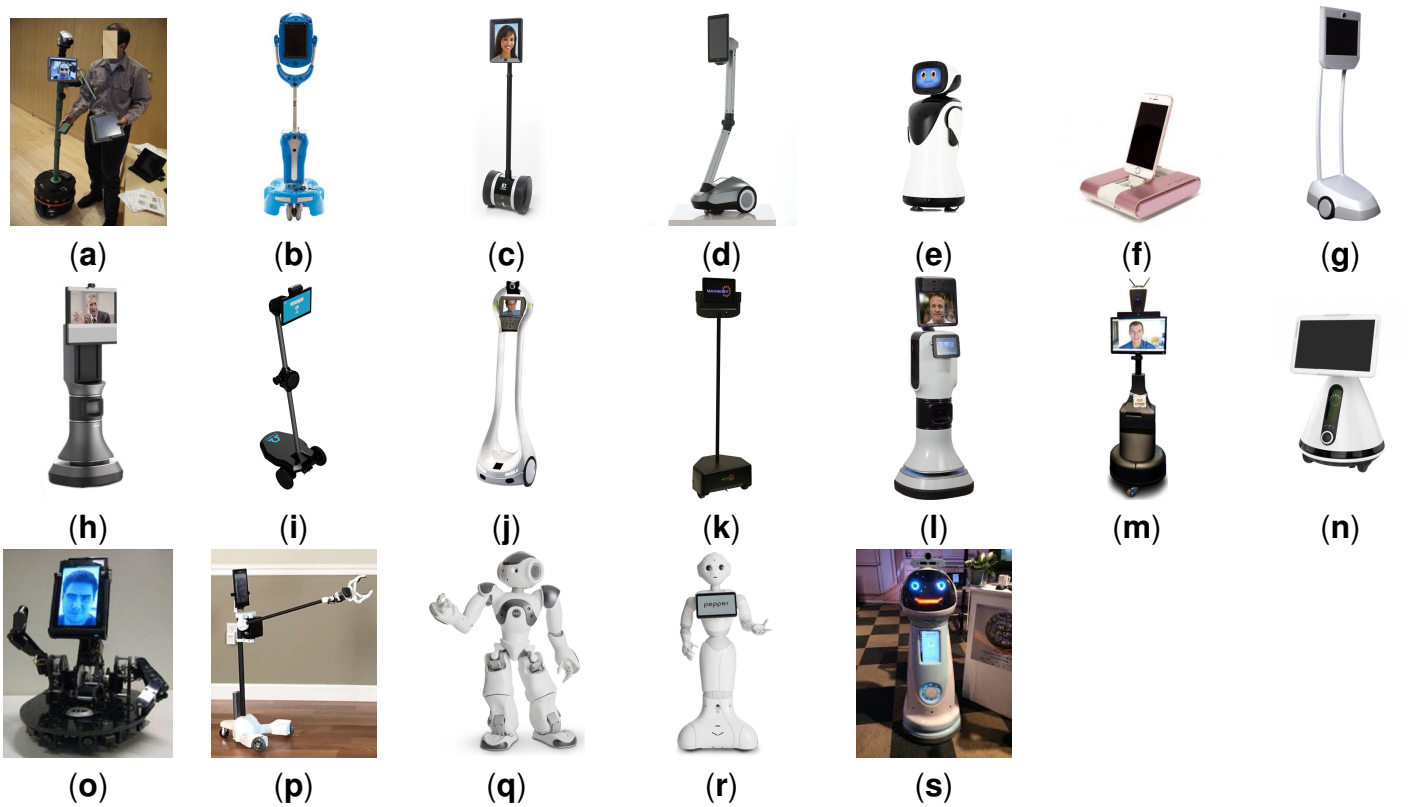


Figure 6.3: Mobile robotic telepresence (MRP) systems: (**a**) PRoP, (**b**) Giraff, (**c**) Double 2, 3, (**d**) PadBot 2, (**e**) PadBot 3, (**f**) PadBot T1, (**g**) Beam Pro, (**h**) Ava 500, (**i**) Ohmni SuperCam, (**j**) VGo, (**k**) TeleMe, (**l**) RP-Vita, (**m**) Teleporter, (**n**) FURo-i, (**o**) MeBot, (**p**) Origitbot2, (**q**) Nao, (**r**) Pepper, and (**s**) GrowMeUp.

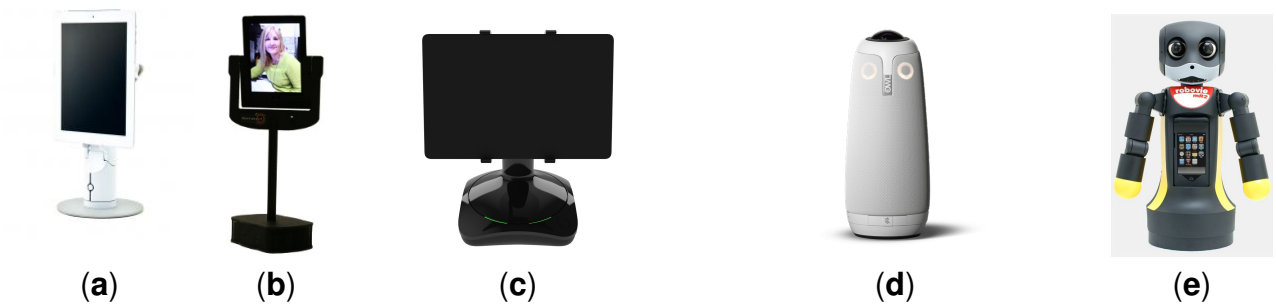


Figure 6.4: Unmovable robotic telepresence (RP) systems: (**a**) Kubi, (**b**) TableTop TeleMe, (**c**) SelfieBot, (**d**) Meeting Owl Pro, and (**e**) Robovie mR2.

Sensing

Robotic sensing technologies are becoming more efficient, lighter, and cheaper. Early human–robot interfaces used to integrate few sensors and relied mainly on video and audio data and low-resolution proximity sensors (e.g., sonars). Current robots can be equipped with 3D or 2D cameras (e.g., low-cost RGB-D cameras), pressure sensors, touch sensors, directional sound sensors (arrays), high-precision proximity sensors (e.g., range laser finder (lidar)), and robot pose and position sensors (e.g., gyroscopes, accelerometers, and GPS). The fusion of these sensors combined with high-accuracy robots, person localization algorithms (e.g., simultaneous localization and mapping (SLAM) or Open Pose), and deep learning approaches, have improved robot operations in an environment, enhancing HRI between operators and bystanders. Valuable information can be extract due to advances in sensor technologies and software, such as sound locations [149, 152, 215], speech segregation [420, 509] and recognition [33, 359], attention [275], gesture recognition [293, 473], human action analysis [59, 235], human intentions [246, 409], object recognition [425], and scene understanding [285, 523].

Action Capabilities

Advances in robot software and hardware, lighter and stronger materials, component miniaturization, and lighter and more powerful batteries have broadened robots' capabilities. Robot mobility has improved significantly, enabling robust navigation in an unstructured environment and in rough terrain [137, 175], and they can climb stairs, walk fast, and run, such as the Boston Dynamics ATLAS robot [136] or Honda ASIMO Robot [214]. Advances in humanoid mobility and equilibrium are remarkable, including compliant interactions and variable speed [220]. Having arms, hands, and fingers with more degrees of freedom (DOFs) enabled new types of interactions such as high-fidelity gestures, grasping objects smoothly [430], or even open doors and pass though them [136, 188]. Whole-body expressive movements [502], facial features to support expression synthesis [3], and speech synthesis technologies are enabling better HRIs.

Reasoning

Robots are designed to perform several tasks, but task execution is not always perfect (e.g., motion constraints, impaired sensory, control, and communication delays). Thus, advancements in software reasoning processes have been developed to supervise tasks, aiming not for perfect execution but optimum performance. Namely, notable advances have been made in localization and mapping [349, 478] and in grasping [430]. In [391], the authors explored approaches for a telepresence robot to detect and position itself with a group of people for social interactions (maintaining an egocentric perspective). The inclusion of these autonomous algorithms can help operators and bystanders in their interactions with the robot, simplifying the control, reducing the effort, and improving the intu-

itiveness.

Appearance

The acceptability of the robots enrolled in a human assistive task also depends on their appearance. Designers have created robots with human-like appearances [230]. The Geminoid robot has an incredibly realistic head and facial features [432]. This approach enables more effective communication through facial expressions and natural gestures. Additionally, given the human-like robot morphology, it is simpler to map the human gestures and movements in the robot. The search for realism, however, suggests some warnings regarding Mori's "uncanny valley" [58, 351]; that is, if a robot or agent is an imperfect replica of a human being, people may feel defrauded in their expectations regarding the affinity as a pair, triggering strange, familiar feelings of unease and revulsion.

Managing Robot Autonomy in Telepresence Systems

Advances in robot autonomy do not eliminate the role of human operators. Human skills remain crucial in an unstructured environment or when dealing with unpredicted events. The integration of autonomous mechanisms aims for process simplification, and it changes the nature of human–robot interaction (HRI). However, there are cases where the complexity increases [476] (e.g., 2019 Boeing 737 Max autopilot problems with deadly consequences). The availability of automated behaviors for telepresence or in humanoid robots may lead people to use them indiscriminately, diverting attention from the interaction essentials. Nevertheless, autonomous mechanisms aim to reduce users' mental workload, performing increasingly complex tasks and now being part of our daily lives (e.g., self-driving cars, autonomous vacuum cleaners, and chatbots). The literature refers to methods to integrate autonomy mechanisms in telerobotics [287], and they can be classified into direct control, supervisory control, shared control, traded control, collaborative control, and cooperative control [105, 455].

Direct control: The robot has no autonomy. An operator controls all the robot's functions manually. Mirroring is a type of direct control in which the robot replicates the human's movements and expressions.

Supervisory control: The robot is programmed intermittently according to the continuous information received from the robot. The human and the robot integrate a closed control loop focused on task performance [449].

Shared control: The human operator controls the robot continuously. However, those commands may be strictly followed by the robot (similar to direct control) or be modified by the robot's system to improve performance or run safely.

Collaborative control: The operator and the robot work together as peers to determine the robot's behavior. In Fong's work [160], there is an explicit semantic dialogue between humans and robots to mediate the sharing of control.

Traded control: The human operator starts a behavior or task that is autonomously

performed by the robot. At any time, the operator can stop that behavior or task and start a new one.

Cooperative control: The behavior of a single robot results from the controlled cooperation of several operators using any of the aforementioned methodologies.

In the shared control method, the operator provides continuous commands to the robot, aiming for high-level behavior from the robot. However, the robot may change those inputs to reach the perceived system goals [32]. The method assumes that the operator knows how to direct the robot's high-level behaviors but may not be sufficiently skilled to express the right commands due to a lack of situation awareness, embodiment, telepresence, or lack of robot motor accuracy and sensor information. Typically, the shared control method includes "safeguard" mechanisms, in which the operators' command actions are overwritten if they violate the robot's safety rules, such as collision with a wall or person or losing balance [161]. The software of HPR-1s or HRP-5P humanoid robots was developed to discard commands that could make the robot lose balance, limiting joint angles [243, 454, 471]. In Almeida et al. [23], given the robot's height, the wheel's initial acceleration provided by the operator had to be supervised by the robot to avoid its falling down. In Crandall and Goodrich's works, the robot's desired trajectory was provided through a joystick as the intended general direction and not as low-level position commands [113].

In the traded control method, a task or subtask is performed autonomously by the robot, but it is initiated by the operator and may be stopped at any time. The method is useful for simultaneously controlling multiple robot's appendages, such as in teleoperation of humanoids [200, 303, 431]. The Geminoid HI-1 robot [431] relies on a traded control known as state-based control, in which the operator selects the state from a library of states. It includes five conscious behaviors, namely right looking, left looking, listening, speaking, and being idle. For each state, the robot assumes autonomous behaviors (i.e., motion files), avoiding an explicit operator's control of 50 robot actuators. The integration of multiple semi-autonomous mechanisms is essential while controlling the eyes, head, torso, arms, hands, and fingers simultaneously in a humanoid robot. Quite often, operators need to control low-level robot behaviors and additionally focus their attention on high-level tasks, such as (1) robot navigation, (2) obstacle avoidance, (3) triggering robot's unconscious and conscious behaviors [431], (4) object and scene understanding [176, 285, 337], (5) mission planning, or (6) people's interaction. Osawa et al. [385] evaluated the automation of involuntary and voluntary movements using a teleoperated telepresence robot (robovie-mR2). The implemented behavior generation architecture (bi-layered architecture [488]) enabled the combination of autonomous movements and manual movements controlled by a remote operator. The results showed that bystander users evaluated both the involuntary and voluntary movements positively but also revealed that from the remote operator's point of view, the automation of voluntary movements should require additional care (agency issue conflicts).

Time Delay Mitigation

The dynamic nature of the communication medium has an impact on the complexity of teleoperated systems. Time delay, jitter, distance, bandwidth constraints, packet loss, or blackout in internet-based solutions can delay or distort interactions. This can affect the synchronicity (rate of message exchange between operator and bystander), compromising the social presence [116] (e.g., degradations in audio and video streams, control streams, or haptic feedback). Traditional methods to mitigate time delay in telerobotics involved user interface design and control theory-based models (e.g., supervisory control or passivity-based teleoperation) and evolved into predictive displays and control [449]. Recent solutions for time delay issues use *time series prediction* methods to predict the time delay, robot movements, and user intentions (e.g., user's gaze prediction [36]). These new adaptive-based control methods make use of nonlinear statistical models and neural network (NN) or machine learning (ML) techniques.

Ferrell and Sheridan [153] determined that a time delay affects human operators' performance while teleoperating manipulators. They realized that the person within the control loop of teleoperated systems under time delays used to adopt a *move-and-wait* strategy to accomplish certain tasks. To address this problem, they proposed supervisory control [154], in which the robot is preprogrammed or programmed online to perform certain subtasks autonomously. By transmitting only high-level commands, there is a data communication reduction, and task time completion improves. Meanwhile, several extensions of supervisory control were developed, including specific languages to chain tasks or predictive displays (i.e., visualization of a phantom robot model that predicts the motion of the real robot) [60, 61].

Control-based approaches for time delay mitigation in teleoperation systems can be clustered into two classes [145, 498]: (1) *predictive control-based methods* (e.g., a discrete linear quadratic Gaussian (LQG) controller for teleoperation acting on the sampling rate or output feedback control of multiple-input and multiple-output (MIMO) systems) and (2) *passivity-based methods* that model the master–slave operator systems and unsure stability and performance under time delay variability (e.g., a two-port network, hybrid matrix, impedance matrix, constant time delay, scattering approach, wave variable, scaling, and geometric scattering).

Time series prediction approaches for time delay mitigation in teleoperation systems try to compensate for the time delay, observing past intrinsic patterns to predict the future values [145, 287]. They integrate trends, seasonality, and white noise and can be clustered into two types: *statistical methods* and *neural network (NN) or machine learning (ML) methods*:

(1) *Statistical methods* (e.g., moving average (MA), linear auto-regression (AR), auto-regression + moving average (ARMA), and auto-regression + moving average + nonlinear component (ARIMA) [301]);

(2) *NN or AI methods* (e.g., recurrent neural networks (RNNs) [342, 477], long short-term memory networks (LSTMs) [210, 292], sequence to sequence (Seq2Seq) [318, 482], and generative adversarial networks (GANs) [182, 533]).

Statistical methods have the advantage of not requiring training with data and are simpler to implement. Although times series prediction traditionally relied on statistical approaches, it has difficulties in modeling the entire set of nonstationary signals. Nevertheless, methods like ARIMA can cope with nonstationary signals. Statistical methods are not appropriate for modeling complex tasks, being more suitable for short-term predictions. Neural networks, on the other hand, have an advantage over statistical approaches in that they enable data description without explicit knowledge of its distribution and can model more complex time series data based on past observations. Neural networks are more prone to adapt their behaviors as the input data increases [145].

6.3.2 User-Adaptive Systems Taxonomy

Social robots aim to assist people, enable telesurveillance of elderly people, guide people on tours, promote physical and mental exercise, keep company, or entertain [41, 228, 451]. In short, they contribute to the user's well-being, adapting to people, to the environment, and ultimately to the context. Case studies include interaction of a service robot for 1 week in an elderly care center [404], or telepresence gaze-controlled robots accessible to persons unable to use their hands because of a motor disability [532]. Several types of user-adaptive mechanisms are described in the robotics literature [8, 145, 202, 203, 321, 368].

Typically, a framework for a user-adaptive system comprises two components (see Figure 6.5): the *interface* layer that is used for the exchange of information between the user and the system (It integrates sensors for the system to perceive the user and actuators to provide stimuli.) and the *decision-making module* which, based on perceived information, makes algorithmic decisions and generates response actions to be synthesized by the interface.

Robot systems with autonomous and semi-autonomous behaviors can be classified with the following taxonomy [321]:

Autonomous and semi-autonomous behaviors supported by	- Adaptive systems with no user model - Systems based on static user models - Systems based on dynamic user models

1. Adaptive systems with no user model: systems with reactive behavior regarding the user's immediate feedback and with no cache of the user's information (see Figure 6.6);
2. Adaptive systems based on static user models: systems that rely on preloaded knowledge retrieved from the relevant attributes of the user and used to adjust the system's behavior (see Figure 6.7);
3. Adaptive systems based on dynamic user models: similar to the previous example, these systems explicitly maintain user models. They are task-oriented models, updated with users' information during their interactions (see Figure 6.8).

User-adaptive systems require information about the user which is typically stored in the form of a user model [202, 369]. As reported in an early survey [330], a new field of research emerged concerning with acquisition, organization, and representation of the system's user.

