University of Trento

Department of

Industrial Engineering

~ · ~

Academic Year 2022–2023

Measurement technologies to enhance human capabilities via Mixed Reality

Supervisor
Prof. Mariolino De Cecco
Prof. Paolo Bosetti
Prof. Andrea Del Prete

Graduate Student
Alessandro Luchetti
223399

Thesis submitted for the degree of Doctor of Philosophy - October 2023

To all the people who never stopped believing in me.

Declaration

I declare that, except for specific references to the work of others, the content of this thesis is original and has not been submitted, in whole or in part, for consideration for any other degree or qualification in this or any other university. This thesis includes scientific papers that I have co-authored and that have been published in journals or presented at international conferences:

> De Cecco, M., Luchetti, A., Butaslac III, I., Pilla, F., Guandalini, G. M. A., Bonavita, J., ... & Hirokazu, K. (2023). Sharing Augmented Reality between a Patient and a Clinician for Assessment and Rehabilitation in Daily Living Activities. Information, 14(4), 204.
>
> Luchetti, A., Zanetti, M., Kalkofen, D., & De Cecco, M. (2022). Omnidirectional camera pose estimation and projective texture mapping for photorealistic 3D virtual reality experiences. Acta IMEKO, 11(2).
>
> Zanetti, M., Luchetti, A., Maheshwari, S., Kalkofen, D., Ortega, M. L., & De Cecco, M. (2022). Object Pose Detection to Enable 3D Interaction from 2D Equirectangular Images in Mixed Reality Educational Settings. Applied Sciences, 12(11), 5309.
>
> Luchetti, A., Tomasin, P., Fornaser, A., Tallarico, P., Bosetti, P., & De Cecco, M. (2019). The human being at the center of smart factories thanks to augmented reality. In 2019 IEEE 5th International forum on Research and Technology for Society and Industry (RTSI) (pp. 51-56). IEEE.

It also includes demos presented at international conferences:

> Luchetti, A., Zaninotto, S., De Cecco M., Guandalini, G. M., Fujimoto Y., & Hirokazu K. (2023). Augmented Reality-based demo for Immersive training in horticultural therapy. In 2023 IEEE International Symposium on Mixed and Augmented Reality
>
> Luchetti, A., & De Cecco M (2022), Interactive Augmented Reality loaded pallet shape checking experience. In 2022 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence, and Neural Engineering.

Alessandro Luchetti
October 2023

Abstract

In the relentless pursuit of human progress, technological innovations have always played a key role in shaping our society. In recent years, the convergence of innovative measurement technologies with Mixed Reality (MR) has emerged as a groundbreaking paradigm, offering transformative solutions to enhance individuals across multiple domains. This dissertation focuses on MR-based applications designed and optimized to restore human centrality, fostering advances in healthcare, education, and industry. The primary purpose is to provide end users with the best tools to enhance their perception-action loop in work or education, empowering them to better achieve and control their final goals. Through the synergy between immersive visualization technologies and a framework based on innovative measurement systems, unique environments are created to enhance end users at different levels of the perception-action loop, leading to improved outcomes and overall well-being. Measurement technologies include three-dimensional cameras, wearable sensors, inertial sensors, thermal cameras, and pressure matrices. Many challenges were overcome in this dissertation, such as designing and testing the proper measurement frame and interface, finding new calibration procedures for measurement systems, and developing original data processing techniques in computer vision and machine learning.

Contents

Nomenclature

$2D$	Two-dimensional
$3D$	Three-dimensional
AAE	Augmented AutoEncoder
ADL	Activity of Daily Living
AI	Artificial Intelligence
$AMPS$	Assessment of Motor and Process Skills
AR	Augmented Reality
AV	Augmented Virtuality
BLE	Bluetooth Low Energy
CT	Computed Tomography
DNN	Deep Neural Network
DoF	Degrees of Freedom
ECG	Electrocardiogram
EMG	Electromyography
EPM	Extent of Presence Metaphor
EWK	Extent of World Knowledge
FOV	Field of View
HCD	Human-Centered Design
HMD	Head-Mounted Display
IMU	Inertial Measurement Unit
IoT	Internet of Things
$LPIPS$	Learned Perceptual Image Patch Similarity
$MCTS$	Monte Carlo Tree Search

$MDPs$ Markov Decision Processes

$MiroLab$ Measurements, Instrumentations and Robotics Laboratory

$MQTT$ Message Queuing Telemetry Transport

MR Mixed Reality

MRI Magnetic Resonance Imaging

$NeRF$ Neural Radiance Fields

ODS Omni-Directional Stereo

OT Occupational Therapy

PAL Perception-Action Loop

POI Point of Interest

$PSNR$ Peak Signal-to-Noise Ratio

PSO Particle Swarm Optimization

$PTSD$ Post-Traumatic Stress Disorder

RF Reproduction Fidelity

SAR Shared Augmented Reality

SAV Shared Augmented Virtuality

SMR Shared Mixed Reality

SSD Shot Multibox Detector

$SSIM$ Structural Similarity Index Measure

SVD Singular Value Decomposition

ToF Time-of-Flight

UDP User Datagram Protocol

VR Virtual Reality

WSN Wireless Sensor Networks

Chapter 1

Introduction

In the ever-changing landscape of technological advances, emerging technologies have continuously influenced and shaped human enhancement, redefining our relationship with the external environment. One of these transformative paradigms is the convergence of innovative measurement technologies with Mixed Reality (MR), opening up new avenues for empowering individuals. At the heart of this renaissance lies the philosophy of Human-Centered Design (HCD). It is an approach that places human users' needs, preferences, and experiences at the center of the design process [21]. It emphasizes understanding and empathizing with users to create innovative and practical solutions to their problems. In recent years, HCD has been greatly influenced and enhanced by integrating innovative technologies [46].

Innovative technologies, such as Artificial Intelligence (AI), Mixed Reality (MR), the Internet of Things (IoT), and wearable devices, have expanded the possibilities and capabilities of HCD. These technologies enable designers to gather more prosperous and nuanced user data, create immersive and interactive experiences, and develop personalized solutions that meet individual needs.

One key aspect of HCD with innovative technologies is the collection and analysis of user data. Technologies like AI and IoT can capture vast amounts of data from various sources, such as user interactions, biometric measurements, and environmental conditions. This data can provide insights into user behaviour, preferences, and pain points, helping designers understand their target audience more deeply.

MR is another set of technologies that have revolutionized HCD. They allow designers to create realistic simulations of environments and experiences, enabling users to interact and provide feedback in a controlled and immersive setting. It facilitates early-stage testing and iteration, leading to more refined and user-friendly designs.

Wearable devices have also played a significant role in HCD. They can collect real-time data about user activities, health metrics, and environmental factors. This information helps designers understand user contexts and design solutions that seamlessly integrate into users' daily lives.

HCD, with innovative technologies, is about more than just creating user-friendly interfaces or visually appealing products. It aims to create holistic experiences that address user needs and aspirations. By leveraging innovative technologies, designers can create solutions that are not only functional but also intuitive, emotionally engaging, and socially responsible.

To successfully implement HCD with innovative technologies, designers must adopt an iterative and collaborative approach. They should involve users throughout the design process, conducting user research, gathering feedback, and refining their designs based on user insights. Balancing technological advancements with a deep understanding of human behaviour, cultural

factors, and ethical considerations is crucial.

This dissertation addresses several MR-based applications designed and optimized to restore human centrality by enhancing users and fostering progress in different domains. Many challenges were overcome in this dissertation, such as designing and testing the proper measurement frame and interface, finding new calibration procedures for measurement systems, and developing original data processing techniques in computer vision and machine learning. In particular, the sections in this chapter focus on the impact of emerging technologies on the human perception-action loop and how it can be enhanced through these technologies. The following three chapters deal with designing frameworks and identifying solutions to address challenges arising from various MR-based applications in health care, education, and industry domains. Specifically, Chapter 2 presents the results of Shared Augmented Reality (SAR) frameworks to enhance both the therapist and the patient; Chapter 3 discusses computer vision algorithms developed for applications in education that aim to increase immersiveness and interaction; Chapter 4 shows how real-time Augmented Reality (AR) feedback ensures a controlled and efficient work environment in the industry. The final section outlines the conclusions and future work.

1.1 Perception-Action Loop (PAL)

The fundamental pillar on which the entire dissertation is based concerns the concept of the perception-action loop and how to try to augment humans in this loop. The perception-action loop (PAL) [30], often discussed in the literature on sequential decision making [61, 56], is a concept in cognitive science and robotics that describes how humans typically interact with the external world. It is a theoretical framework that highlights the continuous cycle of perceiving sensory information from the environment, processing it, and then taking appropriate action based on it, Figure 1.1.

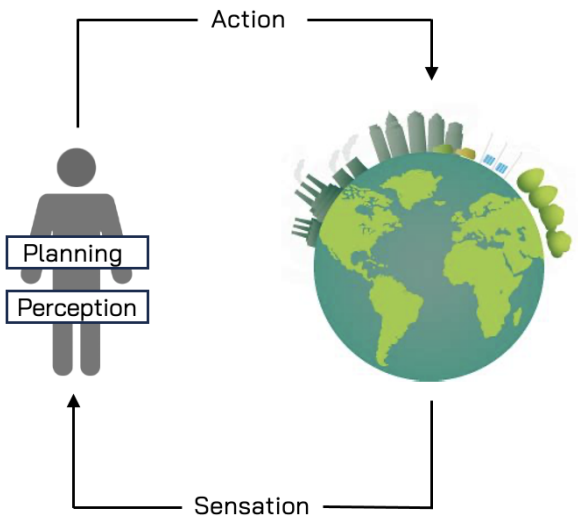


Figure 1.1: The human perception-action loop.

In humans, PAL starts with the sensation stage, where sensory organs such as the eyes, ears, skin, and other senses gather information about the surrounding environment. This sensory

information is processed and interpreted by the brain in the perception stage, allowing us to understand and make sense of what we perceive. Once the information is processed, the planning stage follows. It occurs in the brain and involves cognitive processes that allow us to generate a sequence of actions to achieve a desired goal. This phase incorporates higher-level cognitive functions, such as decision-making, problem-solving, and strategizing. During the planning phase, humans integrate the sensory information received during perception with their knowledge, memories, and understanding of the environment. They evaluate potential courses of action, consider the potential consequences of each option, and select the most suitable plan to achieve their goals. The planning phase also involves setting priorities, weighing different factors, and taking into account constraints or limitations. Humans can anticipate future events, plan multiple steps ahead, and adjust their plans based on evolving circumstances. The loop ends with the actions stage, where the brain generates appropriate motor commands sent to the muscles, enabling us to respond and interact with the environment. This action can be as simple as reaching for an object or as complex as performing intricate movements or making decisions based on perceived information. Importantly, our actions have consequences, which then feedback into the perception stage. For example, if we reach for an object and successfully grasp it, the tactile feedback from our fingers provides additional sensory information that influences our perception. This feedback loop helps refine and adjust our subsequent actions and perceptions, allowing us to learn and adapt to our environment.

The same concept of PAL can be applied to a single device or a complex machine responsible for interacting with and responding to the environment continuously and adaptively. In this case, PAL follows a similar principle but is implemented differently. Instead of human sensory organs, devices or machines use sensors such as cameras, microphones, pressure sensors, or other specialized devices to collect data about the environment. The sensory data is then processed by algorithms and software, which analyze and extract meaningful information from the raw sensor inputs. As with humans, this processed information is then used for the planning stage. It can vary depending on the complexity and capabilities of the system. In simpler systems, the planning may be pre-determined and programmed by human designers. For example, a robotic arm in a manufacturing plant may have a predefined set of actions programmed to perform specific tasks. However, the planning phase can involve sophisticated algorithms and techniques in more advanced systems, such as autonomous robots. These systems can analyze the perceived data, interpret the context, and generate plans based on predefined rules, learned behaviours, or optimization algorithms. Machine planning can range from basic rule-based decision-making to more complex approaches like search algorithms, Markov Decision Processes (MDPs), reinforcement learning, or even advanced planning techniques like Monte Carlo Tree Search (MCTS) [7]. During the planning phase, machines evaluate potential actions, consider their objectives or goals, and select an optimal or near-optimal plan based on the available information. They may also consider environmental factors, resource constraints, safety considerations, and predefined rules or policies. The generated plan is then executed during the action stage through appropriate responses or actions, which can involve controlling motors, actuators, or other mechanisms. Like humans, machines can also receive feedback about the consequences of their actions. This feedback can come from additional sensors or through interactions with the environment. The feedback is then used to adjust subsequent perceptions and actions, enabling machines to improve performance and adapt to changing conditions.

It is important to note that while PAL in devices or machines can be designed and programmed by humans, some advanced systems employ machine learning and AI techniques to enhance their perception and action capabilities. Without explicit human programming, these systems can learn from the data they perceive and optimize their actions over time.

PAL is a fundamental concept in human cognition and machine functionality. It allows

for the continuous flow of information and action, enabling adaptive behaviour and interaction with the environment. The enhancement of these loops can be described as "any attempt to temporarily or permanently overcome the current limitations of the human capabilities (physical and cognitive) through natural and/or artificial means" [4].

1.2 Mixed Reality (MR)

Human enhancement focuses on humans' physical, cognitive, and perceptual augmentation through technology. Cognitive enhancement can be achieved either pharmacologically [29] or, less invasively, via AR or MR in general [28, 66].

The definition of MR can be traced back to 1994 in a paper written by Paul Milgram and Fumio Kishino [74]. Milgram and Kishino define Mixed Reality as blending real and virtual worlds somewhere along the "reality-virtuality continuum", which connects completely real environments to completely virtual ones, Figure 1.2.

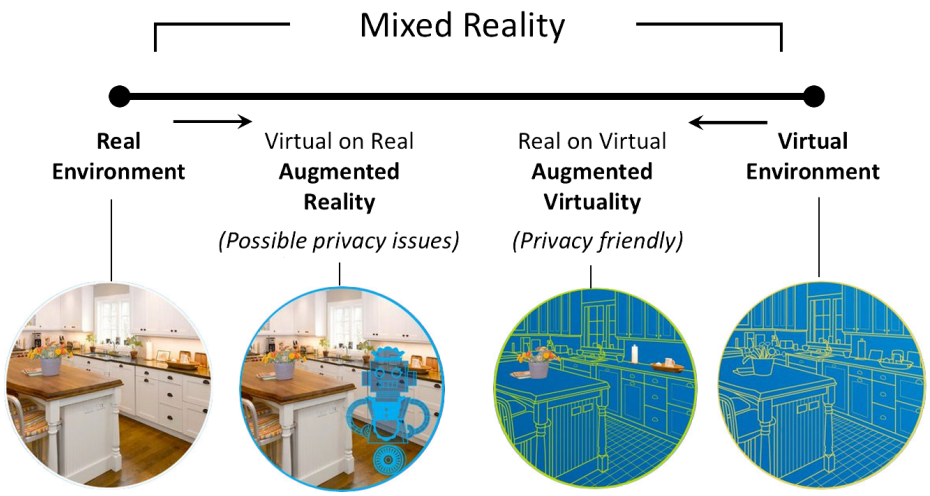


Figure 1.2: Reality-Virtuality continuum.

In other words, the MR continuum comprises the following domains:

- The real world;
- Augmented Reality that can augment reality with digital content. Usually, reality calls for a real-time engine able to elaborate information to provide feedback to a human agent operating in this context;
- Augmented Virtuality (AV) that can insert real (measured) cues into virtual environments. Those real elements can be, in most of the applications, elaborated offline. In the virtual domain, real features related to privacy can be filtered;
- Virtual environments, i.e., Virtual Reality (VR) world.

Within the PAL, users' sensation can be dramatically increased via MR. For example, data acquired with sensor networks of very different sensory types can be simultaneously fed to a user in real-time. Elaborating this large amount of data can, via AI, parallel the human brain

and show the results with low latency in AR. In this way, as shown in Figure 1.3 emerging technologies can enhance human sensation. Among these technologies there are for example three-dimensional (3D) time-of-flight (ToF) cameras, pressure matrices, wearable devices for physiological parameters, MR devices. These technologies together with new data elaboration hardware/techniques such as DNN and machine learning, can augment human capabilities by enhancing his sensation or interact directly with the external world with an action.

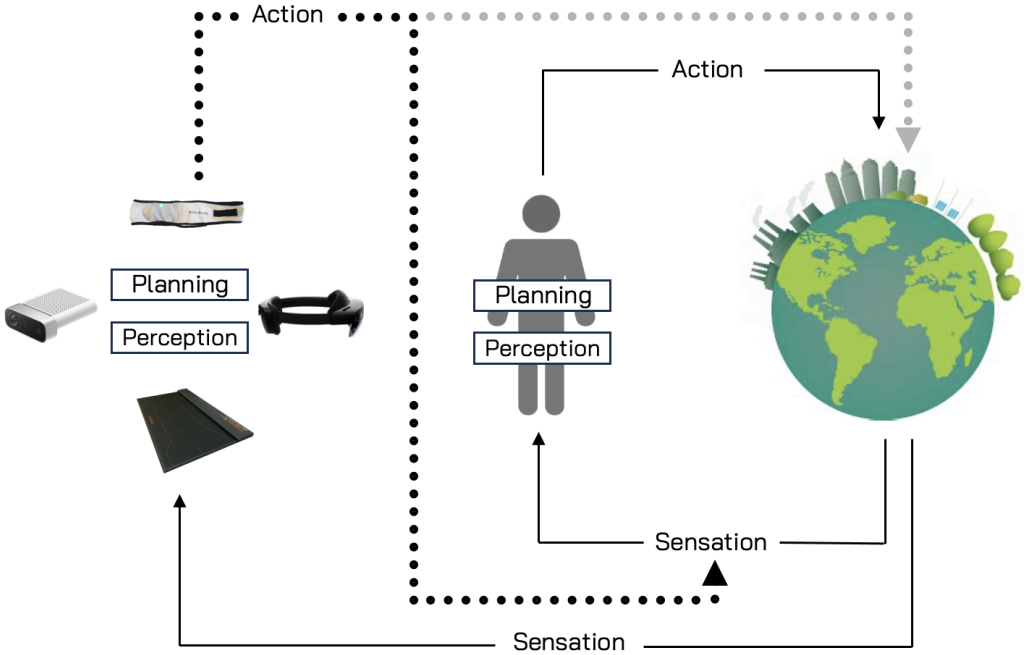


Figure 1.3: Human perception-action loop augmented by a parallel framework composed by measurement and visualization devices.

Thanks to the superimposition of digital content on top of the real cues, human sensation is inherently enhanced with all the basic cognitive functions [94]. The pattern of neuropsychological functions comprises memory, attention, orientation, executive functions, language, visual perception, and motion. Each can be enhanced in several ways, as shown in Table 1.1.