Adaptive systems without user modeling can implicitly map the characteristics of a generic user in the architecture of the decision-making module (Figure 6.6). Nevertheless, it is a reactive adaptation that shapes the system's behavior directly based on the user's feedback. The user's behavior changes are monitored and trigger an immediate switch to a new system's operational state, while no storage or user model update is performed. Table 6.7 summarizes several works that adopt this type of architecture.

Adaptive systems based on static user models assume that the person's profile does not evolve during the interaction. These static models can be built during an initial phase of the interaction (Figure 6.7), similar to the calibration process, or the user's profile can be pre-supplied using external questionnaires. These types of systems are not able to dynamically learn the characteristics of the user. Examples of related works are listed in Table 6.8.

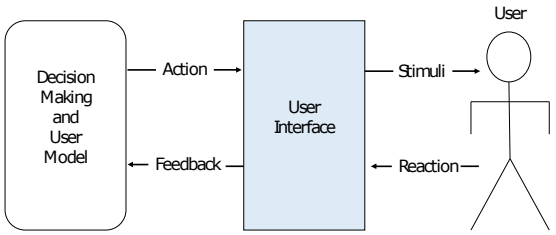


Figure 6.5: Overview of a generic user-adaptive system, which includes a user interface layer and a decision-making module.

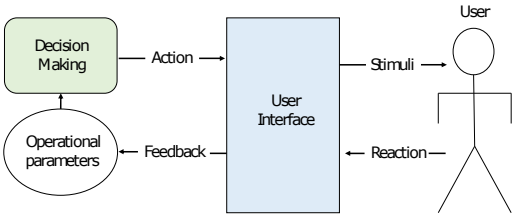


Figure 6.6: General schematic of a user-adaptive system without the user's model. The system's behaviors are direct reactions to the user's feedback, and decisions are made without the user's previous knowledge.

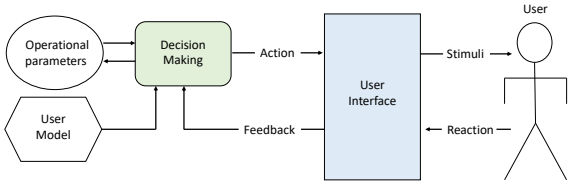


Figure 6.7: General schematic of a user-adaptive system based on static user models.

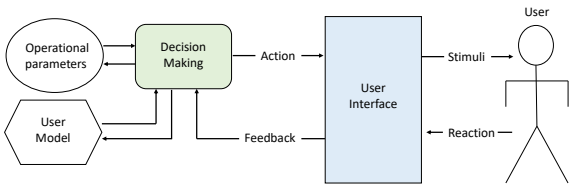


Figure 6.8: General schematic of a user-adaptive system based on dynamic user models. The user's feedback reactions are used to continuously update robot knowledge and consequently tune the system's behavior.

Adaptive systems based on dynamic user models perceive, learn, and update the knowledge regarding the context model and the user model. The stored user model is updated during the interaction based on the user's reactions. This category of systems is considered the best performing user-adaptive solution, although its implementation is more complex [13, 369, 424]. Table 6.9 compiles several references for systems based on dynamic user models.

Additionally, one of the described categories, such as adaptive systems based on dynamic user models, can coexist in a telepresence teleoperated robot [385, 451, 488, 496], thus adding adaptiveness functionalities either for the robot's operator (remote user) or for the local user that is with the robot. The general architecture of a teleoperation system with user adaptiveness is depicted in Figure 6.9.

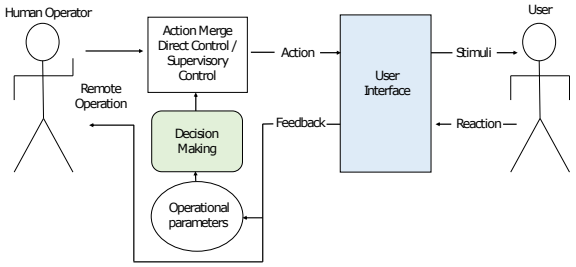


Figure 6.9: A schematic of the general architecture of a teleoperation system that includes an adaptive system.

The decision-making modules of the listed user-adaptive systems include different frameworks, such as the Markov decision process (MDP), partially observable Markov decision process (POMDP) (αPOMDP [319]), Mixed Observability Markov Decision Processes (MOMDP), fuzzy control, rule-based, hidden mode stochastic hybrid system (HMSHS), Bayes-adaptive, dynamic factor graph (DFG), active leaning, or reinforcement learning. Recent approaches for these frameworks are described in [521].

Table 6.7: Adaptive parameters, input modalities, framework of decision, output modalities, and social robot evaluation with no user model.

Adaptive Parameters	**Study**	**Input Modality**	**Decision Making**	**Output Modality**	**Evaluation Process**	**Evaluation Metrics**
Robot's Navigation Goal	[300]	Brain-actuated controls	Rule-based	Robot commands	Measurements, questionnaires,	Robot path
Decisions $takeleftturn$	[88]	Physical controls	POMDP	Image and sound	Measurements, questionnaires	POMDP rewards, perceived control, driving performance, similarity to real world, naturalness, social appropriateness
Robot Speed	[155]	User's pose and speed	MOMDP	Motor control	Measurements	Speed difference and distance to the user
Decisions (object to move)	[219]	Speech, gaze	Rule-based	Robot arm movement	Measurements, questionnaires	Prediction accuracy, projection accuracy, perceived awareness, response time and intentionality
Robot Speed	[102]	Odometry, Physical controls	Fuzzy control	motor controls	-	-
Decisions (warn driver or intervene)	[271]	Physical controls	Hidden mode stochastic hybrid system	Image and sound	Measurements	Time in unsafe and safe states
Decisions (room to clean)	[253]	User locations, task success	Motor control	Rule-based	-	-
Robot's Navigation Goal	[325]	Physical controls	Rule-based	Robot commands	Measurements	Recognition accuracy
Voice Pitch	[308]	User speech	Rule-based	Robot speech	Questionnaires	Persistence and learning gain, rapport, perceived social presence

Table 6.7: *Cont.*

Adaptive Parameters	**Study**	**Input Modality**	**Decision Making**	**Output Modality**	**Evaluation Process**	**Evaluation Metrics**
Robot's Gestures	[448]	Vision, speech	Rule-based	Robot commands	Measurements, questionnaires	Information distance, perceived behavior performance, perceived gesture recognition, enjoyment, perceived social interaction
Robot Speed and Path	[353]	Physical controls	Rule-based	Robot commands	-	-
Decisions (navigation goal)	[436]	Physical controls	POMDP	Robot commands	Measurements	State variables, robot path, destination probabilities
Decisions (what objects to move, when to speak)	[466]	Speech, vision, depth	Rule-based	Speech, robot commands	Measurements	User's speech time

Table 6.8: Adaptive parameters, input modalities, framework of decision, output modalities and social robot evaluation with static user model.

Adaptive Parameters	Study	Input Modality	Decision Making	Output Modality	Evaluation Process	Evaluation Metrics
Robot's Gestures and Speech	[29]	Speech	Rule-based	Gestures, speech	Questionnaires	Preference toward a type of adaptation
Decisions (placement of objects)	[2]	Crowd-sourced data	Rule-based	Robot controls	Measurements	F-scores
Robot Location, Interface Complexity, Warning Levels, Font Size	[134]	-	Rule-based	-	-	-
Decisions (how to dress the user)	[259]	User's pose, speech	Rule-based	Robot commands	Measurements	Task completion speed
Decisions (how to dress users)	[171]	User's pose, speech	Rule-based	Robot commands	Measurements	Classification accuracy
Speech Output Gender, Sound Volume, Robot's Name, Robot Speed	[157]	Speech, touch	Rule-based	Robot commands, speech	Questionnaires	Acceptance, perceived usability
Sequence of dance movements	[422]	User's pose	Rule-based	Robot commands	Questionnaire, manual classification	Gaze position, body language, facial emotion, perceived bond, amusement, satisfaction, enjoyment, anxiety, observed leadership, expectancy

Table 6.9: Adaptive parameters, input modalities, framework of decision, output modalities, and evaluation of the social robots with a dynamic user model.

Adaptive Parameters	**Study**	**Input Modality**	**Decision Making**	**Output Modality**	**Evaluation Process**	**Evaluation Metrics**
Promote Regular Physical Activity Habits	[335]	User's position, pose (exercise performance), speech	Rule-based	Navigation, robot commands, speech (avatar coach)	Measurements, questionnaires	User's exercise performance, flow
Decisions (service, navigate, turn to the person, stop, smile)	[319, 404]	Person detection, speech, emotion recognition, touchscreen	αPOMDP, SOA-based model	Navigation, robot commands, approach, speech, robot expression, recognition, service	Measurements, questionnaires	Usability, appearance, satisfaction
Human Robot Greetings Phase	[96]	Tracking human gestures	MDP	Kendom phase trigger (initiate approach, distance salutation, head dip, approach, final approach, and close salutation)	Measuments, observation	Sequence estimation accuracies
Colors of LEDs	[54]	Physical controls	Rule-based	LED colors	Measurements	Cumulative reward from users, error estimation
Decisions (what interactions to perform with the user)	[446]	Physical controls robot	Rule-based	Commands	Measurements, questionnaires	Child learning rate, human intervention ratio
Reading Difficulty Level	[185]	Speech, touch	Active learning	Number of words learned	Measurements, questionnaires	Images, speech
Decisions (adaptation to user's subtask selection)	[129]	Vision, speech	Rule-based	Robot commands, speech	Measurements	Number of communications required of the user

Table 6.9: *Cont.*

Adaptive Parameters	Study	Input Modality	Decision Making	Output Modality	Evaluation Process	Evaluation Metrics
Decisions (moments to take action, including parameter adjustment and services)	[245]	Gesture sound, projected mages	MDP	Questionnaires	Perceived coherence, user satisfaction, ease of use, perceived helpfulness, originality, perceived adaptivity	
Decisions (when to deploy services)	[190]	Speech, touch	Equilibrium maintenance	Speech, images, robot commands	Measurements	Opportunity relevance for the selected service
Decisions (dialogue to play)	[355]	Tactile sensors, sound, touch	Dynamic factor graph	Image, speech, robot commands	Questionnaires	User's opinion
Decisions (select learning content type)	[288]	Speech, physical controls	Rule-based	LEDs, robot commands	-	-
Decisions (sounds to play)	[435]	Physical controls	Context-free stochastic grammars	Sound (music)	Measurements, questionnaires	Engagement, perceived difficulty, progression, conformity, number of user interventions, speed
Decisions (placing a shared object)	[366]	Vision, physical controls	MAMDP	Robot commands	Measurements, questionnaires	Perceived trustworthiness, ratio of users that change strategies
Decisions (positive, negative, or neutral output)	[44]	Facial expressions, RGBD, electrodermal data, touch screen	Rule-based	Images, speech, gestures	Questionnaires	Understanding, perceived enjoyment, trust
Decisions (where to guide the user)	[444]	Vision, user's attention, robot position, odometry, speech	Rule-based	Robot commands, navigation	Questionnaires	User's opinion (score)

Co-presence mechanisms: the availability of robotic autonomous mechanisms enables a robot's voluntary or involuntary behaviors that contribute to enhancing co-presence [385], such as those listed in Table 6.10.

Table 6.10: Robotic mechanisms to enhance co-presence.

Type	**Voluntary**	**Involuntary**
Eye contact	X	-
Gaze following	X	-
Gazing at the closest face	X	-
Gazing at a random face	X	-
Gazing at the closest object	X	-
Gazing at a random object	X	-
Gazing at a moving object	X	-
Looking around the gazing position	X	-
Joint attention	X	-
Sleeping	X	-
Changing LED colors	X	-
Mouth movement	X	-
Nodding in response to human speech	X	-
Waving both hands at a random human	X	-
Waving left hand at a random human	X	-
Waving right hand at a random human	X	-
Waving left hand in response to palms	X	-
Waving right hand in response to palms	X	-
Waving both hands in response to palms	X	-
Reflexive blinking with eye movement	-	X
Spontaneous blinking	-	X
Avoiding objects at close range	-	X
Eye saccade	-	X
Breathing	-	X

6.3.3 Evaluation Methods

To assess co-presence, telepresence systems require objective and qualitative metrics. Quantitative measures may include physiological signals (such as heart rate, skin temperature, electrodermal activity (EDA), and skin conductance responses (SCRs) [86], eye scan patterns, electroencephalography (EEG), or functional magnetic resonance imaging (fMRI)) [82, 331, 457], as well as other metrics that are simpler to obtain, such as accuracy, time to perform a task, and the number of errors or communication delays. However, given the human factor and the psychological components of interaction, questionnaires remain essential tools. There are methodologies for measuring the *presence*, *social presence* or *co-presence*, and *flow state* of the users using technological devices [89, 193, 297, 400, 416].

Flow is a psychological state that people describe when they are fully engaged

in some events to the point of forgetting time, fatigue, and everything else but the activity itself [414, 429]. Table 6.11 lists the available questionnaires to measure the levels of presence, co-presence, immersion, and flow.

Table 6.11: List of questionnaires to assess presence, flow, and game.

Psychological Phenomena	**Questionnaire**	**Number of Questions**	**Ref.**
Presence	Slater, Usoh and Steed (SUS)	6	[499]
Presence	Temple Presence Inventory (TPI)	42	[299]
Presence	Igroup Presence Questionnaire (IPQ)	14	[439]
Presence	Sense of Presence Inventory (ITC-SOPI)	38	[280]
Presence	Presence Questionnaire, version 3 (PQ)	29	[516]
Presence	Networked Minds Social Presence Inventory (NMSPI)	34	[68, 196]
Presence	Multimodal Presence Scale (MPS)	15	[312]
Presence	Spatial Presence Experience Scale (SPES)	8	[199]
Flow	EduFlow Scale (EFS)	12	[208]
Flow	Flow Short Scale (FSS)	13	[141]
Flow	Reading Flow Short Scale	8	[492]
Game and Flow	EGameFlow (EGF)	42	[167]
Usability	Nielsen Norman Group	-	[364, 365]

Usability, testing and accessibility–Jakob Nielsen, one of the most active proponents of usability processes, referred to the following elements that comprise a definition of usability [363–365]:

1. Ease of use: the use of products or tasks should be natural and easily performed by the user.
2. Simplicity of learning: tasks and product features must be intuitive and present a logical and consistent sequence to simplify learning.
3. Improved reliability: levels of satisfaction and performance are increased when the action's results correspond to the user's expectations.
4. Reduction in errors: usability can be increased if designers attribute the errors to the product or task (rather than the user), redesigning it based on the user's feedback.
5. Enhanced user satisfaction: the user's satisfaction principle must guide all of the design process, making the product or task pleasing to use or perform.

In [6], a taxonomy of usability guidelines for the design of telepresence teleoperated robots (interaction effectiveness and efficiency, information presentation, interface visual design, robot surroundings and environment awareness, robot state awareness, and cognitive factors) is proposed. The usability testing process is an effective use of materials and time [263, 267, 364, 371] that should not be overlooked.

6.4 Conclusions

This work presented a survey of recent works, proposing the development of support for social robotic interactions with applications in health care, elderly assistance, guidance, or office meetings. It focused on enhancing social presence via telepresence robot mediation, in which a user should sense his or her remote interlocutor as being locally present with him or her. The research gathered knowledge to help roboticists design improved user- and environment-adaptive systems and technical methods that contribute to enhancing the sense of presence or co-presence. This literature review aimed to define social presence, identify autonomous "user-adaptive systems" for social robots, and propose a taxonomy for "co-presence" mechanisms. The referred works address robot sensing, perception, action, reasoning, appearance, automation, and cognitive approaches (e.g., statistics models and AI). Additionally, it presents an overview of social robotics systems and application areas and provides directions for telepresence and co-presence robot design, considering the actual and future challenges. Finally, some guidelines for the evaluation of these systems are left, having as reference face-to-face interactions. Based on survey findings in engineering and psychology, our future work includes the design of telepresence and co-presence robots that better emulate or interpret human subjective experiences.

Chapter 7

Conclusion

7.1 Summary of Thesis Achievements

This research explores means to induce the sense of telepresence in human-centered communications and remote robot teleoperations. Given that immersion aims at providing stimuli that illude the sensory system, the proposed solutions maintain the consistency between outside sensory feedback and inside sensory information (proprioceptive, vestibular), and the brain's cognitive models. The space and motion perception and, the consequent interactions with the mediated world (virtual or real) should be as natural as the user was there. This work showed that people can experience and perform actions in remote places, through a robotic agent having the illusion of being physically there. The sensation can be compelled through immersive interfaces; however, technological contingencies can affect human perception. Based on the human factors results of related works, we provide a set of recommendations for the design of immersive teleoperation systems aiming to improve the sense of telepresence for typical tasks (ex. Table 3.5). The mitigation of issues such as system latency, field of view, frame of reference, or frame rate contributes to enhancing the sense of telepresence. The presented evaluation methodology enables analyzing how perceptual issues affect task performance. By decoupling the flows of an immersive teleoperation system, we start to understand how vision and interaction fidelity affects spatial cognition. Task experiments with participants using traditional vs. immersive interfaces allowed quantifying the disturbance introduced by each component of the system. For example, taking as a reference a simple manual pick-and-place task, the introduction of a visual see-through HMD increased the time to perform it by 78%; the introduction of a manual gripper tool increased that time by 252%; and the combination of visual and tool mediation increased the overall time by 372%. Decoupling the flows of an immersive teleoperation system allowed a separate analysis of visual feedback disturbances (e.g., limited FOV) without the influence of other factors that affect the frame of reference for motor-action. Our findings show that misalignment between the frame of reference for vision and motor-action or the use of tools that affect the sense of body position or sense of body movement leads to a higher mental workload and has a higher effect on spatial cognition. Misalignment between kinesthetic and visual feedback increases the

mental workload and compromises the sense of telepresence and the embodiment feeling. The mental workload to control the proposed video feedback component is considerably lower (in the immersive interface); however, the combination of both requires a higher effort (i.e., active visual mediation plus tools). Thus, a recommendation is to keep activities at skill-based behaviour levels, where familiar perceptual signals are essential to lower cognitive effort.

The human role in the teleoperation control loop is fundamental because it is the operator who can decide, react, and adjust operations in the presence of noisy and incomplete data (especially in unstructured and unpredictable scenarios). This fact made human factor analysis an essential tool to design new teleoperation interfaces aiming simultaneously for better performances and decreasing the number of failures caused by operator faults. One recommendation is that the interface systems should be developed so that the operator (surgeon, pilot, or other) receives the necessary information to perform the task without the need to search for it in unusual places. Vital information should always be placed in a visible and in salient way so that the user can perceive it immediately.

Regarding immersive teleoperation, research has demonstrated that it is possible to generate the remote physical embodiment feeling by letting the user perceive the robot's structure as his/her own body. To evolve from teleoperation to embodied operation this research proposes a view transfer using an HMD (i.e., an egocentric controlled view in which the user will see what the robot can see), and the use of natural commands, that is implicit commands instead explicit ones. A key development of this research is the *cockpit concept* in which the user feels inside the robot, perceiving and acting naturally. To this end, a system has been developed to virtually place the operator on board the remote robot and let him/her do the driving tasks from there. The immersive system combines the images, obtained from an orientable camera on the robot, with virtual instruments. This combination is displayed to the user using an HMD, which tracks the user's head movements to modify the POV camera. Additionally, the system can be improved by adding to the scene the user body representation. The evaluation results showed that the immersive system was the one preferred by the users. Furthermore, when compared with the traditional interfaces, the use of this immersive system had a positive effect on teleoperation performance.

By exploring computer graphics, spatial audio, computer vision and reconstruction techniques were demonstrated the potential of inducing sensations of being physical in the *presence* of other people. Namely, regarding human-centered mediated communications, this research proposes a low-cost framework to support three-dimensional conferencing through augmented reality (AR) based on telepresence. It aims to achieve the real face-to-face meeting benefits, in which important social cues such as eye-to-eye contact establishment, gesture reconnaissance, body language or facial expressions are transmitted, (presently not supported by commodity conferencing technologies such as Zoom, Teams or Skype). The contribution is a free viewpoint system framework that synthesizes views of an online 3D reconstructed model dependent on the observer's point of view. The approach explores virtual view synthesis through motion body estimation and hybrid sensors composed of video cameras and a low-cost depth camera based

on structured light. The solution addresses the geometry reconstruction challenge from traditional video cameras array, that is, the lack of accuracy in low-texture or repeated pattern regions. We present a full 3D body reconstruction system that combines visual features and shape-based alignment. The modelling is based on meshes computed from dense depth maps to minimize processed data resulting in a global 3D mesh representation that is independent of the viewpoint. Research contributions include an incremental version of the Crust algorithm that efficiently adds new vertices to an already existing surface without having to recompute previously generated meshes and, a topological incremental reconstruction approach based on confidence measures that avoid redundant data information computation. With this online reconstructed 3D model, it is possible to provide a synchronous point of view for an observer that moves in front of a display of a face-to-face meeting application, thus enhancing the presence sensation.

Additionally, a wearable glove-based method was also developed for natural task execution in virtual interaction scenarios.

Moreover, this work presents a literature review on developments supporting robotic social interactions, contributing to improving the sense of presence and co-presence via robot mediation. It aims to gather knowledge to help roboticists design improved user- and environment-adaptive systems and technical methods. Reviews have addressed user-adaptive systems [2,8] and environment-adaptive systems [8] for social robotics (in which the robot is generally an autonomous agent serving the bystander user). However, we further explore telepresence social robotics, emphasizing the relationship between the robot's operator and the bystander user.

7.2 Future Work

Future work includes the evaluation of the traditional interface setup, considering the control of the remote camera orientation with a joystick, and the evaluation of the proposed immersive interface to control a robotic arm with haptic feedback. With this on-line reconstructed 3D model, we can provide synchronous point of view for an observer that moves in front of a display of a face-to-face meeting application, thus enhancing the presence sensation. Future work includes framework usability tests for a telepresence meeting application. The goal is improve the online incremental 3D reconstruction framework to be widely used on low-cost telepresence applications, augmented reality (AR) or human robot interaction applications. Additionally, the goal is to exploit and contribute to designing new teleoperation interfaces based on natural interaction while providing the synchronous position-movement sensation that eludes proprioceptive sense through technological means. Keep researching inducing the sensation of being there: in the presence of other people and in a remote environment through a robot.

References

[1] (2015). VRPN: Virtual reality peripheral network. `http://www.cs.unc.edu/Research/vrpn/`.

[2] Abdo, N., Stachniss, C., Spinello, L., and Burgard, W. (2015). Robot, organize my shelves! tidying up objects by predicting user preferences. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1557–1564. IEEE.

[3] Abdollahi, H., Mahoor, M., Zandie, R., Sewierski, J., and Qualls, S. (2022). Artificial emotional intelligence in socially assistive robots for older adults: A pilot study. *IEEE Transactions on Affective Computing*, pages 1–1.

[4] Abolmaesumi, P., Fichtinger, G., Peters, T. M., Sakuma, I., and Yang, G. Z. (2013). Introduction to special section on surgical robotics. *IEEE Transactions on Biomedical Engineering*, 60(4):887–891.

[5] 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.

[6] Adamides, G., Christou, G., Katsanos, C., Xenos, M., and Hadzilacos, T. (2014). Usability guidelines for the design of robot teleoperation: A taxonomy. *IEEE Transactions on Human-Machine Systems*, 45(2):256–262.

[7] Ahn, D., Seo, Y., Kim, M., Kwon, J. H., Jung, Y., Ahn, J., and Lee, D. (2014). The effects of actual human size display and stereoscopic presentation on users' sense of being together with and of psychological immersion in a virtual character. *Cyberpsychology, Behavior, and Social Networking*, 17(7):483–487.

[8] Akalin, N. and Loutfi, A. (2021). Reinforcement learning approaches in social robotics. *Sensors*, 21(4).

[9] Akbarzadeh, A., Frahm, J.-M., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Merrell, P., Phelps, M., Sinha, S. N., Talton, B., Wang, L., Yang, Q., Stewénius, H., Yang, R., Welch, G., Towles, H., Nistér, D., and Pollefeys, M. (2006). Towards urban 3d reconstruction from video. In *3DPVT*, pages 1–8. IEEE Computer Society.

[10] Al-Hathal, T. and Fetais, N. (2018). Virtual reality glove for falconry. In *2018 Int. Conf. on Computer and Applications (ICCA)*.

[11] Alabdulkareem, A., Alhakbani, N., and Al-Nafjan, A. (2022). A systematic review of research on robot-assisted therapy for children with autism. *Sensors*, 22(3).

[12] Albert, R., Patney, A., Luebke, D., and Kim, J. (2017). Latency requirements for foveated rendering in virtual reality. *ACM Trans. Appl. Percept.*, 14(4).

[13] Albrecht, S. V. and Stone, P. (2018). Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95.

[14] Aliakbarpour, H., Almeida, L., Menezes, P., and Dias, J. (2011). Multi-sensor 3d volumetric reconstruction using cuda. *3D Research*, 2:1–14. 10.1007/3DRes.04(2011)6.

[15] Aliakbarpour, H. and Dias, J. (2010a). Human silhouette volume reconstruction using a gravity-based virtual camera network. In *Proceedings of the 13th International Conference on Information Fusion, 26-29 July 2010 EICC Edinburgh, UK*.

[16] Aliakbarpour, H. and Dias, J. (2010b). Imu-aided 3d reconstruction based on multiple virtual planes. In *DICTA'10 (the Australian Pattern Recognition and Computer Vision Society Conference), IEEE Pr., 1-3 December 2010, Sydney, Australia.*

[17] Aliakbarpour, H. and Dias, J. (2011). Inertial-visual fusion for camera network calibration. In *IEEE 9th International Conference on Industrial Informatics (INDIN 2011), July 2011.*

[18] Almeida, L., Lopes, E., Yalçinkaya, B., Martins, R., Lopes, A., Menezes, P., and Pires, G. (2019). Towards natural interaction in immersive reality with a cyber-glove. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 2653–2658.

[19] Almeida, L., Menezes, P., and Dias, J. (2011a). Stereo vision head vergence using gpu cepstral filtering. In *VISAPP 2011 - Fifth International Conference on Computer Vision Theory and Applications*, pages 665–670. SciTePress, Vilamoura, Algarve, Portugal.

[20] Almeida, L., Menezes, P., and Dias, J. (2013). Augmented reality framework for the socialization between elderly people. In *Handbook of Research on ICTs for Human-Centered Healthcare and Social Care Services*, pages 430–448. IGI Global.

[21] Almeida, L., Menezes, P., and Dias, J. (2015). Incremental Reconstruction Approach for Telepresence or AR Applications. In Dias, P. and Menezes, P., editors, *22o Encontro Português de Computação Gráfica e Interaçãoo 2015*. The Eurographics Association.

[22] Almeida, L., Menezes, P., and Dias, J. (2016). 3D Modelling Framework: an Incremental Approach. In Magalhaes, L. G. and Mantiuk, R., editors, *EG 2016 - Posters*. The Eurographics Association.

[23] Almeida, L., Menezes, P., and Dias, J. (2017). Improving robot teleoperation experience via immersive interfaces. In *2017 4th Experiment@International Conference (exp.at'17)*, pages 87–92. IEEE.

[24] Almeida, L., Menezes, P., and Dias, J. (2020). Interface transparency issues in teleoperation. *Applied Sciences*, 10(18).

[25] Almeida, L., Menezes, P., and Dias, J. (2022). Telepresence social robotics towards co-presence: A review. *Applied Sciences*, 12(11).

[26] Almeida, L., Menezes, P., Seneviratne, L., and Dias, J. (2011b). Incremental 3d body reconstruction framework for robotic telepresence applications. In *Robo 2011: The 2nd IASTED International Conference on Robotics*. Pittsburgh, USA.

[27] Almeida, L., Patrão, B., Menezes, P., and Dias, J. (2014). Be the robot: Human embodiment in tele-operation driving tasks. In *Ro-Man 2014: The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 477–482, Edinburgh, UK.

[28] Almeida, L., Vasconcelos, F., Barreto, J., Menezes, P., and Dias, J. (2011c). On-line incremental 3d human body reconstruction for hmi or ar applications. In *CLAWAR 2011: 14th International Conference on Climbing and Walking Robots and the Support Technologies for Mobile Machine*. Paris, France.

[29] Aly, A. and Tapus, A. (2013). A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 325–332. IEEE.

[30] Amenta, N., Bern, M., and Kamvysselis, M. (1998). A new voronoi-based surface reconstruction algorithm. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '98, pages 415–421. ACM, New York, NY, USA.

[31] Amin, M. S. (2016). Vestibuloocular reflex testing.

[32] Anderson, R. (1996). Autonomous, teleoperated, and shared control of robot systems. In *Proceedings of IEEE International Conference on Robotics and Automation*, volume 3, pages 2025–2032 vol.3.

[33] Anusuya, M. and Katti, S. (2011). Front end analysis of speech recognition: A review. *International Journal of Speech Technology*, 14:99–145.

[34] AP4ISR (2022). Artificial perception for intelligent systems and robotics - ap4isr. `https://ap.isr.uc.pt/`. (Last Accessed: november 2022).

[35] Appel, J., von der Pütten, A., Krämer, N. C., and Gratch, J. (2012). Does humanity matter? analyzing the importance of social cues and perceived agency of a computer system for the emergence of social reactions during human-computer interaction. *Adv. in Hum.-Comp. Int.*, 2012.

[36] Arita, R. and Suzuki, S. (2019). Maneuvering assistance of teleoperation robot based on identification of gaze movement. In *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*, volume 1, pages 565–570.

[37] Arthur, K. W. (2000). *Effects of field of view on performance with head-mounted displays*. PhD thesis, University of North Carolina.

[38] Arun, K. S., Huang, T. S., and Blostein, S. D. (1987). Least-squares fitting of two 3-d point sets. *IEEE Trans. Pattern Anal. Mach. Intell.*, 9:698–700.

[39] Assuncao, G., Patrao, B., Castelo-Branco, M., and Menezes, P. (2022). An overview of emotion in artificial intelligence. *IEEE Transactions on Artificial Intelligence*, pages 1–1.

[40] Aubot (2022). Teleporter robot, aubot inc. https://aubot.com/.

[41] Avelino, J., Garcia-Marques, L., Ventura, R., and Bernardino, A. (2021). Break the ice: a survey on socially aware engagement for human–robot first encounters. *International Journal of Social Robotics*, 13(8):1851–1877.

[42] Avgousti, S., Christoforou, E., Panayides, A., Voskarides, S., Novales, C., Nouaille, L., Pattichis, C., and Vieyres, P. (2016). Medical telerobotic systems: Current status and future trends. *BioMedical Engineering OnLine*, 15.

[43] Aykut, T., Zou, C., Xu, J., Van Opdenbosch, D., and Steinbach, E. (2018). A delay compensation approach for pan-tilt-unit-based stereoscopic 360 degree telepresence systems using head motion prediction. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3323–3330.

[44] Aylett, R., Kappas, A., Castellano, G., Bull, S., Barendregt, W., Paiva, A., and Hall, L. (2015). I know how that feels—an empathic robot tutor. In *eChallenges e-2015 Conference*, pages 1–9. IEEE.

[45] Azevedo, T. C. S., Tavares, J. M. R. S., and Vaz, M. A. P. (2009). 3d object reconstruction from uncalibrated images using an off-the-shelf camera. *Advances in Computational Vision and Medical Image Processing; in series of Computational Methods in Applied Sciences, Springer Netherlands*, 13:117–136. Universidade do Porto.

[46] Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., and MacIntyre, B. (2001). Recent advances in augmented reality. *IEEE Comput. Graph. Appl.*, 21:34–47.

[47] B. Oving, A. and B.F. van Erp, J. (2001). Driving with a head-slaved camera system. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 45:1372–1376.

[48] Baddeley, A. (2010). Working memory. *Current Biology*, 20(4):R136 – R140.

[49] Bailenson, J., Patel, K., Nielsen, A., Bajscy, R., Jung, S.-H., and Kurillo, G. (2008). The Effect of Interactivity on Learning Physical Actions in Virtual Reality. *Media Psychology*, 11(3):354–376.

[50] Bailenson, J. N., Blascovich, J., Beall, A. C., and Loomis, J. M. (2001). Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence: Teleoperators & Virtual Environments*, 10(6):583–598.

[51] Bailenson, J. N., Swinth, K., Hoyt, C., Persky, S., Dimov, A., and Blascovich, J. (2005). The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. *Presence: Teleoperators & Virtual Environments*, 14(4):379–393.

[52] Baker, W., Kingston, Z., Moll, M., Badger, J., and Kavraki, L. E. (2017). Robonaut 2 and you: Specifying and executing complex operations. In *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, pages 1–8.

[53] Bansal, G., Rajgopal, K., Chamola, V., Xiong, Z., and Niyato, D. (2022). Healthcare in metaverse: A survey on current metaverse applications in healthcare. *IEEE Access*, 10:119914–119946.

[54] Baraka, K. and Veloso, M. (2015). Adaptive interaction of persistent robots to user temporal preferences. In *International Conference on Social Robotics*, pages 61–71. Springer.

[55] Barakova, E. I. and Lourens, T. (2010). Expressing and interpreting emotional movements in social games with robots. *Personal Ubiquitous Comput.*, 14:457–467.

[56] Baumgartner, T., Valko, L., Esslen, M., and Jäncke, L. (2006). Neural correlate of spatial presence in an arousing and noninteractive virtual reality: an eeg and psychophysiology study. *CyberPsychology & Behavior*, 9(1):30–45.

[57] Bay, H., Tuytelaars, T., and Gool, L. V. (2006). Surf: Speeded up robust features. In *In Computer Vision – ECCV 2006*, pages 404–417.

[58] Becker-Asano, C., Ogawa, K., Nishio, S., and Ishiguro, H. (2010). Exploring the uncanny valley with geminoid hi-1 in a real-world application. In *Proceedings of IADIS International conference interfaces and human computer interaction*, pages 121–128.

[59] Beddiar, D. R., Nini, B., Sabokrou, M., and Hadid, A. (2020). Vision-based human activity recognition: a survey. *Multimedia Tools and Applications*, 79(41):30509–30555.

[60] Bejczy, A. K. and Kim, W. S. (1990). Predictive displays and shared compliance control for time-delayed telemanipulation. In *EEE International Workshop on Intelligent Robots and Systems, Towards a New Frontier of Applications*, pages 407–412. IEEE.

[61] Bejczy, A. K., Kim, W. S., and Venema, S. C. (1990). The phantom robot: predictive displays for teleoperation with time delay. In *Proceedings., IEEE International Conference on Robotics and Automation*, pages 546–551. IEEE.

[62] Bente, G., Rüggenberg, S., Krämer, N. C., and Eschenburg, F. (2008). Avatar-mediated networking: Increasing social presence and interpersonal trust in net-based collaborations. *Human communication research*, 34(2):287–318.

[63] Bernardini, F. and Rushmeier, H. E. (2002). The 3d model acquisition pipeline. *Comput. Graph. Forum*, 21(2):149–172.

[64] Besl, P. J. and McKay, N. D. (1992). A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14:239–256.