1.2.1 Taxonomy

In the MR continuum, a set of distinctions can be analyzed, which are also evident from the different classes of MR technologies already presented. Generically, the distinctions made were based on whether the primary world comprises real or virtual objects (AR or AV), whether real objects are viewed directly or non-directly, whether the viewing is exocentric or egocentric, and whether or not there is an orthoscopic mapping between the real and virtual worlds. In the paper of Milgram and Kishino [74], they extended those considerations to classify MR experiences by transforming them into a formalized taxonomy, which attempts to address the following questions:

1. How much do we know about the world being displayed?
2. How realistically are we able to display it?

Macro function	Specific functions	What is it?	How can MR enhance human?
Orientation	Spatial Orientation	Ability to organize the self-perception in space and in time.	Using AR, it is possible to combine street view data with GPS position and your phone's camera to determine exactly where you are and what direction you are facing.
	Temporal Orientation	Orient oneself of "where and when" in specific situations	
Attention	Vigilance/Alertness	Activations level of arousal.	MR can provide specific stimuli to trigger a proper reaction in time.
		Physiological quickness in stimulus responses.	
	Selective attention	Selection of one attention target and inhibition of distractors.	MR can highlight the human field of view exactly on the current target by means of object recognition systems.
	Sustained attention	To preserve attention in a long time.	
	Divided attention	Division of attentive resources between many simultaneous stimulus/tasks.	In general, division of attentive resources can be easily achieved via multitasking software architectures while MR can switch human attention by haptic feedback/visual cues/animation etc.
Memory	Short Term Memory (MBT) and	MBT: memory and recall of just presented information	Working memory can be improved via step-by-step textual information or visual cues that exploit, for example, spatially aligned animations.
	Working memory (WM)	WM: ability of keep in memory all the information necessary to finalize a task	
	Long Term Memory (MLT)	Capacity to organize events, learning, situation awareness, ability to retrieve information when necessary.	Long term memory can be improved having the possibility to recall whichever information stored in modern databases or annotations left by other users.
Executive functions	Planning	To be able to organize own actions in relation to the environmental requests while inhibiting automated behaviour	VR can be used to evaluate person's ability to carry out tasks that depend on executive functions and offer the opportunity to train such functions in virtual reality in a safety way through the use of serious games.
	Attention Control – To inhibit inappropriate responses		
	Set Shifting – Cognitive flexibility	To plan strategies for problem solving	
	Abstraction	Abstraction and classification of stimulus and events.	
	Motivation	Willingness to begin many actions.	
Language	Verbal production	Ability to produce understandable verbal messages	Language, both in verbal production and oral comprehension, can benefit from the current technologies of "text to speech with natural sounding voices" and "automatic speech recognition & translation online".
	Oral comprehension	Ability to understand verbal messages	
Visual perception	Object	Ability to recognize objects, through visual channel	Human visual perception is limited to the visible light. This can be a limitation in certain scenarios e.g. firefighting, i.e. smoky views during a fire or occluded, foggy, cloudy. Fortunately, visual perception can be improved by using high sensitivity cameras or multispectral imaging fed to a human operator via MR technologies (AR or virtual reality glasses, projectors, tablets etc). Furthermore, 3D vision can be used to improve objects recognition in combination with MR
	Space	Ability to elaborate (recognize and map) the surrounding space	
Motion	Executive	Patterns of motor behaviour in relation of space and living agents	MR is able to enhance human motor behaviour by feeding in real time the output of a motion capture system.
	Strategical		

Table 1.1: Pattern of neuropsychological functions enhanced via Mixed Reality [66].

3. What is the extent of the illusion that the observer is present within that world?

The three questions described three dimensions:

1. Extent of World Knowledge
2. Reproduction Fidelity
3. Extent of Presence Metaphor (immersive technologies)

While the first dimension deal with the amount of information available about the relevant data such as environment, user viewpoint, gesture, objects location, and classification, the second and third dimensions both attempt to deal with the issue of realism in MR displays, but in different ways: in terms of image quality and terms of immersion, or presence, within the display.

Extent of world knowledge dimension

The Extent of World Knowledge (EWK) dimension is illustrated in Figure 1.4, where it has been broken down into three main groups. These divisions are due to the different amounts of knowledge the display computer holds about object shapes and locations within the two global worlds being presented.

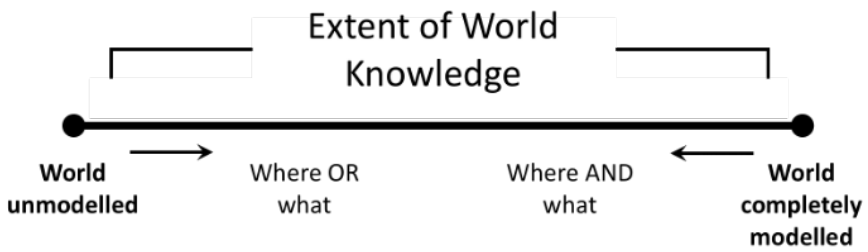


Figure 1.4: Extent of World Knowledge dimension.

At one extreme, on the left, is the case in which nothing is known about the world being displayed. This end of the continuum is reserved for images of objects blindly scanned and synthesized for non-direct "manipulation" in the MR world. The other end of the EWK dimension defines the conditions necessary for displaying a completely virtual world on top of the real one, which can be achieved only when the computer has complete knowledge about the environment, the identification/classification of each real object, its location, the location and viewpoint of the observer and, when relevant, the viewer's attempts to change that world by manipulating objects within it. The mid-section of the EWK continuum is the portion that covers all cases between the two extrema. The different types of subcases are based on two interrogative particles. The first, "Where", refers to cases in which some quantitative data about locations in the remote world are available. Imagine an obstacle avoidance application where raw scanned data obtained, for example, from sonar scanners, detect a blob of something that can be generically classified as an obstacle. The second, "What", refers to cases in which the control software has instead identified/classified objects in the image but has only a vague estimation of their location. This information can be easily achieved via real-time classifiers such as Deep Neural Network (DNN). In the AR domain, to have the digital content, such as indications, structure, and annotations, accurately superimposed in real-time with the environment, a high EWK value is needed. An

example of high EWK hardware can be the Microsoft HoloLens. It is an AR headset capable of performing SLAM (Simultaneous Localization and Mapping) that enables users to view and interact with digital content in the context of their physical surroundings, creating immersive and interactive experiences. On a lower EWK, there is, for example, the DreamGlass Air. It is a simple AR streaming device that makes the final viewing more immersive or private, regardless of the surrounding space.

Reprodution fidelty dimension

The term "Reproduction Fidelity" (RF) refers to the quality with which the synthesizing display can faithfully reproduce real and virtual objects. It lumps together several different factors that are shown in Figure 1.5 through two progressions:

1. the progression above the axis is meant to show a rough progression of video reproduction technology;
2. the one below is towards more sophisticated computer graphic modeling and rendering techniques.

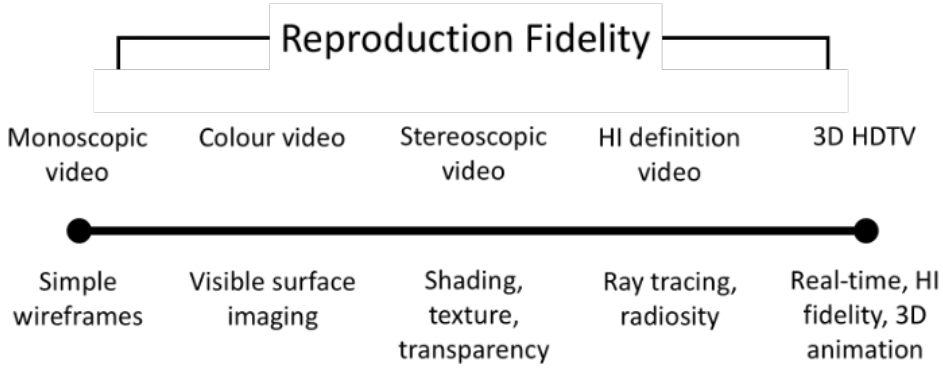


Figure 1.5: Reprodution Fidelty dimension.

Extent of presence metaphor

The third dimension in Figure 1.6 is the Extent of Presence Metaphor (EPM) axis, that is, the extent to which the observer is intended to feel "present" within the displayed scene. In other words, this dimension quantifies sensorial immersion.

As already noted, the EPM axis is not entirely orthogonal to RF since each dimension independently tends towards an extremum which ideally is indistinguishable from viewing reality directly. In the case of EPM, the axis spans a range of cases extending from the metaphor by which the observer peers from outside into the world from a single fixed monoscopic viewpoint up to the metaphor of "real-time imaging", by which the observer's sensations are ideally no different from those of unmediated reality thanks to a multiscopic viewpoint dependent imaging.

1.2.2 MR technologies

With the development of more high-performing and accessible MR technologies, the potential for their use in different applications is advancing continuously. In most situations, using AR

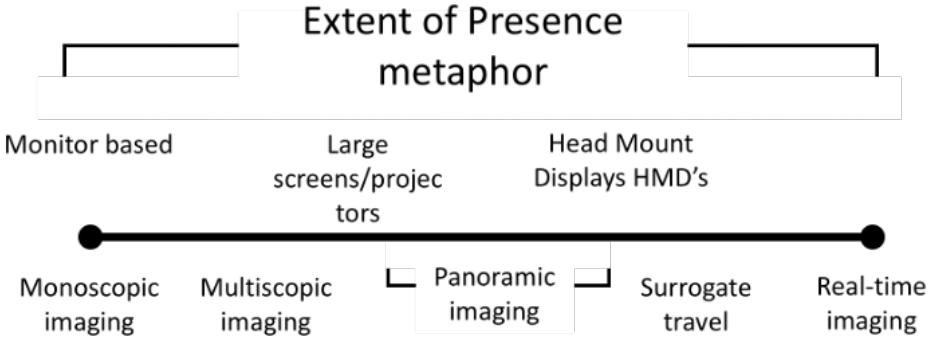


Figure 1.6: Extent of Presence Metaphor dimension.

technologies rather than other technologies in the MR spectrum is based on the context and the intended application [53]. Here are some reasons why AR is often used:

- Interaction with the real world: AR overlays virtual elements onto the real world, allowing users to interact with their physical environment while enhancing it with digital content. This integration with reality can benefit various fields, such as real-time assessment, training, and productivity, where users must engage with their surroundings. Moreover, in terms of self-to-environment-related movements as hand-eye coordination in AR involves much less cognitive load than, for example, VR.