[65] Bevan, C. and Fraser, D. S. (2015). Shaking hands and cooperation in tele-present human-robot negotiation. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 247–254. IEEE.

[66] Biocca, F. (1997a). The cyborg's dilemma: embodiment in virtual environments. In *Proc. Conf. Second Int Cognitive Technology 'Humanizing the Information Age'*, pages 12–26.

[67] Biocca, F. (1997b). The cyborg's dilemma: Embodiment in virtual environments. In *Proceedings of the 2nd International Conference on Cognitive Technology (CT '97)*, CT '97, page 12, USA. IEEE Computer Society.

[68] Biocca, F. and Harms, C. (2002). Defining and measuring social presence: Contribution to the networked minds theory and measure. *Proceedings of PRESENCE*, 2002:1–36.

[69] Biocca, F., Harms, C., and Burgoon, J. K. (2003). Toward a more robust theory and measure of social presence: Review and suggested criteria. *Presence: Teleoperators & virtual environments*, 12(5):456–480.

[70] Biocca, F., Harms, C., and Gregg, J. (2001). The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In *4th annual international workshop on presence, Philadelphia, PA*, pages 1–9.

[71] Biocca, F. A., Inoue, Y., Lee, A., Polinsky, H., and Tang, A. (2002). Visual cues and virtual touch: Role of visual stimuli and intersensory integration in cross-modal haptic illusions and the sense of presence. In *Proceedings of Presence 2002*.

[72] Blackler, A., Popovic, V., and Mahar, D. (2010). Investigating users' intuitive interaction with complex artefacts. *Applied ergonomics*, 41(1):72–92.

[73] Blanke, O. (2012). Multisensory brain mechanisms of bodily self-consciousness. *Nat Rev Neurosci*, 13(8):556–571.

[74] Blanke, O. and Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in cognitive sciences*, 13(1):7–13.

[75] Blanke, O., Slater, M., and Serino, A. (2015). Behavioral, neural, and computational principles of bodily self-consciousness. *Neuron*, 88(1):145–166.

[76] Blascovich, J. (2002). *Social influence within immersive virtual environments*. Springer.

[77] Blascovich, J. and Bailenson, J. (2011). *Infinite reality: Avatars, eternal life, new worlds, and the dawn of the virtual revolution*. William Morrow & Co, New York, USA.

[78] Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., and Bailenson, J. N. (2002a). Immersive virtual environment technology as a methodological tool for social psychology.

[79] Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., and gnson, J. N. (2002b). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological inquiry*, 13(2):103–124.

[80] Bogdanova, R., Boulanger, P., and Zheng, B. (2016). Depth perception of surgeons in minimally invasive surgery. *Surgical Innovation*, 23.

[81] Bohil, C., Owen, C., Jeong, E., Alicea, B., and Biocca, F. (2009). *Virtual Reality and presence, 21st Century Communication: A reference handbook*. SAGE Publications, Inc.

[82] Bohil, C. J., Alicea, B., and Biocca, F. A. (2011). Virtual reality in neuroscience research and therapy. *Nat Rev Neurosci*.

[83] Boian, R., Sharma, A., Han, C., Merians, A., Burdea, G., Adamovich, S., Recce, M., Tremaine, M., and Poizner, H. (2002). Virtual reality-based post-stroke hand rehabilitation. In *Medicine Meets Virtual Reality 02/10*, pages 64–70. IOS Press.

[84] Boitard, R., Mantiuk, R. K., and Pouli, T. (2015). Evaluation of color encodings for high dynamic range pixels. In *Human Vision and Electronic Imaging*, volume 9394 of *SPIE Proceedings*, page 93941K. SPIE.

[85] Bouguet, J.-Y. (2003). Camera calibration toolbox for matlab. In *www.vision.caltech.edu/bouguetj*.

[86] Braithwaite, J. J., Watson, D. G., Jones, R., and Rowe, M. (2013). A guide for analysing electrodermal activity (eda) & skin conductance responses (scrs) for psychological experiments. *Psychophysiology*, 49(1):1017–1034.

[87] Breedveld, P., Stassen, H., Meijer, D., and Stassen, L. (2009). Theoretical background and conceptual solution for depth perception and eye-hand coordination problems in laparoscopic surgery. *Minimally Invasive Therapy & Allied Technologies*, 8:227–234.

[88] Broz, F., Nourbakhsh, I., and Simmons, R. (2013). Planning for human–robot interaction in socially situated tasks. *International Journal of Social Robotics*, 5(2):193–214.

[89] Bulu, S. T. (2012). Place presence, social presence, co-presence, and satisfaction in virtual worlds. *Computers & Education*, 58(1):154–161.

[90] Burdea, G., Rabin, B., Chaperon, A., and Hundal, J. (2011). Emotive, cognitive and motor rehabilitation post severe traumatic brain injury a new convergent approach. In *International Conference on Virtual Rehabilitation*.

[91] Burdea, G. C. and Coiffet, P. (2003). *Virtual reality technology (2. ed.)*. Wiley.

[92] Calbi, A., Regazzoni, C. S., and Marcenaro, L. (2006). Dynamic scene reconstruction for efficient remote surveillance. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'06)*.

[93] Cao, Q., Gao, C., Wang, Y., and Li, T. (2016). The motion capture data glove device for virtual surgery. In *2016 IEEE Int. Nanoelectronics Conf. (INEC)*.

[94] Carney, T. and Klein, S. A. (1997). Resolution acuity is better than vernier acuity. *Vision Research*, 37(5):525 – 539.

[95] Carranza, J., Theobalt, C., Magnor, M. A., and Seidel, H.-P. (2003). Free-viewpoint video of human actors. *ACM Trans. Graph.*, 22(3):569–577.

[96] Carvalho, M., Avelino, J., Bernardino, A., Ventura, R. M. M., and Moreno, P. (2021). Human-robot greeting: tracking human greeting mental states and acting accordingly. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2021, Prague, Czech Republic, September 27 - Oct. 1, 2021*, pages 1935–1941. IEEE.

[97] Challis, J. (1995). A procedure for determining rigid body transformation parameters. *Journal of Biomechanics*, 28(6):733–737.

[98] Chen, D. J. Y. C., Durlach, P. J., Sloan, J. A., and Bowens, L. D. (2008). Human–robot interaction in the context of simulated route reconnaissance missions. *Military Psychology*, 20(3):135–149.

[99] Chen, J., Haas, E., and Barnes, M. (2007). Human performance issues and user interface design for teleoperated robots. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 37(6):1231–1245.

[100] Chen, J. Y. C. and Joyner, C. T. (2009). Concurrent performance of gunner's and robotics operator's tasks in a multitasking environment. *Military Psychology*, 21(1):98–113.

[101] Chen, J. Y. C. and Thropp, J. E. (2007). Review of low frame rate effects on human performance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 37(6):1063–1076.

[102] Chiang, H.-H., Chen, Y.-L., and Lin, C.-T. (2013). Human-robot interactive assistance of a robotic walking support system in a home environment. In *2013 IEEE International Symposium on Consumer Electronics (ISCE)*, pages 263–264. IEEE.

[103] Choi, J. J. and Kwak, S. S. (2017). Who is this?: Identity and presence in robot-mediated communication. *Cognitive Systems Research*, 43:174–189.

[104] Chow, J. and Lichti, D. (2013). Photogrammetric bundle adjustment with self-calibration of the primesense 3d camera technology: Microsoft kinect. *Access, IEEE*, 1:465–474.

[105] Clabaugh, C. and Matarić, M. (2019). Escaping oz: Autonomy in socially assistive robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:33–61.

[106] Claypool, K. T. and Claypool, M. (2007). The effects of resolution on users playing first person shooter games. In *Electronic Imaging*.

[107] Colledanchise, M. and Ögren, P. (2018). *Behavior trees in robotics and AI: An introduction*. CRC Press, Boca Raton, FL, USA.

[108] COMMITTEE, V. F. (1988). Visual acuity measurement standard. *Italian Journal of Ophthalmology*, II:1–15.

[109] Cornells, N. and Van Gool, L. (2005). Real-time connectivity constrained depth map computation using programmable graphics hardware. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition CVPR 2005*, volume 1, pages 1099–1104.

[110] Cortese, J. and Seo, M. (2012). The role of social presence in opinion expression during ftf and cmc discussions. *Communication Research Reports*, 29(1):44–53.

[111] Costantini, M. and Haggard, P. (2007). The rubber hand illusion: Sensitivity and reference frame for body ownership. *Consciousness and Cognition*, 16(2):229 – 240.

[112] Costanza, E., Kunz, A., and Fjeld, M. (2009). Mixed reality: A survey. In Lalanne, D. and Kohlas, J., editors, *Human Machine Interaction*, volume 5440 of *Lecture Notes in Computer Science*, pages 47–68. Springer.

[113] Crandall, J. W. and Goodrich, M. A. (2002). Characterizing efficiency of human robot interaction: A case study of shared-control teleoperation. In *IEEE/RSJ international conference on intelligent robots and systems*, volume 2, pages 1290–1295. IEEE.

[114] Cuervo, E., Chintalapudi, K., and Kotaru, M. (2018). Creating the perfect illusion: What will it take to create life-like virtual reality headsets? In *Proceedings of the 19th International Workshop on Mobile Computing Systems & Applications*, HotMobile '18, pages 7–12, New York, NY, USA. ACM.

[115] Cummings, J. J. and Bailenson, J. N. (2016). How immersive is enough? a meta-analysis of the effect of immersive technology on user presence. *Media Psychology*, 19(2):272–309.

[116] Cummings, J. J. and Wertz, B. (2018). Technological predictors of social presence: A foundation for a meta-analytic review and empirical concept explication. In *Proceedings of the 10th Annual International Workshop on Presence (Prague)*.

[117] Curless, B. and Levoy, M. (1996). A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, SIGGRAPH '96, pages 303–312, New York, NY, USA. ACM.

[118] CyberGlove Systems LLC (2019). Cyberglove III. http://www.cyberglovesystems.com/cyberglove-iii. [Online; accessed 21-July-2019].

[119] Daft, R. L. and Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management science*, 32(5):554–571.

[120] Damasio, A. (1999). *The Feeling of what Happens: Body and Emotion in the Making of Consciousness*. Harvest book. Harcourt Brace.

[121] Daniilidis, K., Mulligan, J., Mckendall, R., Majumder, A., Kamberova, G., Schmid, D., Bajcsy, R., and Fuchs, H. (1999). Towards the holodeck: An initial testbed for real-time 3d-teleimmersion. In *ACM SIGGRAPH*.

[122] Darken, R. P. and Cevik, H. (1999). Map usage in virtual environments: orientation issues. In *Proceedings IEEE Virtual Reality (Cat. No. 99CB36316)*, pages 133–140.

[123] Darken, R. P., Kempster, K., Kempster, M. K., and Peterson, B. (2001). Effects of streaming video quality of service on spatial comprehension in a reconnaisance task. In *Proceedings of the Meeting of I/ITSEC*.

[124] de Greef, H. (2014). Video communication best for female friends? In *ISPR 2014: 15th International Workshop on Presence (PRESENCE 2014), March 17-19, 2014, Vienna, Austria*, pages 187–193. ISPR.

[125] De Greef, P. and Ijsselsteijn, W. A. (2001). Social presence in a home tele-application. *CyberPsychology & Behavior*, 4(2):307–315.

[126] DeDonato, M., Dimitrov, V., Du, R., Giovacchini, R., Knoedler, K., Long, X., Polido, F., Gennert, M. A., Padır, T., Feng, S., Moriguchi, H., Whitman, E., Xinjilefu, X., and Atkeson, C. G. (2015). Human-in-the-loop control of a humanoid robot for disaster response: A report from the darpa robotics challenge trials. *Journal of Field Robotics*, 32(2):275–292.

[127] DeJong, B. P., Colgate, J. E., and Peshkin, M. A. (2011). *Mental Transformations in Human-Robot Interaction*, pages 35–51. Springer Netherlands, Dordrecht.

[128] DeLucia, P. R. and Griswold, J. A. (2011). Effects of camera arrangement on perceptual-motor performance in minimally invasive surgery. *Journal of Experimental Psychology: Applied*, 17(3):210–232.

[129] Devin, S. and Alami, R. (2016). An implemented theory of mind to improve human-robot shared plans execution. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 319–326. IEEE.

[130] Dicianno, B. E., Sibenaller, S., Kimmich, C., Cooper, R. A., and Pyo, J. (2009). Joystick use for virtual power wheelchair driving in individuals with tremor: Pilot study. *Journal of Rehabilitation Research & Development*, 46(2).

[131] Dicke, C., Aaltonen, V., Rämö, A., and Vilermo, M. (2010). Talk to me: The influence of audio quality on the perception of social presence. *Proceedings of HCI 2010 24*, pages 309–318.

[132] Dipietro, L., Sabatini, A. M., and Dario, P. (2008). A survey of glove-based systems and their applications. *IEEE Trans. on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(4).

[133] Double and Robotics (2022). Double robotics, inc. https://www.doublerobotics.com/.

[134] Duque, I., Dautenhahn, K., Koay, K. L., Christianson, B., et al. (2013). A different approach of using personas in human-robot interaction: Integrating personas as computational models to modify robot companions' behaviour. In *2013 IEEE RO-MAN*, pages 424–429. IEEE.

[135] Durand, F. and Dorsey, J. (2002). Fast bilateral filtering for the display of high-dynamic-range images. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '02, page 257–266, New York, NY, USA. Association for Computing Machinery.

[136] Dynamics, B. (2022a). Atlas - the most dynamic humanoid robot | boston dynamics. https://www.bostondynamics.com/atlas.

[137] Dynamics, B. (2022b). Spot - the agile mobile robot | boston dynamics. https://www.bostondynamics.com/products/spot.

[138] Eggert, D. W., Lorusso, A., and Fisher, R. B. (1997). Estimating 3D rigid body transformations: a comparison of four major algorithms. *MAchine Vision and Applications*, 9:272–290.

[139] Ellis, S., Adelstein, B., Baumeler, S., J. Jense, G., and H. Jacoby, R. (1999). Sensor spatial distortion, visual latency, and update rate effects on 3d tracking in virtual environments. pages 218–221.

[140] Ellis, S. R., Mania, K., Adelstein, B. D., and Hill, M. I. (2004). Generalizeability of latency detection in a variety of virtual environments. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 48(23):2632–2636.

[141] Engeser, S. and Rheinberg, F. (2008). Flow, performance and moderators of challenge-skill balance. *Motivation and Emotion*, 32:158–172.

[142] Ernst, M. O. and Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in cognitive sciences*, 8(4):162–169.

[143] Essmaeel, K., Gallo, L., Damiani, E., De Pietro, G., and Dipanda, A. (2014). Comparative evaluation of methods for filtering kinect depth data. *Multimedia Tools and Applications*, pages 1–24.

[144] European, C. (2010). Overview of the european strategy in ict for ageing well. Technical report.

[145] Farajiparvar, P., Ying, H., and Pandya, A. (2020). A brief survey of telerobotic time delay mitigation. *Frontiers in Robotics and AI*, 7.

[146] Feldmann, T., Mihailidis, I., Schulz, S., Paulus, D., and Worner, A. (2010). Online full body human motion tracking based on dense volumetric 3d reconstructions from multi camera setups. In Dillmann, R., Beyerer, J., Hanebeck, U., and Schultz, T., editors, *KI 2010, Advances in Artificial Intelligence*, volume 6359 of *Lecture Notes in Computer Science*, pages 74–81. Springer Berlin - Heidelberg.

[147] Feng, B., Li, S., and Li, N. (2016). Is a profile worth a thousand words? how online support-seeker's profile features may influence the quality of received support messages. *Communication Research*, 43(2):253–276.

[148] Fernando, C., Furukawa, M., Kurogi, T., Kamuro, S., Sato, K., Minamizawa, K., and Tachi, S. (2012). Design of telesar v for transferring bodily consciousness in telexistence. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 5112–5118.

[149] Ferreira, J., Lobo, J., Bessiere, P., Castelo-Branco, M., and Dias, J. (2013). A bayesian framework for active artificial perception. *IEEE Transactions on Cybernetics*, 43(2):699–711.

[150] Ferreira, J., Lobo, J., and Dias, J. (2011). Bayesian real-time perception algorithms on gpu. *Journal of Real-Time Image Processing*, 6:171–186. 10.1007/s11554-010-0156-7.

[151] Ferreira, J. F., Lobo, J., and Dias, J. (2010). Real-time perception algorithms on gpu: Real-time implementation of bayesian models for multimodal perception using cuda. In *Journal of Real-Time Image Processing, Springer Berlin/Heidelberg*.

[152] Ferreira, J. F., Pinho, C., and Dias, J. (2009). Implementation and calibration of a bayesian binaural system for 3d localisation. In *2008 IEEE International Conference on Robotics and Biomimetics*, pages 1722–1727.

[153] Ferrell, W. R. (1965). Remote manipulation with transmission delay. *IEEE Transactions on Human Factors in Electronics*, HFE-6(1):24–32.

[154] Ferrell, W. R. and Sheridan, T. B. (1967). Supervisory control of remote manipulation. *IEEE spectrum*, 4(10):81–88.

[155] Fiore, M., Khambhaita, H., Milliez, G., and Alami, R. (2015). An adaptive and proactive human-aware robot guide. In *International Conference on Social Robotics*, pages 194–203. Springer.

[156] Fisch, A., Mavroidis, C., Melli-Huber, J., and Bar-Cohen, Y. (2003). Haptic devices for virtual reality, telepresence, and human-assistive robotics. *Biologically inspired intelligent robots*, 73.

[157] Fischinger, D., Einramhof, P., Papoutsakis, K., Wohlkinger, W., Mayer, P., Panek, P., Hofmann, S., Koertner, T., Weiss, A., Argyros, A., et al. (2016). Hobbit,

a care robot supporting independent living at home: First prototype and lessons learned. *Robotics and Autonomous Systems*, 75:60–78.

[158] Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395.

[159] Folds, D. J. and Gerth, J. M. (1994). Auditory monitoring of up to eight simultaneous sources.

[160] Fong, T., Thorpe, C. E., and Baur, C. (2003). Robot, asker of questions. *Robotics Auton. Syst.*, 42:235–243.

[161] Fong, T. W., Thorpe, C., and Baur, C. (2001). A safeguarded teleoperation controller. In *Proceedings of IEEE International Conference on Advanced Robotics (ICAR '01)*, Budapest, Hungary.

[162] Francescato, D., Porcelli, R., Mebane, M., Cuddetta, M., Klobas, J., and Renzi, P. (2006). Evaluation of the efficacy of collaborative learning in face-to-face and computer-supported university contexts. *Computers in human behavior*, 22(2):163–176.

[163] Franco, J.-S. and Boyer, E. (2005). Fusion of multi-view silhouette cues using a space occupancy grid. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV05)*.

[164] Frank, L. H., Casali, J. G., and Wierwille, W. W. (1987). Effects of visual display and motion system delays on operator performance and uneasiness in a driving simulator. *Proceedings of the Human Factors Society Annual Meeting*, 31(5):492–496.

[165] Freedman, B., Shpunt, A., Machline, M., and Arieli, Y. (2010). Depth mapping using projected patterns.

[166] Freeman, J., Avons, S. E., Meddis, R., Pearson, D. E., and IJsselsteijn, W. A. (2000). Using behavioral realism to estimate presence: A study of the utility of postural responses to motion-stimuli. *Presence*, 9(2):149–164.

[167] Fu, F.-L., Su, R.-C., and Yu, S.-C. (2009). Egameflow: A scale to measure learners' enjoyment of e-learning games. *Comput. Educ.*, 52:101–112.

[168] Fuchs, H., Bishop, G., Arthur, K., McMillan, L., Bajcsy, R., Lee, S. W., Farid, H., and Kanade, T. (1994). Virtual space teleconferencing using a sea of cameras. In *Proc. First International Conference on Medical Robotics and Computer Assisted Surgery*, pages 161–167.

[169] FutureRobot (2022). Furo-i, future robot co., lda. http://www.futurerobot.com/default/.

[170] Gandy, M., Catrambone, R., MacIntyre, B., Alvarez, C., Eiriksdottir, E., Hilimire, M., Davidson, B., and McLaughlin, A. (2010). Experiences with an ar evaluation test bed: Presence, performance, and physiological measurement. In *Mixed and Augmented Reality (ISMAR), 2010 9th IEEE International Symposium on*, pages 127 –136.

[171] Gao, Y., Chang, H. J., and Demiris, Y. (2015). User modelling for personalised dressing assistance by humanoid robots. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1840–1845. IEEE.

[172] Garau, M., Slater, M., Vinayagamoorthy, V., Brogni, A., Steed, A., and Sasse, M. A. (2003). The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In Cockton, G. and Korhonen, P., editors, *Proceedings of the 2003 Conference on Human Factors in Computing Systems, CHI 2003, Ft. Lauderdale, Florida, USA, April 5-10, 2003*, pages 529–536. ACM.

[173] García, J., Fernández, J., Sanz, P., and Marín, R. (2010). Increasing autonomy within underwater intervention scenarios: The user interface approach. In *2010 IEEE International Systems Conference*, pages 71–75, San Diego, CA, USA.

[174] Garcia, J. C., Patrão, B., Almeida, L., Perez, J., Menezes, P., Dias, J., and Sanz, P. J. (2017). A natural interface for remote operation of underwater robots. *IEEE Computer Graphics and Applications*, 37(1):34–43.

[175] Garcia Bermudez, F. L., Julian, R. C., Haldane, D. W., Abbeel, P., and Fearing, R. S. (2012). Performance analysis and terrain classification for a legged robot over rough terrain. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 513–519.

[176] Garg, S., Sünderhauf, N., Dayoub, F., Morrison, D., Cosgun, A., Carneiro, G., Wu, Q., Chin, T.-J., Reid, I., Gould, S., Corke, P., and Milford, M. (2020). Semantics for robotic mapping, perception and interaction: A survey. *Foundations and Trends® in Robotics*, 8(1–2):1–224.

[177] Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Houghton Mifflin Company.

[178] Girão, P., Paulo, J., Garrote, L., and Peixoto, P. (2018). Real-time multi-view grid map-based spatial representation for mixed reality applications. In *Augmented Reality, Virtual Reality, and Computer Graphics*. Springer Int. Pub.

[179] Gnatzig, S., Chucholowski, F., Tang, T., and Lienkamp, M. (2013). A system design for teleoperated road vehicles. In Ferrier, J.-L., Gusikhin, O. Y., Madani, K., and Sasiadek, J. Z., editors, *ICINCO (2)*, pages 231–238. SciTePress.

[180] Goffman, E. (1963). *Behavior in Public Places*. A Free Press paperback. Free Press.

[181] Goldstein, E. and Brockmole, J. (2016). *Sensation and Perception*. Cengage Learning.

[182] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

[183] Goodrich, M. A., Crandall, J. W., and Barakova, E. (2013). Teleoperation and beyond for assistive humanoid robots. *Reviews of Human Factors and Ergonomics*, 9(1):175–226.

[184] Goodrich, M. A. and Schultz, A. C. (2008). Human–robot interaction: A survey. *Foundations and Trends® in Human–Computer Interaction*, 1(3):203–275.

[185] Gordon, G. and Breazeal, C. (2015). Bayesian active learning-based robot tutor for children's word-reading skills. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[186] Gorini, A., Gaggioli, A., and Riva, G. (2010). *Virtual Reality as an Experiential Tool The Role of Virtual Worlds in Psychological Interventions, Ubiquitous Health and Medical Informatics: The Ubiquity 2.0 Trend and Beyond*. IGI Global.

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

[188] Gray, S., Chitta, S., Kumar, V., and Likhachev, M. (2013). A single planner for a composite task of approaching, opening and navigating through non-spring and spring-loaded doors. In *2013 IEEE International Conference on Robotics and Automation*, pages 3839–3846. IEEE.

[189] Griesser, A., Roeck, S. D., Neubeck, A., and Gool, L. V. (2005). Gpu-based foreground-background segmentation using an extended colinearity criterion. In *In Proc. Vision, Modeling, and Visualization (VMV) 2005. Amsterdam, The Netherlands: IOS, Nov. 2005.*, pages 319–326.

[190] Grosinger, J., Pecora, F., and Saffiotti, A. (2016). Making robots proactive through equilibrium maintenance. In *IJCAI*, pages 3375–3381.

[191] Guerchouche, R., Bernier, O., and Zaharia, T. (2008). Multiresolution volumetric 3d object reconstruction for collaborative interactions. *Pattern Recognition and Image Analysis*, 18:621–637. 10.1134/S1054661808040147.

[192] Ha Park, S. and C. Woldstad, J. (2000). Multiple two-dimensional displays as an alternative to three-dimensional displays in telerobotic tasks. *Human factors*, 42:592–603.

[193] Han, J. and Conti, D. (2020). The use of utaut and post acceptance models to investigate the attitude towards a telepresence robot in an educational setting. *Robotics*, 9(2).

[194] Hanley, J. R. and Thomas, A. (1984). Maintenance rehearsal and the articulatory loop. *British Journal of Psychology*, 75(4):521–527.

[195] Harders, M. (2008). *Surgical Scene Generation for Virtual Reality-Based Training in Medicine*. Springer-Verlag TELOS, Santa Clara, CA, USA, 1 edition.

[196] Harms, C. and Biocca, F. (2004). Internal consistency and reliability of the networked minds social presence measure. In *Seventh Annual International Workshop: Presence 2004. Valencia: Universidad Politecnica de Valencia.*

[197] Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA.

[198] Hartley, R. and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. CAMBRIDGE UNIVERSITY PRESS.

[199] Hartmann, T., Wirth, W., Schramm, H., Klimmt, C., Vorderer, P., Gysbers, A., Böcking, S., Ravaja, N., Laarni, J., Saari, T., Gouveia, F., and Sacau, A. (2015). The spatial presence experience scale (spes). *Journal of Media Psychology: Theories, Methods, and Applications*, 1:1–15.

[200] Harutyunyan, V., Manohar, V., Gezehei, I., and Crandall, J. W. (2013). Cognitive telepresence in human-robot interactions. *Journal of Human-Robot Interaction*, 1(2):158–182.

[201] Hasenfratz, J. M., Lapierre, M., and Sillion, F. (2004). A real-time system for full body interaction with virtual worlds. In *Eurographics*.

[202] Hellou, M., Gasteiger, N., Lim, J. Y., Jang, M., and Ahn, H. S. (2021). Personalization and localization in human-robot interaction: A review of technical methods. *Robotics*, 10(4).

[203] Hemminahaus, J. and Kopp, S. (2017). Towards adaptive social behavior generation for assistive robots using reinforcement learning. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 332–340. IEEE.

[204] Henry, P., Krainin, M., Herbst, E., Ren, X., and Fox, D. (2010). RGB-D Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments. In *RSS Workshop on Advanced Reasoning with Depth Cameras*.

[205] Henry, P., Krainin, M., Herbst, E., Ren, X., and Fox, D. (2012). Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. *I. J. Robotic Res.*, 31(5):647–663.

[206] Herath, D. C., Jochum, E., and Vlachos, E. (2017). An experimental study of embodied interaction and human perception of social presence for interactive robots in public settings. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):1096–1105.

[207] Herrera C, D., Kannala, J., and Heikkila, J. (2012). Joint depth and color camera calibration with distortion correction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(10):2058–2064.

[208] Heutte, J., Fenouillet, F., Boniwell, I., Martin-Krumm, C., and Csikszentmihalyi, M. (2014). Eduflow: Proposal for a new measure of flow in education. *Previous paper*, (Previous paper).

[209] Hilton, A. and Illingworth, J. (2000). Geometric fusion for a hand-held 3d sensor. *Mach. Vision Appl.*, 12(1):44–51.

[210] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[211] Hoda, M., Hafidh, B., and El Saddik, A. (2015). Haptic glove for finger rehabilitation. In *2015 IEEE Int. Conf. on Multimedia Expo Workshops (ICMEW)*.

[212] Hodges, L. F., Kooper, R., Meyer, T. C., Rothbaum, B. O., Opdyke, D., de Graaff, J. J., Williford, J. S., and North, M. M. (1995). Virtual environments for treating the fear of heights. *Computer*, 28:27–34.

[213] Hoffman, H. G., Richards, T. L., Bills, A. R., Oostrom, T. V., Magula, J., Seibel, E. J., and Sharar, S. R. (2006). Using fmri to study the neural correlates of virtual reality analgesia. *CNS Spectr*, 11(1):45–51.

[214] Honda (2022). Asimo robot | honda. https://asimo.honda.com/default.aspx.

[215] Hornstein, J., Lopes, M., Santos-Victor, J., and Lacerda, F. (2006). Sound localization for humanoid robots - building audio-motor maps based on the hrtf. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1170–1176.

[216] 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.

[217] Hosseini, M., Sengul, A., Pane, Y., De Schutter, J., and Bruyninckx, H. (2018). Haptic perception of virtual spring stiffness using exoten-glove. In *2018 11th Int. Conf. on Human System Interaction (HSI)*.

[218] HTC, V. (2022). Htc vive. https://www.vive.com/eu/. (Last Accessed: november 2022).

[219] Huang, C.-M. and Mutlu, B. (2016). Anticipatory robot control for efficient human-robot collaboration. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, pages 83–90. IEEE.

[220] Hubicki, C., Abate, A., Clary, P., Rezazadeh, S., Jones, M., Peekema, A., Van Why, J., Domres, R., Wu, A., Martin, W., et al. (2018). Walking and running with passive compliance: Lessons from engineering: A live demonstration of the atrias biped. *IEEE Robotics & Automation Magazine*, 25(3):23–39.

[221] IEEE (2022). Robots: Your guide to the world of robotics | ieee 2022. `https://robots.ieee.org/`. (Last Accessed: 31 Oct 2022).

[222] IJsselsteijn, W. (2003). *Presence in the past: what can we learn from media history?*, pages 17 – 40. Emerging communication: studies in new technologies and practices in communication; 5. IOS Press, Amsterdam.

[223] InbotTechnology (2022). Padbot, inbot technology, ltd. https://www.padbot.com/.

[224] Inc., X. (2022). kubi telepresence robots, xandex inc. https://www.kubiconnect.com/.

[225] Institute, S. K. R. (2011). Advanced rehabilitative technologies (art). Web. http://www.allina.com/ahs/ski.nsf/page/art.

[226] InTouchHealth and iRobot (2022). Rp-vita telepresence robot, intouch health and irobot, inc. https://intouchhealth.com/.

[227] Irfan, Q., Jensen, C., Ni, Z., and Hietpas, S. (2018). Building an exoskeleton glove on virtual reality platform. In *2018 IEEE Int. Conf. on Electro/Information Technology (EIT)*.

[228] Isabet, B., Pino, M., Lewis, M., Benveniste, S., and Rigaud, A.-S. (2021). Social telepresence robots: A narrative review of experiments involving older adults before and during the covid-19 pandemic. *International Journal of Environmental Research and Public Health*, 18(7).

[229] Isgro, F., Trucco, E., Kauff, P., and Schreer, O. (2004). Three-dimensional image processing in the future of immersive media. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(3):288 – 303.

[230] Ishiguro, H. (2006). Android science: conscious and subconscious recognition. *Connection Science*, 18(4):319–332.

[231] Ishiyama, H. and Kurabayashi, S. (2016). Monochrome glove: A robust real-time hand gesture recognition method by using a fabric glove with design of structured markers. In *2016 IEEE Virtual Reality (VR)*.

[232] Islam, S., Liu, X., Saddik, A., Seneviratne, L., and Dias, J. (2014). Control schemes for passive teleoperation systems over wide area communication networks with time varying delay. *International Journal of Automation and Computing*, 11(1):100–108.

[233] J Prendergast, C., Ryder, B., Abodeely, A., Muratore, C., P Crawford, G., and Luks, F. (2008). Surgical performance with head-mounted displays in laparoscopic surgery. *Journal of laparoendoscopic & advanced surgical techniques. Part A*, 19 Suppl 1:S237–40.

[234] Janvier, M., Durand, L., Cardinal, M. R., Renaud, I., Chayer, B., Bigras, P., de Guise, J. A., Soulez, G., and Cloutier, G. (2008). Performance evaluation of a medical robotic 3d-ultrasound imaging system. *Medical Image Analysis*, 12(3):275–290.

[235] Ji, Y., Yang, Y., Shen, F., Shen, H. T., and Li, X. (2020). A survey of human action analysis in hri applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7):2114–2128.

[236] Jin, S.-A. A. (2010). Parasocial interaction with an avatar in second life: A typology of the self and an empirical test of the mediating role of social presence. *PRESENCE: Teleoperators and Virtual Environments*, 19:331–340.

[237] Jo, D., Kim, K.-H., and Kim, G. J. (2017). Effects of avatar and background types on users' co-presence and trust for mixed reality-based teleconference systems. In *Proceedings the 30th Conference on Computer Animation and Social Agents, Seoul, South Korea, May 22-24*, pages 27–36.

[238] Johnson, R. D. (2011). Gender differences in e-learning: Communication, social presence, and learning outcomes. *J. Organ. End User Comput.*, 23(1):79–94.

[239] Judd, D. B. and Wyszecki, G. (1975). *Color in business, science, and industry*. Wiley New York, 3d ed. edition.

[240] Jung, S., Roh, S., Yang, H., and Biocca, F. (2017). Location and modality effects in online dating: rich modality profile and location-based information cues increase social presence, while moderating the impact of uncertainty reduction strategy. *Cyberpsychology, Behavior, and Social Networking*, 20(9):553–560.

[241] Kalantari, M., Hashemi, A., Jung, F., and Guedon, J.-P. (2011). A new solution to the relative orientation problem using only 3 points and the vertical direction. *Journal of Mathematical Imaging and Vision*, 39:259–268.

[242] Kanade, T., Rander, P., and Narayanan, P. J. (1997). Virtualized reality: constructing virtual worlds from real scenes. *IEEE_M_MM*, 4(1):34–47.

[243] Kaneko, K., Kaminaga, H., Sakaguchi, T., Kajita, S., Morisawa, M., Kumagai, I., and Kanehiro, F. (2019). Humanoid robot hrp-5p: An electrically actuated humanoid robot with high-power and wide-range joints. *IEEE Robotics and Automation Letters*, 4(2):1431–1438.

[244] Kang, S.-H. and Watt, J. H. (2013). The impact of avatar realism and anonymity on effective communication via mobile devices. *Computers in Human Behavior*, 29(3):1169–1181.

[245] Karami, A. B., Sehaba, K., and Encelle, B. (2016). Adaptive artificial companions learning from users' feedback. *Adaptive Behavior*, 24(2):69–86.

[246] Kelley, R., Nicolescu, M., Tavakkoli, A., King, C., and Bebis, G. (2008). Understanding human intentions via hidden markov models in autonomous mobile robots. In *Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on*, pages 367–374.

[247] Kendon, A. (1990). *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press, Cambridge, U.K.

[248] Kessels, R. P. C., van Zandvoort, M. J. E., Postma, A., Kappelle, L. J., and de Haan, E. H. F. (2000). The corsi block-tapping task: Standardization and normative data. *Applied Neuropsychology*, 7(4):252–258.

[249] Khan, S. M., Yan, P., and Shah, M. (2007). A homographic framework for the fusion of multi-view silhouettes. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*.