- Enhanced situational awareness: AR can provide users with additional information about their environment, enabling them to make informed decisions or perform tasks more effectively. For instance, AR can overlay real-time data, such as directions, instructions, or sensor information, onto the user's view, enhancing their situational awareness and facilitating complex operations.

- Practical applications: AR has numerous practical applications across industries [17]. It can be used in fields like healthcare for surgical planning and visualization, architecture and design for virtual prototyping, retail for virtual try-on experiences, and gaming for immersive gameplay that combines virtual and real-world elements.

- Social interaction: AR can foster social interactions by enabling users to share augmented experiences with others. Multiple users can see and interact with the same digital content in a shared physical space. This aspect opens up collaborative work, multiplayer gaming, and interactive storytelling possibilities.

- Accessibility and portability: AR experiences can be accessed through various devices, including smartphones and tablets, which are widely available to the general public. Compared to VR, which often requires dedicated headsets and setups, AR can reach a broader audience, making it more accessible and portable.

More generally, both AR and VR have strengths and limitations. VR provides highly immersive and fully virtual experiences that can be advantageous in certain contexts, such as simulations, training scenarios, or entertainment where complete immersion is desired. Ultimately, the choice between AR and VR depends on the specific use case, goals, and user requirements.

Several MR technologies span the Reality-Virtuality continuum, Figure 1.7. Starting from the virtual side, we find technologies that immerse the user inside virtual environments. Then we find the AV domain, where the virtual world can be enhanced by the data coming from environmental sensors. Finally, we find the AR domain where the user percepts reality with digital content in overlay through different devices whose immersiveness, i.e., the extent of presence, can differ.

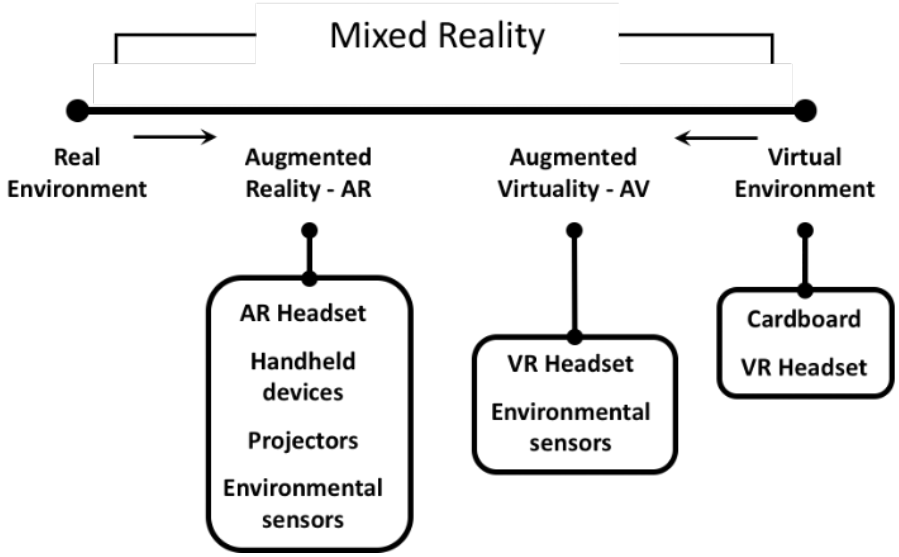


Figure 1.7: Technologies along the Reality-Virtuality continuum.

One of the main differences between AR and AV/VR technologies is that in AR, it is necessary to align the digital content with the user viewpoint and, therefore, a measurement system able to estimate the relative position between the used head and the object of interest must be embedded within the solution. In AV/VR, it is "only" needed to estimate the user viewpoint motion with respect to a fixed reference system that, in turn, pairs the user motion within the AV/VR domain.

1. AV-VR technologies:
   - Google Cardboard, Figure 1.8, and Samsung's VR Headset are VR platforms that use a head mount board with a smartphone.

Figure 1.8: Cardboard. © Google via the Google Cardboard website

   - VR headsets replace the user's natural environment with virtual reality content consisting of a full 360° reconstructed VR environment (3D mesh with images on top)

that allows the user to turn and look around, just as in the physical world. Examples of VR headsets are Oculus Quest and HTC Vive (Figure 1.9).

Figure 1.9: HTC Vive headset, two controllers, and two motion capture. © HTC via HTC VIVE website

2. AV technologies:

- Environmental sensors can be whichever kind of sensor that can be integrated into the actual environment in order to monitor the environment and capture both human interaction and the human state, Figure 1.10. The most crucial technological issue is how sensor data are collected in the MR domain. One of the possible architectures to collect sensor data is Wireless Sensor Networks (WSN).

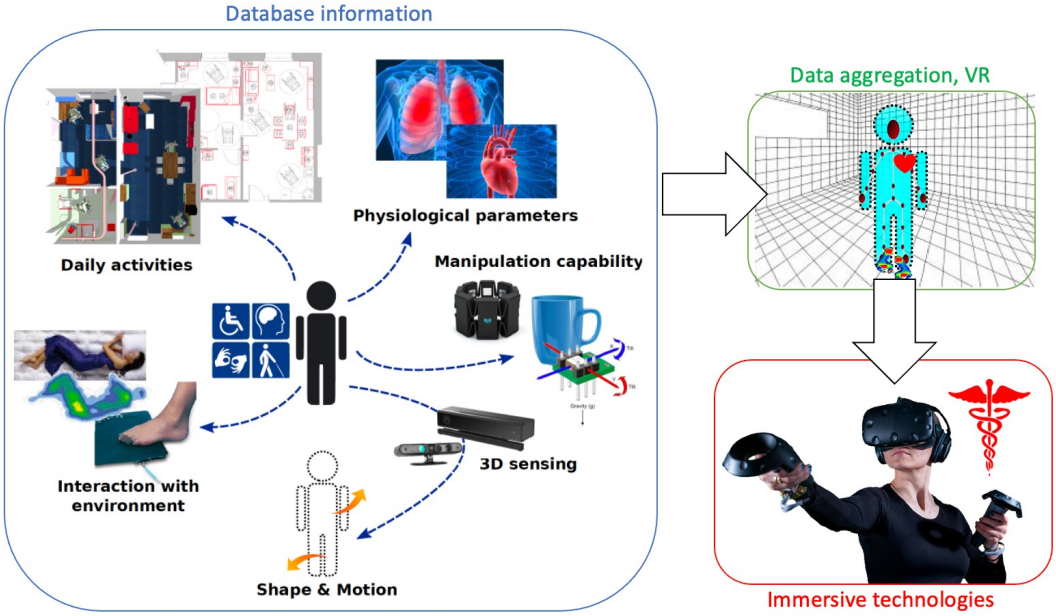


Figure 1.10: Example of data collection and visualization with immersive technologies in the Italian AUSILIA project [120].

3. AR technologies:

   - AR can be experienced via a wearable glass device, head-mounted device (HMD), or through handheld (such as with smartphone) applications. One of the best-known examples of HMD is the Microsoft HoloLens, Figure 1.11.

Figure 1.11: HoloLens 2. © Microsoft via Microsoft website

   - Heads-up displays (HUDs) are another category of devices that can support AR experiences, Figure 1.12. HUDs are designed to present information or digital content to the user in a way that allows them to keep their attention on the real-world environment. While HUDs are often associated with vehicles, such as cars and aircraft, they can also be used in other contexts to provide AR information.

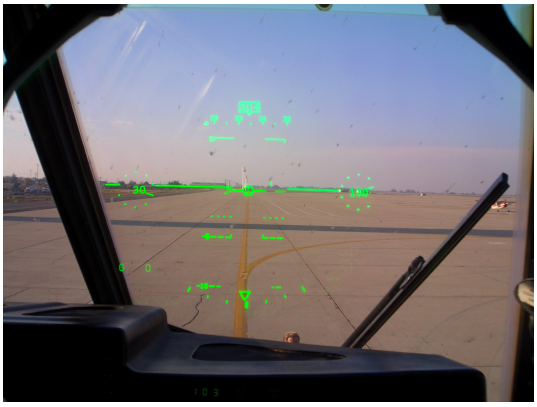


Figure 1.12: C-130J: Co-pilot's head-up display. © Telstar Logistics via Telstar Logistics website

   - Video projector creates AR with no bulky headset. It is possible by projecting the digital content directly on top of the relevant scenario. In 2019, the system Lightform, thanks to projection mapping, enables the projected patterns to superimpose exactly over real-world objects, Figure 1.13.

Also, the taxonomy space can be used to differentiate between the available AR displays based on their features and capabilities. In the plot of Figure 1.14 different technologies are compared: Smartphone, Lightform, and HoloLens.

Figure 1.13: Lightform. © Volkswagen via project MARTA

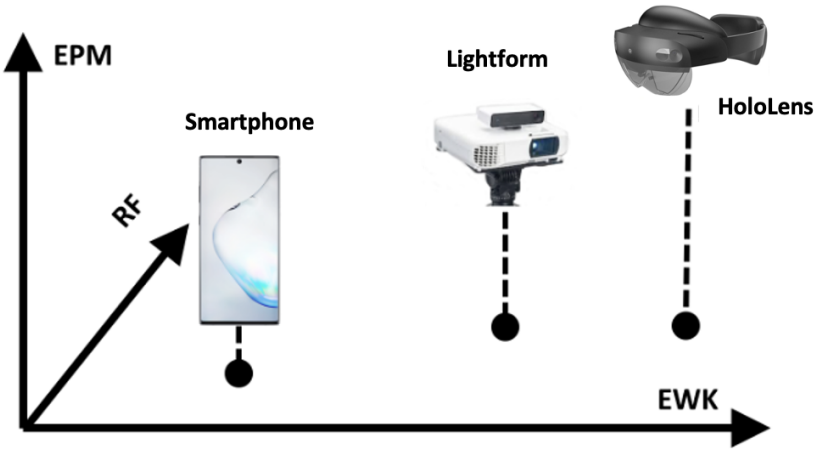


Figure 1.14: AR devices in the taxonomy space.