[250] Khoshelham, K. and Elberink, E. O. (2012). Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, 12:1437–1454.

[251] Kim, D. and Jo, D. (2022). Effects on co-presence of a virtual human: A comparison of display and interaction types. *Electronics*, 11(3).

[252] Kim, H., Suh, K.-S., and Lee, U.-K. (2013a). Effects of collaborative online shopping on shopping experience through social and relational perspectives. *Information & Management*, 50(4):169–180.

[253] Kim, H.-G., Yang, J.-Y., and Kwon, D.-S. (2014). Experience based domestic environment and user adaptive cleaning algorithm of a robot cleaner. In *2014 11th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pages 176–178. IEEE.

[254] Kim, H.-S. and Shyam Sundar, S. (2014). Can online buddies and bandwagon cues enhance user participation in online health communities? *Computers in Human Behavior*, 37:319–333.

[255] Kim, K., Schubert, R., Hochreiter, J., Bruder, G., and Welch, G. (2019). Blowing in the wind: Increasing social presence with a virtual human via environmental airflow interaction in mixed reality. *Computers & Graphics*, 83:23–32.

[256] Kim, K. J., Park, E., and Sundar, S. S. (2013b). Caregiving role in human–robot interaction: A study of the mediating effects of perceived benefit and social presence. *Computers in Human Behavior*, 29(4):1799–1806.

[257] Kim, T. and Biocca, F. (1997). Telepresence via television: Two dimensions of telepresence may have different connections to memory and persuasion. *Journal of Computer-mediated Communication*, 3.

[258] Kim, Y. M., Theobalt, C., Diebel, J., Kosecka, J., Miscusik, B., and Thrun, S. (2009). Multi-view image and tof sensor fusion for dense 3d reconstruction. In *Proc. IEEE 12th Int Computer Vision Workshops (ICCV Workshops) Conf*, pages 1542–1549.

[259] Klee, S. D., Ferreira, B. Q., Silva, R., Costeira, J. P., Melo, F. S., and Veloso, M. (2015). Personalized assistance for dressing users. In *International Conference on Social Robotics*, pages 359–369. Springer.

[260] Klein, S. A. and Levi, D. M. (1985). Hyperacuity thresholds of 1 sec: theoretical predictions and empirical validation. *J. Opt. Soc. Am. A*, 2(7):1170–1190.

[261] 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: Teleoper. Virtual Environ.*, 13(5):572–577.

[262] Knoblauch, D. and Kuester, F. (2009). Focused volumetric visual hull with color extraction. In Bebis, G., Boyle, R., Parvin, B., Koracin, D., Kuno, Y., Wang, J., Pajarola, R., Lindstrom, P., Hinkenjann, A., Encarnacao, M., Silva, C., and Coming, D., editors, *Advances in Visual Computing*, volume 5876 of *Lecture Notes in Computer Science*, pages 208–217. Springer Berlin-Heidelberg.

[263] Kristoffersson, A., Severinson Eklundh, K., and Loutfi, A. (2013). Measuring the quality of interaction in mobile robotic telepresence: A pilot perspective. *International Journal of Social Robotics*, 5(1):89–101.

[264] Kumcu, A., Vermeulen, L., Elprama, S., Duysburgh, P., Platisa, L., Nieuwenhove, Y., Van De Winkel, N., Jacobs, A., Looy, J., and Philips, W. (2016). Effect of video lag on laparoscopic surgery: correlation between performance and usability at low latencies. *International Journal of Medical Robotics and Computer Assisted Surgery*, 13.

[265] Kurillo, G., Koritnik, T., Bajd, T., and Bajcsy, R. (2011). Real-time 3d avatars for tele-rehabilitation in virtual reality. *Stud Health Technol Inform*, 163:290–6.

[266] Kurillo, G., Vasudevan, R., Lobaton, E., and Bajcsy, R. (2008). A framework for collaborative real-time 3d teleimmersion in a geographically distributed environment. In *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*, pages 111 –118.

[267] Kurosu, M. and Hashizume, A. (2021). Erm-at applied to social aspects of everyday life. In Kurosu, M., editor, *Human-Computer Interaction. Theory, Methods and Tools, 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29*, pages 280–290, Cham, Switzerland. Springer International Publishing.

[268] Ladikos, A., Benhimane, S., and Navab, N. (2008). Efficient visual hull computation for real-time 3d reconstruction using cuda. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, pages 1 –8.

[269] Lai, P.-L. and Yilmaz, A. (2008a). Efficient object shape recovery via slicing planes. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –6.

[270] Lai, P.-L. and Yilmaz, A. (2008b). Projective reconstruction of building shape from silhouette images acquired from uncalibrated cameras. In *ISPRS Congress Beijing 2008, Proceedings of Commission III*.

[271] Lam, C., Yang, A. Y., Driggs-Campbell, K., Bajcsy, R., and Sastry, S. S. (2015). Improving human-in-the-loop decision making in multi-mode driver assistance systems using hidden mode stochastic hybrid systems. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5776–5783.

[272] Lane, J. C., Carignan, C. R., Sullivan, B. R., Akin, D. L., Hunt, T., and Cohen, R. (2002). Effects of time delay on telerobotic control of neutral buoyancy vehicles. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, volume 3, pages 2874–2879.

[273] Lange, B., Requejo, P., Flynn, S., Rizzo, A., Valero-Cuevas, F., Baker, L., and Winstein, C. (2010). The potential of virtual reality and gaming to assist successful aging with disability. *Physical Medicine and Rehabilitation Clinics of North America*, 21(2):339 – 356.

[274] Lanier, J. (2001). Virtually there. *j-SCI-AMER*, 284(4):66–75.

[275] Lanillos, P., Ferreira, J. F., and Dias, J. (2015). Designing an artificial attention system for social robots. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4171–4178.

[276] LaValle, S. (2019). *Virtual Reality.* Cambridge University Press, Avaiable at http://vr.cs.uiuc.edu/.

[277] Lee, H. and Yilmaz, A. (2010). 3d reconstruction using photo consistency from uncalibrated multiple views. In *VISAPP 2010 - The International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*.

[278] Lee, K.-M. and Nass, C. (2005). Social-psychological origins of feelings of presence: Creating social presence with machine-generated voices. *Media Psychology*, 7(1):31–45.

[279] Lee, S. and Kim, G. J. (2008). Effects of haptic feedback, stereoscopy, and image resolution on performance and presence in remote navigation. *International Journal of Human-Computer Studies*, 66(10):701 – 717.

[280] Lessiter, J., Freeman, J., Keogh, E., and Davidoff, J. (2001). A cross-media presence questionnaire: The itc-sense of presence inventory. *Presence: Teleoper. Virtual Environ.*, 10:282–297.

[281] Lester, D. and Thronson, H. (2011). Human space exploration and human spaceflight: Latency and the cognitive scale of the universe. *Space Policy*, 27(2):89 – 93.

[282] Lewis, J. R. (1993). Ibm computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. Technical report, Boca Raton, FL.

[283] Lewis, J. R. (1995). Ibm computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *Int. J. Hum.-Comput. Interact.*, 7(1):57–78.

[284] Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77:23–37.

[285] Li, L.-J., Socher, R., and Fei-Fei, L. (2009). Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2036–2043.

[286] Li, S., Feng, B., Li, N., and Tan, X. (2015). How social context cues in online support-seeking influence self-disclosure in support provision. *Communication Quarterly*, 63(5):586–602.

[287] Lichiardopol, S. (2007). A survey on teleoperation. *Technische Universitat Eindhoven, DCT report*, 20:40–60.

[288] Lim, G. H., Hong, S. W., Lee, I., Suh, I. H., and Beetz, M. (2013). Robot recommender system using affection-based episode ontology for personalization. In *2013 IEEE RO-MAN*, pages 155–160. IEEE.

[289] Lim, J. and Richardson, J. C. (2016). Exploring the effects of students' social networking experience on social presence and perceptions of using snss for educational purposes. *The Internet and Higher Education*, 29:31–39.

[290] Lin, H.-Y. and Wu, J.-R. (2008). 3d reconstruction by combining shape from silhouette with stereo. In *IEEE*.

[291] Lion, D. M. (1993). *Three dimensional manual tracking using a head-tracked stereoscopic display (Technical Report).* Human Interface Technology Lab, WA.

[292] Lipton, Z. C., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*.

[293] Liu, H., Fang, S., Zhang, Z., Li, D., Lin, K., and Wang, J. (2021). Mfdnet: Collaborative poses perception and matrix fisher distribution for head pose estimation. *IEEE Transactions on Multimedia*, pages 1–1.

[294] Livingstone, M. and Hubel, D. (1988). Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, 240 4853:740–9.

[295] Lobo, J., Almeida, L., Alves, J., and Dias, J. (2003). Registration and segmentation for 3d map building: A solution based on stereo vision and inertial sensors. In *ICRA*, pages 139–144. IEEE.

[296] Lobo, J. and Dias, J. (2007). Relative pose calibration between visual and inertial sensors. *International Journal of Robotics Research, Special Issue 2nd Workshop on Integration of Vision and Inertial Sensors*, 26:561–575.

[297] Lombard, M., Biocca, F., Freeman, J., IJsselsteijn, W., and Schaevitz, R. (2015). *Immersed in Media: Telepresence Theory, Measurement & Technology*. Springer International Publishing, Cham.

[298] Lombard, M. and Ditton, T. (1997). At the heart of it all: The concept of presence. *Journal of Computer-Mediated Communication*, 3.

[299] Lombard, M., Ditton, T. B., Crane, D., Davis, B., Gil-Egui, G., Horvath, K., Rossman, J., and Park, S. (2000). Measuring presence: A literature-based approach to the development of a standardized paper-and-pencil instrument. In *Third international workshop on presence, delft, the netherlands, March 27-28*, volume 240, pages 2–4.

[300] Lopes, A. C., Pires, G., and Nunes, U. (2013). Assisted navigation for a brain-actuated intelligent wheelchair. *Robotics and Autonomous Systems*, 61(3):245–258.

[301] Lorek, K. S. and Willinger, G. L. (2016). A multivariate time-series prediction model for cash-flow data. *The Accounting Review*, 71:81–102.

[302] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110.

[303] Lu, Y., Huang, Q., Li, M., Jiang, X., and Keerio, M. (2008). A friendly and human-based teleoperation system for humanoid robot using joystick. In *2008 7th World Congress on Intelligent Control and Automation*, pages 2283–2288. IEEE.

[304] Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2*, pages 674–679. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[305] Lum, M. J. H., Rosen, J., Lendvay, T. S., Sinanan, M. N., and Hannaford, B. (2009). Effect of time delay on telesurgical performance. In *2009 IEEE International Conference on Robotics and Automation*, pages 4246–4252.

[306] Ma, R. and Kaber, D. B. (2006). Presence, workload and performance effects of synthetic environment design factors. *International Journal of Human-Computer Studies*, 64(6):541 – 552.

[307] Macedo, J. A., Kaber, D. B., Endsley, M. R., Powanusorn, P., and Myung, S. (1998). The effect of automated compensation for incongruent axes on teleoperator performance. *Human Factors*, 40(4):541–553. PMID: 9974228.

[308] Madureira, A., Cunha, B., Pereira, J. P., Gomes, S., Pereira, I., Santos, J. M., and Abraham, A. (2014). Using personas for supporting user modeling on scheduling systems. In *2014 14th International Conference on Hybrid Intelligent Systems*, pages 279–284. IEEE.

[309] Maimone, A., Bidwell, J., Peng, K., and Fuchs, H. (2012). Enhanced personal autostereoscopic telepresence system using commodity depth cameras. *Computers & Graphics*, 36(7):791–807.

[310] Maimone, A. and Fuchs, H. (2011). Encumbrance-free telepresence system with real-time 3d capture and display using commodity depth cameras. In *10th IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2011, Basel, Switzerland, October 26-29, 2011*, pages 137–146.

[311] Maimone, M., Cheng, Y., and Matthies, L. (2007). Two years of visual odometry on the mars exploration rovers. *Journal of Field Robotics*, 24(3):169–186.

[312] Makransky, G., Lilleholt, L., and Aaby, A. (2017). Development and validation of the multimodal presence scale for virtual reality environments: A confirmatory factor analysis and item response theory approach. *Computers in Human Behavior*, 72.

[313] Mania, K., Adelstein, B. D., Ellis, S. R., and Hill, M. I. (2004). Perceptual sensitivity to head tracking latency in virtual environments with varying degrees of scene complexity. In *Proceedings of the 1st Symposium on Applied Perception in Graphics and Visualization*, APGV '04, pages 39–47, New York, NY, USA. ACM.

[314] Mania, K. and Chalmers, A. (2000). A user-centered methodology for investigating presence and task performance. Technical report, Bristol, UK, UK.

[315] Mania, K., Wooldridge, D., Coxon, M., and Robinson, A. (2006). The effect of visual and interaction fidelity on spatial cognition in immersive virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, 12(3):396–404.

[316] MantaroBot1 (2022). Teleme - telepresence robot, mantaro inc. http://www.mantarobot.com/.

[317] MantaroBot2 (2022). Tabletop teleme - telepresence robot, mantaro inc. http://www.mantarobot.com/.

[318] Mariet, Z. and Kuznetsov, V. (2019). Foundations of sequence-to-sequence modeling for time series. In *The 22nd international conference on artificial intelligence and statistics*, pages 408–417. PMLR.

[319] Martins, G. S., Al Tair, H., Santos, L., and Dias, J. (2019a). αpomdp: Pomdp-based user-adaptive decision-making for social robots. *Pattern Recognition Letters*, 118:94–103. Cooperative and Social Robots: Understanding Human Activities and Intentions.

[320] Martins, G. S., Santos, L., and Dias, J. (2015). The growmeup project and the applicability of action recognition techniques. In *Third workshop on recognition and action for scene understanding (REACTS). Ruiz de Aloza*.

[321] Martins, G. S., Santos, L., and Dias, J. (2019b). User-adaptive interaction in social robots: A survey focusing on non-physical interaction. *International Journal of Social Robotics*, 11(1):185–205.

[322] Martins, H. and Ventura, R. (2009). Immersive 3-d teleoperation of a search and rescue robot using a head-mounted display. In *2009 IEEE conference on emerging technologies & factory automation*, pages 1–8. IEEE.

[323] Masia, L., Casadio, M., Sandini, G., and Morasso, P. (2009). Eye-hand coordination during dynamic visuomotor rotations. *PLOS ONE*, 4(9):1–11.

[324] Massimino, M. J. and Sheridan, T. B. (1994). Teleoperator performance with varying force and visual feedback. *Human factors*, 36(1):145–157.

[325] Matsubara, T., Miro, J. V., Tanaka, D., Poon, J., and Sugimoto, K. (2015). Sequential intention estimation of a mobility aid user for intelligent navigational assistance. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 444–449. IEEE.

[326] Matsumura, R., Shiomi, M., Nakagawa, K., Shinozawa, K., and Miyashita, T. (2016). A desktop-sized communication robot:“robovie-mr2”. *Journal of Robotics and Mechatronics*, 28(1):107–108.

[327] Matthies, L. and Shafer, S. A. (1986). Error modelling in stereo navigation. In *FJCC*, pages 114–122. IEEE Computer Society.

[328] May, S., Droeschel, D., Holz, D., Fuchs, S., Malis, E., Nüchter, A., and Hertzberg, J. (2009). Three-dimensional mapping with time-of-flight cameras. *J. Field Robot.*, 26:934–965.

[329] MBARI (2015). MBARI Ridges 2005 Expedition. `http://www.mbari.org/expeditions/ridges2005/august_15.htm`. (Last Accessed: Oct 2022).

[330] McTear, M. F. (1993). User modelling for adaptive computer systems: a survey of recent developments. *Artificial intelligence review*, 7(3-4):157–184.

[331] Meehan, M., Insko, B., Whitton, M., and Brooks, Jr, F. (2002). Physiological measures of presence in stressful virtual environments. *ACM Transactions on Graphics*, 21:645–652.

[332] Meehan, M., Razzaque, S., Whitton, M. C., and Brooks, F. P. (2003). Effect of latency on presence in stressful virtual environments. In *IEEE Virtual Reality, 2003. Proceedings.*, pages 141–148.

[333] Menezes, P., Gouveia, N., and Patrão, B. (2018). *Touching Is Believing - Adding Real Objects to Virtual Reality*.

[334] Menezes, P., Lerasle, F., and Dias, J. (2011). Towards human motion capture from a camera mounted on a mobile robot. *IVC*, 29(6):382–393.

[335] Menezes, P. and Rocha, R. P. (2021). Promotion of active ageing through interactive artificial agents in a smart environment. *SN Applied Sciences*, 3(5):1–15.

[336] Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. MIT Press.

[337] Mi, J., Tang, S., Deng, Z., Goerner, M., and Zhang, J. (2019). Object affordance based multimodal fusion for natural human-robot interaction. *Cognitive Systems Research*, 54:128–137.

[338] Michoud, B., Guillou, E., and Bouakaz, S. (2007). Real-time and markerless 3d human motion capture using multiple views. *Human Motion-Understanding, Modeling, Capture and Animation, Springer Berlin/Heidelberg.*, 4814/2007:88–103.

[339] Michoud, B., Saida, B., Erwan, G., and Hector, B. (2008). Largest silhouette-equivalent volume for 3d shapes modeling without ghost object. In *M2SFA2 2008: Workshop on Multi-camera and Multi-modal Sensor Fusion, Marseille, France.*

[340] Microsoft (2014a). Kinect for windows sdk. http://msdn.microsoft.com/en-us/library/dn772637.aspx. (Last Accessed: 2014).

[341] Microsoft (2014b). Kinect for windows sensor components and specifications. http://msdn.microsoft.com/en-us/library/jj131033.aspx. (Last Accessed: 2014).