1.2.3 MR framework design

When designers need to develop a framework in which innovative technologies such as those based on MR visualization are present, regardless of the specific application, they need to answer three questions:

1. What kind of data would be helpful for the user?
2. How should these data be visualized?
3. Which is the best medium for visualization?

About the first question, designers should aim to gather the information that enables them to understand users' needs, preferences, behaviours, and contexts. To design user-centered solutions, they need to manage different data, such as:

- Final purpose of the framework: Understanding the goals and challenges can guide designers in developing solutions that align with users' desired outcomes and effectively overcome their pain points.

- User's biometric and physiological data: Wearable devices or sensors can collect biometric data such as breathing rate, heart rate, or stress levels. These data can provide insights into users' physical and emotional well-being, enabling designers to create solutions that promote health and wellness.

- User demographic data: Understanding users' demographics, such as age, gender, location, and level of instruction, can provide insights into their backgrounds, preferences, and motivations.

- User behaviour data: Collecting data on the existing methods by which users operate can offer valuable insights. It can include tracking user actions, patterns used to perform tasks, time spent on specific tasks, and frequency of use.

- Contextual data: Understanding the environment and context in which users operate is crucial. It can involve gathering data on location, time of day, environmental conditions, or other contextual factors that impact user experiences.

While emphasizing the value of data collection, it is equally crucial that designers prioritize user privacy and data security. Transparency and consent are essential when gathering personal or sensitive information, and data should be anonymized and aggregated whenever possible.

Ultimately, the data that would be helpful for users will depend on the specific context and design objectives. By combining multiple data sources, designers can comprehensively understand users and create solutions that meet their needs.

When choosing the data to show, the designer must focus on including only the essential information needed to minimize the cognitive load on the end user. Neglecting this consideration can lead to a scenario reminiscent of London designer Keiichi Matsuda's depiction in his 2016 hyper-reality film [70], where a nightmarish sci-fi future unfolds, inundating every surface, appliance, and peripheral vision with an overwhelming amount of data, Figure 1.15. This hyper-visualization results in such a high cognitive load that the designed interface becomes unusable.

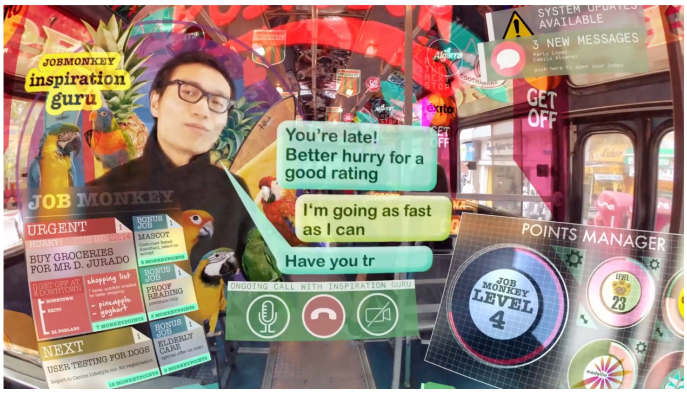


Figure 1.15: An example of hyper-visualization from a frame of Keiichi Matsuda's film about our future life saturated with inescapable streams of information, advertising, and data [70].

The second question about data visualization involves a combination of objective principles and subjective design choices. While there are established best practices and guidelines for data visualization, individuals' interpretation and design decisions can introduce subjective elements. When information must be added, whether in a real environment or a reconstructed virtual environment, it is essential to keep in mind the principles of visual perception and clarity when visualizing data:

- Data accuracy: Visualizations should accurately represent the underlying data without distorting or misrepresenting information.
- Clarity and simplicity: Visualizations should be clear, easy to understand, and free from unnecessary clutter.
- Location: All data should be placed in a way that simplifies understanding without confusing the user. For example, all information about a particular person or machinery can be displayed on or near them to make it easier to understand who they relate to.
- Consistency: Visual elements such as colour, scale, and labelling should be used consistently to avoid confusion.
- Contextual relevance: The visualizations should provide appropriate context to help users understand the data's meaning and significance.
- Accessibility: Consider accessibility guidelines to ensure that visualizations are usable by individuals with different abilities.

Additionally, subjective design choices can be adopted:

- Visual style: The choice of visual styles, such as colour palettes, typography, and layout, can introduce subjective elements and reflect the designer's aesthetic preferences.
- Emphasis and hierarchy: Designers emphasize specific data points or patterns over others based on their understanding of the insights and story they want to convey.
- Interaction design: The selection and implementation of interactive features within visualizations can be subjective, influenced by the intended user experience and the designer's creative approach.
- Interpretation and storytelling: Designers interpret the data and decide how to tell the story best or highlight critical insights. It can introduce subjectivity based on their understanding, expertise, and intended message.

While subjectivity plays a role in data visualization, designers must validate their choices through user feedback, usability testing, and iterative design processes. It helps ensure that the visualizations effectively communicate the intended information and are understood by the target audience. Collaboration and multidisciplinary approaches can also help balance subjective design choices with objective principles and domain expertise.

The last design question concerns the best medium for visualization. It depends on several factors, including the immersiveness required, the nature of the data, and the users' experience with the selected technology. It often results in a trade-off among these factors, and one crucial aspect to consider is accessibility to ensure that visualizations are usable by individuals with different abilities. For example, AR is suitable for situations where users must simultaneously

interact with physical and virtual elements. A good level of immersiveness can be provided by smart glasses such as Microsoft HoloLens, with the advantage of leaving hands free. They also allow the manipulation of virtual objects with gestures or voice commands, although more complex than traditional interaction methods on tablets or smartphones. While gestures and voice commands offer more immersive and interactive experiences, they often require a learning curve for users to become comfortable with the new interaction paradigms. Accessibility problems then increase for users with physical disabilities or those with speech impairments.

For this reason, one solution adopted in some AR applications developed and analyzed in this thesis is to use HoloLens only for visualization and to use traditional interaction methods, such as tablets or smartphones, to interact with the data, e.g., to decide which ones to display and which ones not, Figure 1.16.

Figure 1.16: Example of combined use of HoloLens and a smartphone to simplify interaction with virtual cues.

1.3 PAL in a Shared MR

One of the many advantages offered by MR technologies is their potential use as a medium for communication [97]. This enables new possibilities, such as multiple users' simultaneous experience of an augmented environment. A categorization of Computer-Supported Cooperative Work (CSCW) allows these technologies and, more generally, any form of computer-based medium to be classified in a temporal and space dimension [95]. Regarding the temporal dimension, collaboration among multiple users can occur synchronously (simultaneously) or asynchronously (at different times and thus independently). Regarding the spatial dimension, users can be co-located (in the same space) or remote (in different locations). When Virtual and Augmented Reality technologies converge to foster interactive and immersive collaborative experiences among multiple agents [12] it is called shared MR (SMR). Examples are also found in driving cars [86], in industrial settings [55] and, even in human-robot interaction while sharing the same virtual-real (mirrored) environment [42, 90]. Within the context of the PAL, SMR presents significant possibilities for enhancing how humans perceive, process, and respond to their environment by improving communication, collaboration, and information sharing. This extends PAL to a new

level of perception. In this dissertation, I focused on co-located and synchronous SMR applications, Figure 1.17.

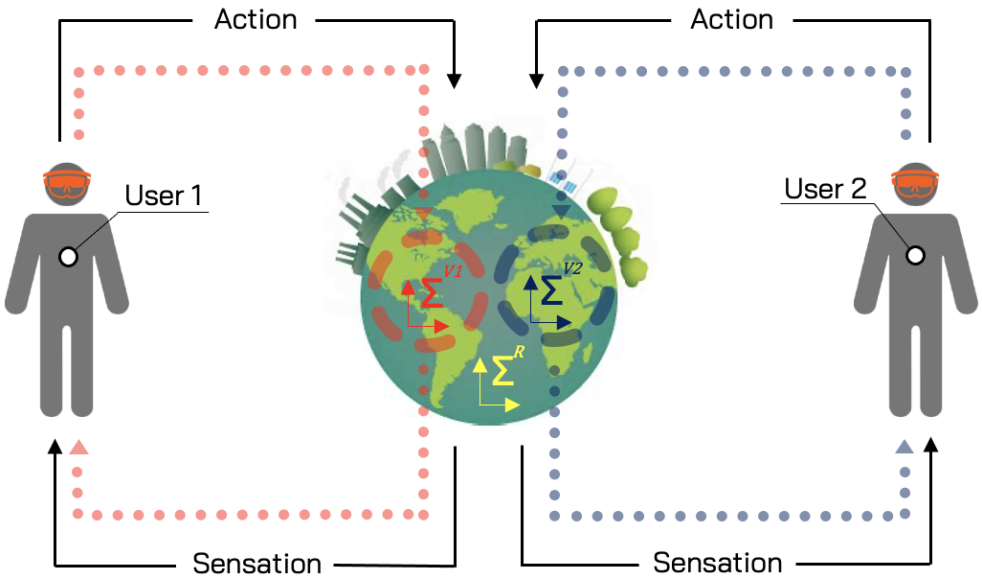


Figure 1.17: The second level of the perception-action loop: collaboration between two users in a shared virtual environment with the virtual environments V_1 and V_2 aligned in space and time with each other and the real world R.