[342] Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In Kobayashi, T., Hirose, K., and Nakamura, S., editors, *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048. ISCA.

[343] Milgram, P., Takemura, H., Utsumi, A., and Kishino, F. (1994). Augmented reality: A class of displays on the reality-virtuality continuum. pages 282–292.

[344] Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97.

[345] Minsky, M. (1980). Telepresence. *Omni*, pages 45–51.

[346] Mirisola, L. G. B., Dias, J., and de Almeida, A. T. (2007a). Trajectory recovery and 3d mapping from rotation-compensated imagery for an airship. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems San Diego, CA, USA, Oct 29 - Nov 2, 2007*.

[347] Mirisola, L. G. B., Lobo, J., and Dias, J. (2007b). 3d map registration using vision/laser and inertial sensing. In *EMCR*.

[348] Monteiro, F., Rocha, P., Menezes, P., Silva, A., and Dias, J. (1997). Teleoperating a mobile robot. a solution based on java language. In *Industrial Electronics, 1997. ISIE '97., Proceedings of the IEEE International Symposium on*, volume 1, pages SS263–SS267 vol.1.

[349] Montemerlo, M. and Thrun, S. (2007). *FastSLAM: A scalable method for the simultaneous localization and mapping problem in robotics*, volume 27. Springer.

[350] Moore, J. W. and Fletcher, P. C. (2012). Sense of agency in health and disease: a review of cue integration approaches. *Consciousness and cognition*, 21(1):59–68.

[351] Mori, M., MacDorman, K. F., and Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100.

[352] Mostofa, N., Avendano, I., McMahan, R. P., Conner, N. E., Anderson, M., and Welch, G. F. (2021). Tactile telepresence for isolated patients. In *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), Bari, Italy, Oct. 4 - Oct. 8*, pages 346–351, Los Alamitos, CA, USA. IEEE Computer Society.

[353] Moustris, G. P., Geravand, M., Tzafestas, C., and Peer, A. (2016). User-adaptive shared control in a mobility assistance robot based on human-centered intention reading and decision making scheme. In *IEEE International Conference on Robotics and Automation Workshop: Human-Robot Interfaces for Enhanced Physical Interactions*.

[354] MTR Corporation, Hong Kong (2020). MTR deploys new “vapourised hydrogen peroxide robot to further enhance disinfection of stations and trains. https://www.mtr.com.hk/archive/corporate/en/press_release/PR-20-020-E.pdf.

[355] Müller, S., Sprenger, S., and Gross, H.-M. (2014). Online adaptation of dialog strategies based on probabilistic planning. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 692–697. IEEE.

[356] Murphy, R. (2014). Real robots to help fight ebola.

[357] Nahrstedt, K., Yang, Z., Wu, W., Arefin, M. A., and Rivas, R. (2011). Next generation session management for 3d teleimmersive interactive environments. *Multimedia Tools Appl.*, 51(2):593–623.

[358] Nakanishi, H., Tanaka, K., and Wada, Y. (2014). Remote handshaking: touch enhances video-mediated social telepresence. In Jones, M., Palanque, P. A., Schmidt, A., and Grossman, T., editors, *CHI Conference on Human Factors in Computing Systems, CHI'14, Toronto, ON, Canada - April 26 - May 01, 2014*, pages 2143–2152. ACM.

[359] Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., and Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE Access*, 7:19143–19165.

[360] Nations, U. (2022). World population prospects 2022: Summary of results. un desa/pop/2022/tr/no. 3. Technical report.

[361] Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohli, P., Shotton, J., Hodges, S., and Fitzgibbon, A. (2011). Kinectfusion: Real-time dense surface mapping and tracking. In *Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality*, ISMAR '11, pages 127–136, Washington, DC, USA. IEEE Computer Society.

[362] Nguyen, C., Izadi, S., and Lovell, D. (2012). Modeling kinect sensor noise for improved 3d reconstruction and tracking. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*, pages 524–530.

[363] Nielsen, J. (1996). Usability metrics: tracking interface improvements. *IEEE Software*, 13(6):1–2.

[364] Nielsen, J. (2012). Usability 101: Introduction to usability. nielsen norman group. *Retrived from http://www. nngroup. com/articles/usability-101-introduction-to-usability/[Octubre 27, 2014]*.

[365] Nielsen, J. and Budiu, R. (2012). *Mobile Usability*. Pearson Education.

[366] Nikolaidis, S., Kuznetsov, A., Hsu, D., and Srinivasa, S. (2016). Formalizing human-robot mutual adaptation: A bounded memory model. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 75–82. IEEE.

[367] Nitschke, C., Nakazawa, A., and Takemura, H. (2007). Real-time space carving using graphics hardware. *IEICE - Trans. Inf. Syst.*, E90-D:1175–1184.

[368] Nocentini, O., Fiorini, L., Acerbi, G., Sorrentino, A., Mancioppi, G., and Cavallo, F. (2019). A survey of behavioral models for social robots. *Robotics*, 8(3).

[369] Norcio, A. F. and Stanley, J. (1989). Adaptive human-computer interfaces: A literature survey and perspective. *IEEE Transactions on Systems, Man, and cybernetics*, 19(2):399–408.

[370] North, M. M., North, S. M., D, E., Coble, J. R., D, P., and D, P. (1997). Virtual reality therapy - an effective treatment for psychological disorders. In *International Journal of Virtual Reality*, pages 2–6.

[371] Nowak, K. L. and Biocca, F. (2003). The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, 12(5):481–494.

[372] Nowak, K. L. and Fox, J. (2018). Avatars and computer-mediated communication: a review of the definitions, uses, and effects of digital representations. *Review of Communication Research*, 6:30–53.

[373] NVIDIA (2022). Nvidia corporate. `https://www.nvidia.com/`. (Last Accessed: november 2022).

[374] OceanRobotics (2022). Beam pro, gobe robots, oceanrobotics, inc. https://gobe.blue-ocean-robotics.com/robots.

[375] Oculus, R. (2018). Oculus rift cv1. https://www.oculus.com/rift.

[376] Oculus, R. (2019). Pc sdk developer guide. https://developer.oculus.com/documentation/pcsdk/latest/concepts/book-dg/.

[377] Oculus, R. (2020). Pc sdk developer guide, guidelines for vr performance optimization. https://developer.oculus.com/documentation/native/pc/dg-performance-guidelines/.

[378] Oh, C. S., Bailenson, J. N., and Welch, G. F. (2018). A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI*, 5:114.

[379] OhmniLabs (2022). Ohmni telepresence robot, ohmnilabs, inc. https://ohmnilabs.com/.

[380] OpenCV (2022). Opencv (open source computer vision library). `https://opencv.org/`. (Last Accessed: november 2022).

[381] OpenKinect (2014). Openkinect. `https://github.com/OpenKinect`.

[382] OpenNI (2014). Openni. `https://github.com/openni`.

[383] OriginRobotics (2022). Origibot telepresence robot, origin robotics, inc. https://www.originrobotics.com/.

[384] Orlandini, A., Kristoffersson, A., Almquist, L., Björkman, P., Cesta, A., Cortellessa, G., Galindo, C., Gonzalez-Jimenez, J., Gustafsson, K., Kiselev, A., Loutfi, A., Melendez, F., Nilsson, M., Hedman, L. O., Odontidou, E., Ruiz-Sarmiento, J.-R., Scherlund, M., Tiberio, L., von Rump, S., and Coradeschi, S. (2016). ExCITE Project: A Review of Forty-Two Months of Robotic Telepresence Technology Evolution. *Presence: Teleoperators and Virtual Environments*, 25(3):204–221.

[385] Osawa, M., Okuoka, K., Takimoto, Y., and Imai, M. (2020). Is automation appropriate? semi-autonomous telepresence architecture focusing on voluntary and involuntary movements. *International Journal of Social Robotics*, pages 1–16.

[386] OwlLabs (2022). Owl pro, owl labs, inc. https://owllabs.com/.

[387] Pan, X., Gillies, M., and Slater, M. (2008). The impact of avatar blushing on the duration of interaction between a real and virtual person. In *Presence 2008: The 11th Annual International Workshop on Presence*, pages 100–106.

[388] Paris, S. and Durand, F. (2006). A fast approximation of the bilateral filter using a signal processing approach. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part IV*, ECCV'06, pages 568–580, Berlin, Heidelberg. Springer-Verlag.

[389] Park, J.-H., Shin, Y.-D., Bae, J.-H., and Baeg, M.-H. (2012). Spatial uncertainty model for visual features using a kinectâ„¢ sensor. *Sensors*, 12(7):8640–8662.

[390] Park, S. and Woldstad, J. C. (2006). Design of visual displays for teleoperation. In Karwowski, W., editor, *International encyclopedia of ergonomics and human factors*, page 1579–1583, Boca Raton, FL, USA. CRC Press, Inc.

[391] Pathi, S. K., Kiselev, A., and Loutfi, A. (2022). Detecting groups and estimating f-formations for social human-robot interactions. *Multimodal Technologies and Interaction*, 6(3).

[392] Patrao, B. and Menezes, P. (2013). A virtual reality system for training operators. *International Journal of Online Engineering (iJOE)*, 9:53–55.

[393] Patrão, B. and Menezes, P. (2013). A virtual reality system for training operators. *International Journal of Online Engineering*, 9(8):53–55.

[394] Paulos, E. and Canny, J. (1997). Ubiquitous tele-embodiment: Applications and implications. *International Journal of Human-Computer Studies*, 46:861–877.

[395] Paulos, E. and Canny, J. (1998). Prop: personal roving presence. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 296–303.

[396] Paulos, E. and Canny, J. F. (2001). Social tele-embodiment: Understanding presence. *Auton. Robots*, 11(1):87–95.

[397] Pazuchanics, S. L. (2006). The effects of camera perspective and field of view on performance in teleoperated navigation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(16):1528–1532.

[398] Perret, J. and Vander Poorten, E. (2018). Touching virtual reality: A review of haptic gloves. In *ACTUATOR 2018; 16th International Conference on New Actuators, Bremen, Germany, 25-27 June*, pages 1–5.

[399] Petit, B., Lesage, J.-D., Menier, C., Allard, J., Franco, J.-S., Raffin, B., Boyer, E., and Faure, F. (2009). Multicamera real-time 3d modeling for telepresence and remote collaboration. *INTERNATIONAL JOURNAL OF DIGITAL MULTIMEDIA BROADCASTING*, 2010:247108–12.

[400] Pianzola, F. (2021). Presence, flow, and narrative absorption questionnaires: a scoping review. *Open Research Europe*, 1(11):11.

[401] Pimentel, D. and Vinkers, C. (2021). Copresence with virtual humans in mixed reality: The impact of contextual responsiveness on social perceptions. *Frontiers in Robotics and AI*, 8:25.

[402] Pito, R. (1996). Mesh integration based on co-measurements. In *Image Processing, 1996. Proceedings., International Conference on*, volume 1, pages 397 –400 vol.2.

[403] Plüss, C., Ranieri, N., Bazin, J.-C., Martin, T., Laffont, P.-Y., Popa, T., and Gross, M. (2016). An immersive bidirectional system for life-size 3d communication. In *Proceedings of the 29th International Conference on Computer Animation and Social Agents*, pages 89–96.

[404] Portugal, D., Alvito, P., Christodoulou, E., Samaras, G., and Dias, J. (2019). A study on the deployment of a service robot in an elderly care center. *International Journal of Social Robotics*, 11(2):317–341.

[405] Prasad, V., Stock-Homburg, R., and Peters, J. (2021). Human-robot handshaking: A review. *International Journal of Social Robotics*, pages 1–17.

[406] Prats, M., Pérez, J., Fernández, J., and Sanz, P. (2012). An open source tool for simulation and supervision of underwater intervention missions. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2577–2582.

[407] Prewett, M. S., Johnson, R. C., Saboe, K. N., Elliott, L. R., and Coovert, M. D. (2010). Managing workload in human-robot interaction: A review of empirical studies. *Computers in Human Behavior*, 26(5):840 – 856.

[408] PrimeSense (2011). The primesensor (tm) reference design 1.08. `https://www.reuters.com/article/idUS167050511820110112`. (Last Accessed: April 2011).

[409] Quintas, J., Almeida, L., Brito, M., Quintela, G., Menezes, P., and Dias, J. (2012). Context-based understanding of interaction intentions. In *RO-MAN, 2012 IEEE*, pages 515–520.

[410] Quintas, J., Martins, G. S., Santos, L., Menezes, P., and Dias, J. (2019). Toward a context-aware human–robot interaction framework based on cognitive development. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(1):227–237.

[411] Randall E. Bailey, Jarvis James Arthur, S. P. W. (2004). Latency requirements for head-worn display s/EVs applications. volume 5424, pages 5424–5424– 12.

[412] Raposo, C., Barreto, J., and Nunes, U. (2013). Fast and accurate calibration of a kinect sensor. In *3D Vision - 3DV 2013, 2013 International Conference on*, pages 342–349.

[413] Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13:257–266.

[414] Redaelli, C. and Riva, G. (2011). *Flow for Presence Questionnaire*, pages 3–22. Springer.

[415] Reddy, D., Sankaranarayanan, A. C., Cevher, V., and Chellappa, R. (2008). Compressed sensing for multi-view tracking and 3-d voxel reconstruction. In *ICIP*, pages 221–224. IEEE.

[416] Rhee, T., Thompson, S., Medeiros, D., dos Anjos, R., and Chalmers, A. (2020). Augmented virtual teleportation for high-fidelity telecollaboration. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1923–1933.

[417] Rizzo, A. A. and Kim, G. J. (2005). A swot analysis of the field of virtual rehabilitation and therapy. *Presence*, 14(2):119–146.

[418] Robotics, A. and iRobot (2022). Ava 500, ava robotics and irobot, inc. https://www.avarobotics.com/s.

[419] Röijezon, U., Djupsjöbacka, M., Björklund, M., Häger-Ross, C., Grip, H., and Liebermann, D. (2010). Kinematics of fast cervical rotations in persons with chronic neck pain: a cross-sectional and reliability study. *BMC musculoskeletal disorders*, 11.

[420] Roman, N., Wang, D. L., and Brown, G. J. (2003). Speech segregation based on sound localization. *Journal of the Acoustical Society of America*, 114(4):2236–2252.

[421] ROS (2016). Robot operative system. `www.ros.org`.

[422] Ros, R., Baroni, I., and Demiris, Y. (2014). Adaptive human–robot interaction in sensorimotor task instruction: From human to robot dance tutors. *Robotics and Autonomous Systems*, 62(6):707–720.

[423] Ross, B., Bares, J., Stager, D., Jackel, L., and Perschbacher, M. (2007). An Advanced Teleoperation Testbed. In *6th International Conference on Field and Service Robotics - FSR 2007*, volume 42 of *Springer Tracts in Advanced Robotics*, Chamonix, France. Springer.

[424] Rossi, S., Ferland, F., and Tapus, A. (2017). User profiling and behavioral adaptation for hri: A survey. *Pattern Recognition Letters*, 99:3–12.

[425] Roth, P. M. and Winter, M. (2008). Survey of appearance-based methods for object recognition. *Inst. for computer graphics and vision, Graz University of Technology, Austria, technical report ICGTR0108 (ICG-TR-01/08)*.

[426] Rusinkiewicz, S., Hall-Holt, O., and Levoy, M. (2002). Real-time 3d model acquisition. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '02, pages 438–446, New York, NY, USA. ACM.

[427] Rutishauser, M., Stricker, M., and Trobina, M. (1994). Merging range images of arbitrarily shaped objects. In *In Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 573–580.

[428] Ry Andersen, S. (2002). The history of the ophthalmological society of copenhagen 1900–50. 234:6–17.

[429] Saari, T., Laarni, J., Ravaja, N., Kallinen, K., and Turpeinen, M. (2004). Virtual ba and presence in facilitating learning from technology mediated organizational information flows. In *Annual International Workshop on Presence, Valencia, Spain, October 13-14*, pages 133–140. Technical University of Valencia.

[430] Sahbani, A., El-Khoury, S., and Bidaud, P. (2012). An overview of 3d object grasp synthesis algorithms. *Robotics and Autonomous Systems*, 60(3):326–336. Autonomous Grasping.

[431] Sakamoto, D., Kanda, T., Ono, T., Ishiguro, H., and Hagita, N. (2007). Android as a telecommunication medium with a human-like presence. In *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 193–200. IEEE.

[432] Sakamoto, D., Kanda, T., Ono, T., Ishiguro, H., and Hagita, N. (2018). *Androids as a Telecommunication Medium with a Humanlike Presence*, pages 39–56. Springer Singapore, Singapore.

[433] Sanchez-Vives, M. V. and Slater, M. (2005). From presence to consciousness through virtual reality. *Nature reviews. Neuroscience*, 6(4):332–339.

[434] Sappa, A. and Garcia, M. A. (2007). Incremental integration of multiresolution range images. *The Imaging Science Journal*, 55(3):127–139.

[435] Sarabia, M., Lee, K., and Demiris, Y. (2015). Towards a synchronised grammars framework for adaptive musical human-robot collaboration. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 715–721. IEEE.

[436] Schadenberg, B. R., Neerincx, M. A., Cnossen, F., and Looije, R. (2017). Personalising game difficulty to keep children motivated to play with a social robot: A bayesian approach. *Cognitive systems research*, 43:222–231.

[437] Scharstein, D. (1996). Stereo vision for view synthesis. In *In Proc. Computer Vision and Pattern Recognition Conf*, pages 852–858.

[438] Schloerb, D. W. and Sheridan, T. B. (1995). Experimental investigation of the relationship between subjective telepresence and performance in hand-eye tasks.

[439] Schubert, T. (2003). The sense of presence in virtual environments: A three-component scale measuring spatial presence, involvement, and realness. *Zeitschrift für Medienpsychologie*, 15:69–71.