Depending on the context, the tasks and roles of the two users may be the same or completely different. The alignment in space and time between the two shared virtual environments must be the same for both in a collaborative environment. Indeed, SMR has not yet been fully exploited because of the extrinsic calibration between the two viewpoints, which requires systems that can continuously track their positions and orientations with respect to the environment in real-time. HoloLens, through its depth cameras and inertial measurement units (IMUs), maps the physical environment allowing a spatial understanding of the surroundings. Using these sensors and a technology called "spatial anchors" [77], the reference systems of multiple HoloLens can be aligned in a shared environment. It ensures that virtual objects appear in the correct positions and orientations for all users involved. Moreover, in an AV context, collaborators in different locations can experience the same environment simultaneously, fostering a sense of presence and shared understanding.

Shared Augmented Reality (SAR) and Shared Augmented Virtuality (SAV) enable more prosperous collaboration and communication through visual annotations, gestures, and shared content. While SMR offers advantages, it also introduces cognitive load challenges. Users must simultaneously process physical and virtual information, potentially leading to information overload or distraction. Design considerations are crucial to minimizing cognitive load, such as optimizing the presentation of information and allowing users to adjust the level of augmented content.

SMR for collaborative purposes can be extended to another level when one of the two users becomes the supervisor, Figure 1.18.

The supervisor's perception increases by involving the second user in his virtual world. The supervisor perceives the virtual world as at the collaborative level but with additional information

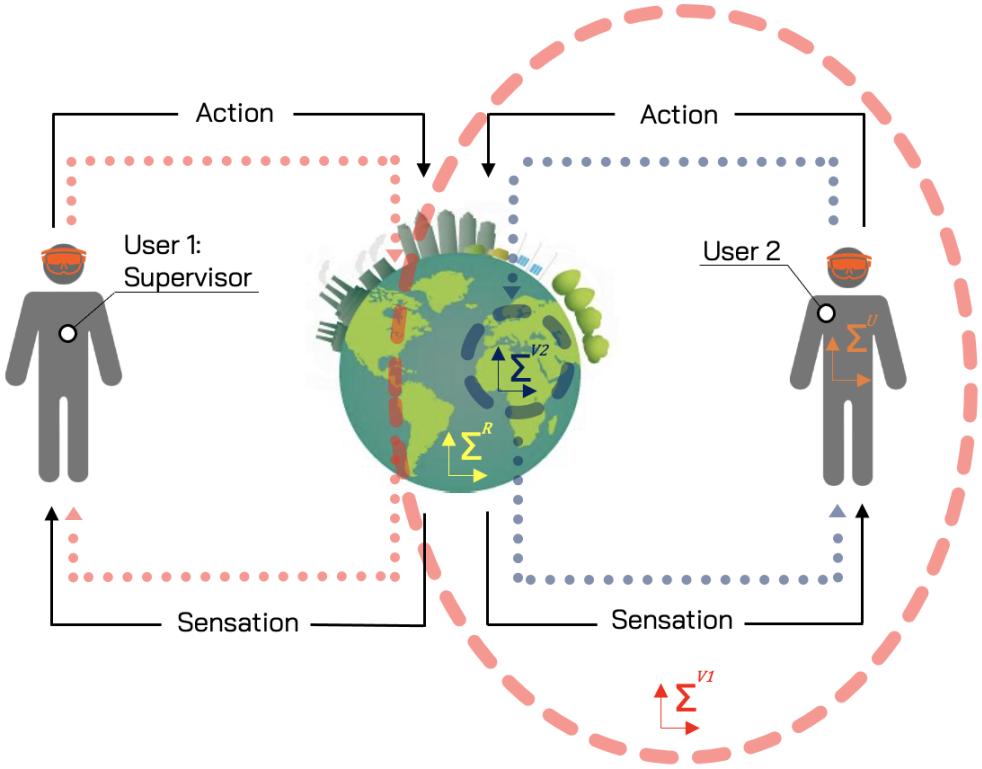


Figure 1.18: The third level of the perception-action loop: supervision of one of the two users. Supervisor's perception is augmented by the second user U included in his virtual environment V_1.

from the second user. The measurement systems of the second user involved different stages of PAL. For example, physiological parameters are related to his perception stage, while the motion capture system or his interaction with the environment is related to the action. Moreover, in AR, all this additional information can be displayed above the second user, expressing his relative reference system with respect to the environment. On the other hand, the second user continues to perceive only his virtual world. The concept of SMR with a supervisory role introduces a dynamic and interactive way for experts to assess, assist and guide others in real-world scenarios. Depending on the context, the supervisor may be a therapist assessing a patient, a teacher evaluating students, and a supervisor monitoring an operator in an industrial environment.

Finally, by embracing human-centered design principles, designers can ensure that SMR applications are intuitive, inclusive, and empathetic, paving the way for a future in which virtual collaboration and supervision become seamless, engaging, and deeply human-centered.

This dissertation explores innovative frameworks in various domains, each involving different levels of PAL. Each of the next chapters will discuss which levels of PAL are involved and enhanced. Not every designed application encompasses all levels of PAL for multiple users in a SMR framework. Nevertheless, all the algorithms and methods developed can always be included in a more general SMR framework with multiple users.

Chapter 2

AR in healthcare

AR has gained popularity in various industries, including gaming [79] and entertainment [44], but it has also made significant advancements in healthcare.

AR offers numerous possibilities for improving patient care, medical training, and surgical procedures in healthcare. By providing real-time, context-specific information, AR enables healthcare professionals to make more accurate diagnoses, perform complex surgeries with precision, and enhance patient outcomes. Here are a few areas where AR is being used in healthcare:

- Medical education and training: AR provides a unique platform for medical students and professionals to learn and practice complex procedures in a safe and controlled environment. It allows them to visualize and interact with anatomical structures, medical devices, and simulations, enhancing their understanding and skill development [124].
- Surgical planning and navigation: AR can assist surgeons in preoperative planning by overlaying patient-specific medical imaging data, such as Computed Tomography (CT) scans or Magnetic Resonance Imaging (MRI), onto the surgical site [6]. It allows surgeons to visualize the internal anatomy in real-time during the procedure, improving accuracy and reducing risks.
- Surgical guidance: During surgeries, AR can provide real-time guidance to surgeons by overlaying relevant information, such as critical structures, blood vessels, or tumor margins, directly onto the patient's body. It helps surgeons navigate complex anatomical areas and perform procedures with increased precision [18].
- Remote consultations and telemedicine: AR can facilitate remote consultations by enabling healthcare professionals to project their expertise onto the patient's location. By using AR glasses or mobile devices, doctors can visualize and guide patients through examinations, diagnostics, and treatments, regardless of their physical location [112].
- Patient education and empowerment: AR applications can help patients better understand their medical conditions, treatment options, and post-operative care. By visualizing the effects of diseases or explaining complex medical concepts, AR empowers patients to participate in their healthcare decisions actively [3].
- Mental health and well-being: AR can create immersive and interactive experiences promoting mental health and well-being. It can be employed in therapies for anxiety disorders, phobias [47], and Post-Traumatic Stress Disorder (PTSD) by providing controlled exposure to triggering situations.

- Rehabilitation and occupational therapy: AR technology can enhance rehabilitation programs by providing patients with interactive exercises and real-time feedback. It can help patients regain motor skills, improve coordination, and track progress [16, 65, 24, 102, 31].

2.1 Co-design

Co-designing AR technology between doctors and engineers can lead to the development of innovative and practical solutions that cater specifically to the needs of healthcare professionals. Collaborating on the design process ensures that the AR technology addresses the challenges faced by doctors and aligns with their clinical requirements [15]. An example of this co-design process is shown in Figure 2.1 between doctors and therapists at Villa Rosa Rehabilitation Hospital in Pergine, TN, Italy, and engineers from the Italian Department of Industrial Engineering at the University of Trento and the Japanese NARA Institute of Science and Technology.

Figure 2.1: Co-design between doctors and engineers in Villa Rosa rehabilitation hospital of Pergine (TN), Italy.

Here is a general framework for co-design applied to the specific case of AR solutions between doctors and engineers:

1. Identify needs and challenges: Doctors and engineers should come together to identify the specific needs, challenges, and opportunities where AR can make a meaningful impact in healthcare. It can involve discussing current pain points, workflow inefficiencies, and areas that could benefit from AR technology.

2. Define Objectives: Establish clear objectives for the AR technology based on the identified needs. Determine the specific goals it should achieve, such as improving surgical precision, enhancing diagnostic accuracy, or streamlining training processes. These objectives will guide the design process.

3. Gather user insights: Involve doctors and other healthcare professionals in the design process as active participants. Conduct interviews, observations, and surveys to gain deep insights into their workflows, preferences, and requirements. Understand how they interact with technology, their pain points, and their vision for augmented reality in their practice.

4. Prototype and iteration: Engineers can develop initial prototypes of the AR technology based on the insights gathered. These prototypes should be shared and tested with doctors for feedback and evaluation. This iterative process allows for continuous refinement and ensures that the technology aligns with the practical needs of doctors.

5. Usability testing: Conduct usability testing sessions with doctors to evaluate the effectiveness and usability of the AR technology. It involves observing how doctors interact with the prototypes in realistic scenarios and gathering feedback on user interface design, comfort, accuracy, and overall user experience.

6. Iterative refinement: Based on the feedback received during usability testing, engineers should refine and iterate the AR technology to address any identified issues or concerns. This process may involve modifying user interfaces, optimizing performance, enhancing ergonomics, or integrating new features based on the specific requirements of doctors.

7. Validation and deployment: Once the AR technology has undergone sufficient iterations and refinement, validate its effectiveness and impact through pilot studies or clinical trials. Collect data on its performance, user satisfaction, and patient outcomes. This evidence will help fine-tune the technology further and gain acceptance from doctors and regulatory bodies.

8. Continuous collaboration: Collaboration between doctors and engineers should continue after the initial design and deployment. Establishing a continuous feedback loop is crucial, allowing doctors to provide ongoing input and suggestions for improvements. It ensures that AR technology remains aligned with the evolving needs of healthcare professionals.

The resulting AR technology will be more effective, practical, and user-friendly by fostering collaboration and co-design between doctors and engineers. This interdisciplinary approach leverages the expertise of both parties, leading to the development of innovative solutions that truly enhance healthcare delivery and patient care.

2.2 Occupational Therapy

Occupational Therapy (OT) is a healthcare profession focused on helping individuals develop, recover, or maintain the skills needed to participate in Activities of Daily Living (ADLs) [51]. These activities can encompass self-care tasks (such as bathing, dressing, and eating), productivity tasks (work, school, or volunteering), and leisure occupations (sports, hobbies, and social interactions).

Occupational therapists work with people of all ages, from infants to the elderly, and across various settings, such as hospitals, schools, rehabilitation centres, and community-based programs. They collaborate with individuals with motor/sensory impairments, cognitive/perceptual insufficiencies, behavioral shortfalls, or visual discrepancies that may impede their ability to engage in activities essential to their well-being and independence. They can be the result of brain injuries such as a stroke, traumatic brain injury, or brain tumor.

The primary goal of OT is to improve a patient's overall quality of life and functional independence. Occupational therapists achieve this through a patient-centred approach that considers each individual's unique needs, goals, and abilities. They conduct comprehensive assessments to identify strengths and challenges, design personalized treatment plans, and implement evidence-based interventions to address specific areas of difficulty.

The evaluation of the patient's abilities is based on a non-standardized method that measures the patient's performance using these markers: safety, efficiency, effort, and independence [27]. In the current workflow, occupational therapists give standardized examinations using manuals such as the Assessment of Motor and Process Skills (AMPS) [19], which makes them more trustworthy and consistent, but still has some remaining flaws. Moreover, in order to gain the ability to conduct the proper AMPS evaluation, continued training is necessary. Consequently, the clinician's expertise impacts an assessment utilizing these modalities and is therefore susceptible to mistakes and misinterpretations for less experienced therapists. In addition, today, the number of health conditions associated with severe disability rates has reached 183 million [80]. The resources needed for addressing rehabilitation needs out-measures accessibility, resulting in inadequacies in these services. To cope with these demands, the education and training of therapists who have just finished schooling may find their knowledge and expertise lacking in the actual field [8]. It is mainly due to the differences between the scope of the theoretical knowledge in the literature about rehabilitation concepts and their application in clinical practice. Developing the clinical eye would take years of practice, so novice therapists would need help making complicated clinical decisions and evaluations. As a solution to this problem, innovative visualization technologies such as AR and enhanced measuring techniques can be support tools therapists use for motor rehabilitation. Introducing these technologies may be more useful for therapists with less experience. Still, it could support veteran therapists by speeding up the evaluation process or showing more detailed patient data. The possibility to have more information in AR contextualized close to the patient simplifies their assessment without losing the exteroception of the scene. In this way, occupational therapists can define more reliable assessment scales based on objective parameters, increasing the effectiveness of clinical observation for more effective rehabilitation programs.

AR assistive technology in clinical settings is widely discussed in the literature for ADLs support [96, 115, 76]. In most situations, the decision to employ AR technologies for ADLs training rather than other technologies in the MR spectrum is based on the notion that generally, subjects perform better in AR than in VR in terms of self-to-environment-related movements, as hand-eye coordination in VR involves a much higher extraneous cognitive load [53]. Additionally, the patient can manipulate physical objects while seeing virtual information in AR to ensure the perception of the surrounding physical surroundings and the weight of those objects. Moreover, these technologies not only increase the clinical eye of occupational therapists and thus their final assessment of patients, but also the engagement of patients in daily life through gamification of some of their daily activities in AR through serious games.

2.3 Serious games

Serious games [2], also known as therapeutic or health-related games, have gained popularity in OT as a valuable tool to engage patients and enhance their therapeutic outcomes. These games are designed with specific therapeutic goals. They are intended to be fun and interactive while addressing various physical, cognitive, emotional, and social challenges faced by individuals receiving OT. Therapy sessions are more enjoyable and less intimidating for patients, particularly children and young adults. By incorporating game elements like rewards, points, and competition, serious games can boost motivation and engagement during therapy sessions, encouraging individuals to participate actively in their rehabilitation process. They can simulate real-life activities, such as cooking, driving, or shopping, to help individuals practice functional tasks in a controlled and supportive environment. In addition, they can also provide therapists with valuable data on patients' progress and performance, enabling personalized treatment plans.

The immersiveness of serious games can be enhanced through AR. Therapists can leverage this technology to design engaging and effective interventions, actively motivating patients to participate in rehabilitation and skill development. Moreover, their perception can be enhanced at different levels of PAL, from simple collaboration to supervision in a SAR framework, Figure 2.2.

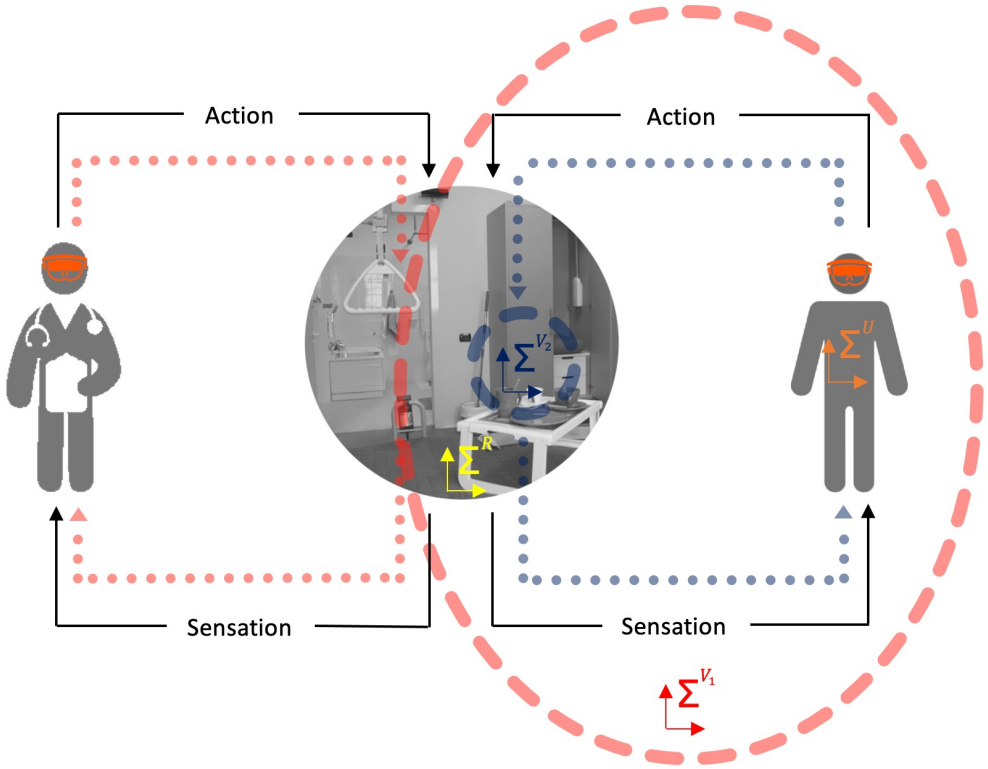


Figure 2.2: The third level of the perception-action loop between therapist, i.e. supervisor, and patient in a shared augmented reality framework. The external environment R is represented in this context by, for example, a domotic apartment.

Within the Measurement, Instrumentation and Robotics Laboratory (MiroLab) of the Department of Industrial Engineering at the University of Trento, several AR-based serious games were developed, demonstrating the application of this technology in OT. They were co-designed with doctors and therapists at the Villa Rosa Rehabilitation Hospital in Pergine, TN (Italy). The purpose of some of them is also to show doctors and therapists the potential that an AR solution may have, and then jointly evaluate its applicability to a specific pathology. Not all of the demos described in the next sections have been tested and validated with real patients because they are still undergoing iterative refinement between engineers and doctors. In Section 2.4, however, the SAR framework for the specific ADL scenario of setting up the table was validated with real patients.

2.3.1 Flower watering game

A first AR-based demo, designed in collaboration with the Interactive Media Design Lab of the Nara Institute of Science and Technology in Japan, aims for immersive training in horticultural therapy, Figure 2.3.

Figure 2.3: AR demo presented to the 22nd IEEE International Symposium on Mixed and Augmented Reality (ISMAR) for watering flowers with virtual objects and two actors: a therapist and a patient.

The proposed AR-based occupational therapy demo has several advantages for patients and therapists. For the patient, it is an immersive and engaging training experience without losing contact and perception of the real environment, thanks to the combination of real and virtual elements. He can train physical (muscle coordination, range of motion) and cognitive functions (short-term memory, planning) [100, 59, 45]. Furthermore, it provides real-time performance feedback with a demo whose difficulty levels can be tailored to their individual skills. The therapist can control almost all parameters and objectively evaluate the patient's performance. In addition, the framework is open to integrating additional sensors on the patient, such as pressure insoles or wearable heart, electromyography (EMG), or respiratory sensors, to enhance the therapist's clinical eye and thus extend the level of PAL to a supervised SAR.

The end user may be a patient with motor disturbances due to upper spinal cord injury (paraparesis, tetraparesis), cerebellar lesion (ataxia), brain stem/basal nuclei lesion (Parkinson's disease and parkinsonisms), brain lesions (stroke with hemiparesis and possible cognitive disorders such as apraxia, inattention, head trauma with executive function problems, tremors, sensory disturbances (of vision such as diplopia, of proprioception) and multiple sclerosis.