[440] 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 and Behavior*, 4(2):183–201.

[441] Scribner, D. R. and Gombash, J. W. (1998). The effect of stereoscopic and wide field of view conditions on teleoperator performance. In *(Technical Report)*, Army Research Laboratory at Aberdeen Proving Grounds, MD.

[442] Seitz, S. and Dyer, C. (1999). Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 25(1).

[443] Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. (2006). A comparison and evaluation of multi-view stereo reconstruction algorithms. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1:519–528.

[444] Sekmen, A. and Challa, P. (2013). Assessment of adaptive human–robot interactions. *Knowledge-based systems*, 42:49–59.

[445] SELFIEBOT.CO (2022). Selfie bot, selfiebot.co. https://www.selfiebot.co/.

[446] Senft, E., Baxter, P., Kennedy, J., and Belpaeme, T. (2015). Sparc: Supervised progressively autonomous robot competencies. In *International Conference on Social Robotics*, pages 603–612. Springer.

[447] Shannon, C. E. (2001). A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55.

[448] Shen, Q., Dautenhahn, K., Saunders, J., and Kose, H. (2015). Can real-time, adaptive human–robot motor coordination improve humans' overall perception of a robot? *IEEE Transactions on Autonomous Mental Development*, 7(1):52–64.

[449] Sheridan, T. (1992a). *Telerobotics, Automation and Human Supervisory Control*. MIT Press, USA.

[450] Sheridan, T. B. (1992b). Musings on telepresence and virtual presence. *Presence*, 1(1):120–126.

[451] Sheridan, T. B. (2020). A review of recent research in social robotics. *Current Opinion in Psychology*, 36:7–12. Cyberpsychology.

[452] Shi, J. and Tomasi, C. (1994). Good features to track. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, pages 593 –600.

[453] Short, J., Williams, E., and Christie, B. (1976). *The social psychology of telecommunications*. John Wiley & Sons.

[454] Sian, N., Yokoi, K., Kajita, S., Kanehiro, F., and Tanie, K. (2002). Whole body teleoperation of a humanoid robot - development of a simple master device using joysticks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 3, pages 2569–2574 vol.3.

[455] Siciliano, B., Khatib, O., and Kröger, T. (2008). *Springer handbook of robotics*, volume 200. Springer.

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

[457] Slater, M., Brogni, A., and Steed, A. (2003). Physiological responses to breaks in presence: A pilot study. In *Presence 2003: The 6th annual international workshop on presence*, volume 157. Citeseer.

[458] Slater, M., Guger, C., Edlinger, G., Leeb, R., Pfurtscheller, G., Antley, A., Garau, M., Brogni, A., and Friedman, D. (2006). Analysis of physiological responses to a social situation in an immersive virtual environment. *Presence: Teleoper. Virtual Environ.*, 15(5):553–569.

[459] Slater, M., Lotto, B., Arnold, M. M., and Sanchez-Vives, M. V. (2009). How we experience immersive virtual environments: the concept of presence and its measurement. *Anuario de Psicologia*, 40:193–210.

[460] Slater, M., Pertaub, and Steed, A. (1999). Public speaking in virtual reality: Facing an audience of atavars.

[461] Slater, M. and Sanchez-Vives, M. V. (2014). Transcending the self in immersive virtual reality. *Computer*, 47(7):24–30.

[462] Slater, M. and Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, 3.

[463] Slater, M., Spanlang, B., Sanchez-Vives, M. V., and Blanke, O. (2010). First Person Experience of Body Transfer in Virtual Reality. *PLoS ONE*, 5(5):e10564+.

[464] Slater, M., Usoh, M., and Steed, A. (1994). Depth of presence in virtual environments. *Presence-Teleoperators and Virtual Environments*, 3(2):130–144.

[465] Smisek, J., Jancosek, M., and Pajdla, T. (2013). 3d with kinect. In Fossati, A., Gall, J., Grabner, H., Ren, X., and Konolige, K., editors, *Consumer Depth Cameras for Computer Vision*, Advances in Computer Vision and Pattern Recognition, pages 3–25. Springer London.

[466] Smith, J. S., Chao, C., and Thomaz, A. L. (2015). Real-time changes to social dynamics in human-robot turn-taking. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3024–3029. IEEE.

[467] Smyth, C., W Gombash, J., and M Burcham, P. (2001). Indirect vision driving with fixed flat panel displays for near-unity, wide, and extended fields of camera view. In *(Technical Report)*, Army Research Laboratory at Aberdeen Proving Grounds, MD.

[468] SoftBankRobotics (2022). Nao and pepper robots, softbank robotics, lda. https://www.softbankrobotics.com/.

[469] Sormann, M., Zach, C., Bauer, J., Karner, K., and Bishof, H. (2007). Watertight multi-view reconstruction based on volumetric graph-cuts. In Ersball, B. and Pedersen, K., editors, *Image Analysis*, volume 4522 of *Lecture Notes in Computer Science*, pages 393–402. Springer Berlin, Heidelberg.

[470] Soucy, M. and Laurendeau, D. (1995). A general surface approach to the integration of a set of range views. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(4):344–358.

[471] Stanton, C., Bogdanovych, A., and Ratanasena, E. (2012). Teleoperation of a humanoid robot using full-body motion capture, example movements, and machine learning. In *Proc. Australasian Conference on Robotics and Automation*, volume 8, page 51.

[472] Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. *JOURNAL OF COMMUNICATION*, 42:73–93.

[473] Stiefelhagen, R., Fugen, C., Gieselmann, R., Holzapfel, H., Nickel, K., and Waibel, A. (2004). Natural human-robot interaction using speech, head pose and gestures. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, volume 3, pages 2422–2427 vol.3.

[474] Stone, S., Haldar, J., Tsao, S., m.W. Hwu, W., Sutton, B., and Liang, Z.-P. (2008). Accelerating advanced mri reconstructions on gpus. *Journal of Parallel and Distributed Computing*, 68(10):1307 – 1318. General-Purpose Processing using Graphics Processing Units.

[475] Strasburger, H., Rentschler, I., and Jüttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11(5):13.

[476] Strauch, B. (2017). Ironies of automation: Still unresolved after all these years. *IEEE Transactions on Human-Machine Systems*, 48(5):419–433.

[477] Su, H., Hu, Y., Karimi, H. R., Knoll, A., Ferrigno, G., and De Momi, E. (2020). Improved recurrent neural network-based manipulator control with remote center of motion constraints: Experimental results. *Neural Networks*, 131:291–299.

[478] Sualeh, M. and Kim, G.-W. (2019). Simultaneous localization and mapping in the epoch of semantics: a survey. *International Journal of Control, Automation and Systems*, 17(3):729–742.

[479] Sun, Q., Patney, A., Wei, L.-Y., Shapira, O., Lu, J., Asente, P., Zhu, S., Mcguire, M., Luebke, D., and Kaufman, A. (2018). Towards virtual reality infinite walking: Dynamic saccadic redirection. *ACM Trans. Graph.*, 37(4).

[480] Sun, T. and Neuvo, Y. (1994). Detail-preserving median based filters in image processing. *Pattern Recognition Letters*, 15(4):341 – 347.

[481] Sun, Y., Dumont, C., and Abidi, M. A. (2000). Mesh-based integration of range and color images. In *Proc. of SPIE*, pages 110–117.

[482] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, Quebec, Canada, 8-13 December 2014*, pages 3104–3112.

[483] Swinnen, S. P. and Wenderoth, N. (2004). Two hands, one brain: cognitive neuroscience of bimanual skill. *Trends in cognitive sciences*, 8(1):18–25.

[484] Tachi, S. (2016). Telexistence: Enabling humans to be virtually ubiquitous. *IEEE Computer Graphics and Applications*, 36(1):8–14.

[485] Tachi, S., Inoue, Y., and Kato, F. (2020). TELESAR VI: telexistence surrogate anthropomorphic robot VI. *Int. J. Humanoid Robotics*, 17(5):2050019:1–2050019:33.

[486] Takahashi, S., Fujishiro, I., Takeshima, Y., and Nishita, T. (2005). A feature-driven approach to locating optimal viewpoints for volume visualization. In *IEEE Visualization*, page 63. IEEE Computer Society.

[487] Takeshita, H., Kihara, K., Yoshida, S., Higuchi, S., ku Nagoya-shi Ito, M. M., Nakanishi, Y., Kijima, T., Ishioka, J., Matsuoka, Y., Numao, N., Saito, K., and Fujii, Y. (2014). Clinical application of a modern high-definition head-mounted display in sonography. *Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine*, 33 8:1499–504.

[488] Takimoto, Y., Hasegawa, K., Sono, T., and Imai, M. (2017). A simple bi-layered architecture to enhance the liveness of a robot. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2786–2792. IEEE.

[489] Tavares, R., Sousa, P. J., Abreu, P., and Restivo, M. T. (2016). Virtual environment for instrumented glove. In *2016 13th Int. Conf. on Remote Engineering and Virtual Instrumentation (REV)*.

[490] telepresencerobots.com (2022). Telepresence robots shop. https://telepresencerobots.com/robots/orbis-robotics-teleporter/.

[491] Thibos, L. N., Cheney, F. E., and Walsh, D. J. (1987). Retinal limits to the detection and resolution of gratings. *J. Opt. Soc. Am. A*, 4(8):1524–1529.

[492] Thissen, B., Menninghaus, W., and Schlotz, W. (2018). Measuring optimal reading experiences: The reading flow short scale. *Frontiers in Psychology*, 9.

[493] Tittle, J. S., Woods, D. D., Roesler, A., Howard, M., and Phillips, F. (2001). The role of 2-d and 3-d task performance in the design and use of visual displays. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 45(4):331–335.

[494] Tomasi, C. and Kanade, T. (1991). Detection and tracking of point features. Technical report.

[495] Tomasi, C. and Manduchi, R. (1998). Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, pages 839–846.

[496] Tsui, K. and Yanco, H. (2013). Design challenges and guidelines for social interaction using mobile telepresence robots. *Reviews of Human Factors and Ergonomics*, 9:227–301.

[497] Turk, G. and Levoy, M. (1994). Zippered polygon meshes from range images. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, SIGGRAPH '94, pages 311–318, New York, NY, USA. ACM.

[498] Uddin, R. and Ryu, J. (2016). Predictive control approaches for bilateral teleoperation. *Annual Reviews in Control*, 42:82–99.

[499] Usoh, M., Catena, E., Arman, S., and Slater, M. (2000). Using presence questionnaires in reality. *Presence: Teleoperators and Virtual Environments*, 9(5):497–503.

[500] Vallines, I. and Greenlee, M. W. (2006). Saccadic suppression of retinotopically localized blood oxygen level-dependent responses in human primary visual area v1. *Journal of Neuroscience*, 26(22):5965–5969.

[501] Vasudevan, R., Kurillo, G., Lobaton, E. J., Bernardin, T., Kreylos, O., Bajcsy, R., and Nahrstedt, K. (2011). High-quality visualization for geographically distributed 3-d teleimmersive applications. *IEEE Transactions on Multimedia*, 13(3):573–584.

[502] Venture, G. and Kulić, D. (2019). Robot expressive motions: A survey of generation and evaluation methods. *J. Hum.-Robot Interact.*, 8(4).

[503] VGo (2022). Vgo robotic telepresence, vecna technologies, inc. https://www.vgocom.com/.

[504] Viciana-Abad, R., Reyes-Lecuona, A., Rosa-Pujazón, A., and Pérez-Lorenzo, J. M. (2014). The influence of different sensory cues as selection feedback and co-location in presence and task performance. *Multim. Tools Appl.*, 68(3):623–639.

[505] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511 – I–518 vol.1.

[506] von der Pütten, A. M. R., Krämer, N. C., Gratch, J., and Kang, S.-H. (2010). "it doesn't matter what you are!" explaining social effects of agents and avatars. *Comput. Hum. Behav.*, 26:1641–1650.

[507] Wada, T., Wu, X., Tokai, S., and Matsuyama, T. (2000). Homography based parallel volume intersection: Toward real-time volume reconstruction using active cameras. In *Computer Architectures for Machine Perception, 2000. Proceedings. Fifth IEEE International Workshop on 11-13 Sept. 2000*, pages 331–339.

[508] Waizenegger, W., Feldmann, I., Eisert, P., and Kauff, P. (2009). Parallel high resolution real-time visual hull on gpu. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 4301 –4304.

[509] Wang, D. and Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726.

[510] Ware, C. and Balakrishnan, R. (1994). Reaching for objects in vr displays: lag and frame rate. *ACM Trans. Comput.-Hum. Interact.*, 1:331–356.

[511] Watson, B., Walker, N., Hodges, L. F., and Reddy, M. (1997). An evaluation of level of detail degradation in head-mounted display peripheries. *Presence: Teleoper. Virtual Environ.*, 6(6):630–637.

[512] Watson, B., Walker, N., Ribarsky, W., and Johnson, V. A. S. (1998). Effects of variation in system responsiveness on user performance in virtual environments. *Human factors*, 40 3:403–14.

[513] Watson, B., Walker, N., Woytiuk, P., and Ribarsky, W. (2003). Maintaining usability during 3d placement despite delay. In *IEEE Virtual Reality, 2003. Proceedings.*, pages 133–140.

[514] Whelan, T., Kaess, M., Fallon, M., Johannsson, H., Leonard, J., and McDonald, J. (2012). Kintinuous: Spatially extended KinectFusion. In *RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras*, Sydney, Australia.

[515] wikipedia (2022). Sagittal planes. `https://en.wikipedia.org/wiki/Sagittal_plane`.

[516] Witmer, B. G., Jerome, C. J., and Singer, M. J. (2005). The factor structure of the presence questionnaire. *Presence: Teleoperators and Virtual Environments*, 14(3):298–312.

[517] Witmer, B. G. and Sadowski, W. J. (1998). Nonvisually guided locomotion to a previously viewed target in real and virtual environments. *Human Factors*, 40(3):478–488.

[518] Witmer, B. G. and Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoper. Virtual Environ.*, 7:225–240.

[519] Woodford, O. J., Reid, I. D., Torr, P. H. S., and Fitzgibbon, A. W. (2007). On new view synthesis using multiview stereo. In *BMVC*. British Machine Vision Association.

[520] Woods, D., Tittle, J., Feil, M., and Roesler, A. (2004). Envisioning human-robot coordination in future operations. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 34(2):210–218.

[521] Xiang, X. and Foo, S. (2021). Recent advances in deep reinforcement learning applications for solving partially observable markov decision processes (pomdp) problems: Part 1—fundamentals and applications in games, robotics and natural language processing. *Machine Learning and Knowledge Extraction*, 3(3):554–581.

[522] Xsens (2022). Xsens motion technologies. http://www.xsens.com. `https://www.xsens.com/`. (Last Accessed: november 2022).

[523] Ye, C., Yang, Y., Mao, R., Fermüller, C., and Aloimonos, Y. (2017). What can i do around here? deep functional scene understanding for cognitive robots. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4604–4611.

[524] Yeh, M. and Wickens, C. (2001). Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human factors*, 43:355–65.

[525] Yguel, M., Aycard, O., and Laugier, C. (Oct. 2006). Efficient gpu-based construction of occupancy grids using several laser range-finders. Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on.

[526] Yi Ma, Stefano Soatta, J. K. and Sastry, S. S. (2004). *An invitation to 3D vision*. Springer.

[527] Yous, S., Laga, H., Kidode, M., and Chihara, K. (2007). Gpu-based shape from silhouettes. In *Proceedings of the 5th international conference on Computer graphics and interactive techniques in Australia and Southeast Asia*, GRAPHITE '07, pages 71–77, New York, NY, USA. ACM.

[528] Zahorik, P. and Jenison, R. L. (1998). Presence as being-in-the-world. *Presence: Teleoper. Virtual Environ.*, 7(1):78–89.

[529] Zhan, Z. and Mei, H. (2013). Academic self-concept and social presence in face-to-face and online learning: Perceptions and effects on students' learning achievement and satisfaction across environments. *Computers & Education*, 69:131–138.

[530] Zhang, B. and Li, Y. (2005). An efficient method for dynamic calibration and 3d reconstruction using homographic transformation. *Sensors and Actuators A: Physical*, 119(2):349 – 357.

[531] Zhang, C. and Zhang, Z. (2011). Calibration between depth and color sensors for commodity depth cameras. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pages 1–6. IEEE.

[532] Zhang, G. (2021). *Gaze-controlled Telepresence: Accessibility, Training and Evaluation*. PhD thesis, DTU Management,Technical University of Denmark.

[533] Zhang, K., Zhong, G., Dong, J., Wang, S., and Wang, Y. (2019). Stock market prediction based on generative adversarial network. *Procedia computer science*, 147:400–406.

[534] Zhang, Q.-B., Wang, H.-X., and Wei, S. (2003). A new algorithm for 3d projective reconstruction based on infinite homography. In *Machine Learning and Cybernetics, 2003 International Conference on, IEEE*.

[535] Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(11):1330–1334.

[536] Zhang, Z. (2012). Microsoft kinect sensor and its effect. *IEEE MultiMedia*, 19(2):4–10.

[537] Zhang, Z. and Hanson, A. R. (1996). 3d reconstruction based on homography mapping. In *In ARPA Image Understanding Workshop*.

[538] Zhao, S. (2003). Toward a taxonomy of copresence. *Presence: Teleoper. Virtual Environ.*, 12:445–455.

[539] Zhu, D., Gedeon, T., and Taylor, K. (2011). Exploring camera viewpoint control models for a multi-tasking setting in teleoperation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 53–62, New York, NY, USA. ACM.

[540] Ziegler, G. (2010). *GPU Data Structures for Graphics and Vision*. PhD thesis, Max-Planck-Institut für Informatik.