The developed framework involves both the therapist and the patient wearing a HoloLens. The demo starts with the therapist configuring parameters via a custom smartphone application. The application allows the therapist to set parameters such as patient name, watering can weight, flower growth time, watering frequency and adjust the complexity of the task. These parameters can be tailored to suit the patient's specific needs and abilities. The smartphone is then placed on the watering can to measure its tilt angle, fusing its gyroscope output with the attitude estimated via a Vuforia marker. Combining the two measurement systems improves accuracy and allows an estimation when the HoloLens camera is obstructed or the watering can is outside its field of view (FOV).

Next, the therapist places the virtual pots in the room with HoloLens, Figure 2.4. This procedure considers objects and obstacles by estimating the environment mesh using the Mixed Reality Toolkit.

Figure 2.4: Game setting according to the therapist's FOV.

Now it is the patient's turn, who, through HoloLens, visualize the pots to water. The patient is instructed to water the virtual pots using a real watering can (without water) to grow virtual flowers. When the pot receives water in its (virtual) area, plants and flowers grow dynamically accordingly to the amount of water received, Figure 2.5. The watering can has the option to add (real) weight to adjust the training intensity according to the patient's physical progress.

Figure 2.5: Patient's task from his FOV.

A countdown timer is displayed on the AR glasses near each virtual pot to guide the patient in the task. In addition, 3D sounds help the patient to identify the next pot, thus improving spatial awareness, focus of attention and engagement.

All data, including watering accuracy, watering time, and task completion speed, are saved. The therapist can decide whether to share the final score with the patient in AR.

During this iterative refinement phase of the co-design process, we received a lot of positive feedback from doctors and therapists. They appreciated the simplicity of setting up the demo, the heightened engagement of potential patients due to the high-quality graphical animation that closely approximates a real-world scenario, and the system's capability to collect very useful data for patient assessment while offering effective patient training.

2.3.2 Balance games

Other demos were presented to therapists as initial prototypes, after identifying with them the need for a tool to train and assess patients' balance and body movements.

The first, Figure 2.6, is a serious game developed for people with torso problems; the requirement was to create a game where the user must move the head to avoid an object. During the game, the subject's feet can stand on top of a baropodometric platform to see how the weight distribution changes during the game. This game can work either standing or sitting.

Figure 2.6: The AR game demo for people with torso problems.

The second demo in Figure 2.7, is a serious game developed to test and train users' balance through an AR balancing game. Users have to hit green virtual capsules with a virtual ball while avoiding red ones by tilting the plane in front of them.

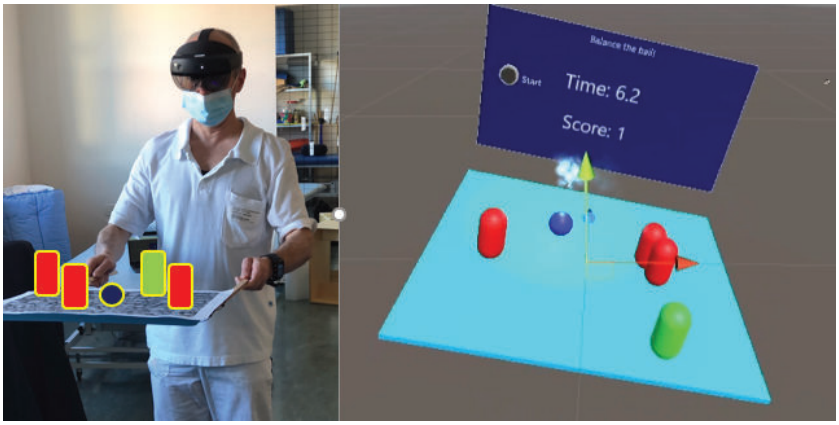


Figure 2.7: The balance-the-ball AR game demo that was presented to the therapists.

Therapists agree that these demos can help people improve their motor skills and coordination. The positive feedback collected again highlights the high level of engagement through AR technologies and the high level of safety in interacting with virtual objects, allowing their use to be extended to a wider range of patients. These demos were also designed to stimulate the imagination of therapists so that they could draw up novel programs to give to their patients.

However, it is essential to use serious games to complement traditional therapeutic methods rather than replace them. Each patient's needs and goals should guide the selection and implementation of serious games, ensuring that they align with the overall treatment plan and contribute effectively to the individual's functional outcomes and well-being.

In addition, when using AR in occupational therapy, it is essential to consider the individual's comfort with technology, the specific therapeutic goals, and potential safety concerns. Therapists should provide proper guidance and supervision, ensuring the technology complements the overall treatment plan and contributes effectively to the client's functional progress and well-being.

As technology continues to evolve, the possibilities for utilizing AR in occupational therapy will likely expand, offering even more innovative and effective ways to support patients in achieving their therapeutic goals.

2.4 Specific ADL in a SAR kitchen environment

Another original framework was designed [24] involving all levels of the PAL in a SAR environment for the specific ADL scenario of setting up the table. The SAR environment is augmenting with the proper elements from the two different perspectives of the therapist and the patient. The fundamental novelty of the proposed framework lies in the enhancement and support of the clinical eye [23] in a SAR environment by increasing empathy between actors [89]. The proposed prototype increases the therapist's involvement and perception of the patient with the ability to access their multidimensional data in AR. Furthermore, it helps improve the patient's engagement by allowing interaction with virtual augmented information and real tools/utensils throughout the ADL exercise. The AR system incorporates a robust, reliable, and accurate computer vision-based technique to assure the high metrological quality of the evaluation. This system does not replace the traditional workflow of the therapist-patient interaction during the ADL but instead promotes and deepens this interaction [37]. This interaction is bridged by having the patient see virtual guides that support their understanding of the ADL task that the therapist describes. Moreover, on the therapist's side, they can see the invisible current conditions of the patient (i.e., body and feet posture, heart rate), and they can better understand the situation and make correct decisions and guidance about the ADL execution.

The prototype was developed in the MiroLab of the University of Trento and set up inside the home automation apartment AUSILIA(Assisted Unit for the Simulation of Independent Activities) [35] at the rehabilitation hospital Villa Rosa in Pergine Valsugana (Italy). Figure 2.8 shows the framework tools used during the ADL assessment.

Visualization devices

Both therapists and patients can see AR cues on Microsoft HoloLens 2 head-mounted displays. In addition, the therapist can manage and interact with information using a handheld device, such as a smartphone.

Distributed measurement system

The measurement system includes the following components:

- Two Time-of-Flight (ToF) depth cameras, such as the Microsoft Kinect v2, with the first one in front of the patient, used to determine where the position of the body joints are in 3D space [81]; the second one, above the table, used to capture an RGB image (1920 $\times$ 1080)

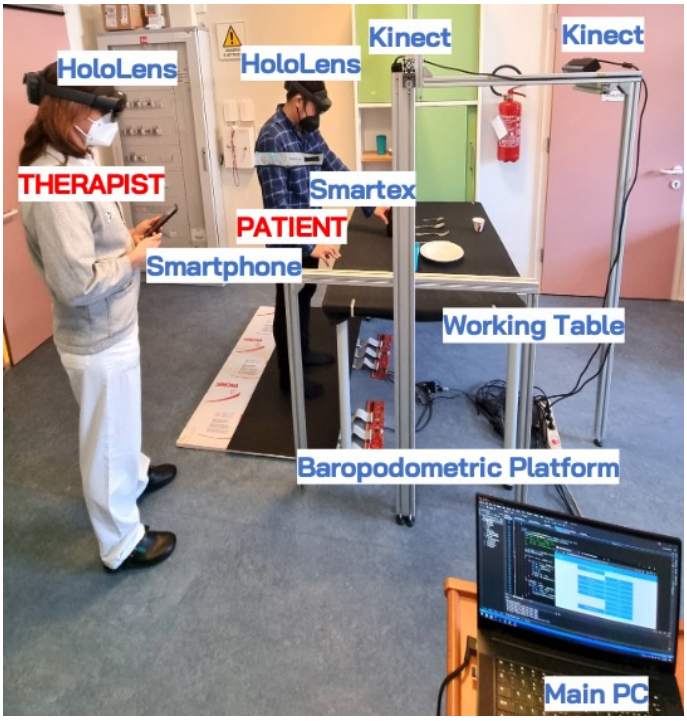


Figure 2.8: Framework setup in the AUSILIA apartment.

for the computer vision based-algorithm and to measure the height from the table and check its orientation during the initial setup phase.

- A wearable band system developed by the company Smartex s.r.l of Navacchio (PI), Italy. It continuously monitors several physiological parameters. In particular, the system can simultaneously acquire the patient's electrocardiographic (ECG) and respiratory signals.
- The baropodometric platform used for non-invasive static and dynamic pressure measurement and body stability analysis is a customized model of the FreeMed family manufactured by the Italian company Sensor Medica of Guidonia Montecelio (RM). The platform, which measures 56 × 120 cm, consists of two units, the sum of which results in 6000 24k gold-coated resistive sensors with frequency acquisition up to 400 Hz.
- The main PC, where all raw sensor data are processed, stored, and sent.

Software development and communication protocols

The control interface for handheld devices such as smartphones was developed with the Node-RED programming tool, Figure 2.9.

This development tool is useful for real-time data management and elaboration for IoT distributed systems. Its advantages include: open-source, visual programming ("flow-based programming"); fast development; lightweight; efficient MQTT (Message Queuing Telemetry Transport) client-server protocol.

All devices, including HoloLens, a smartphone, and the main computer, are connected over the same LAN. The MQTT protocol, based on TCP/IP, thanks to its reliability and lightness,

Figure 2.9: Examples of control interfaces for the therapist's handheld device.

allows the communication of data involving logic control (i.e., interface buttons, switches, and other controls). On the other hand, standard UDP (User Datagram Protocol) broadcasts data that concerns a large and continuous stream of information (i.e., platform data, Kinect data). The data transmission pipeline is shown in Figure 2.10.

The raw ECG and respiratory signal acquired by the Smartex band were processed and analyzed via Bluetooth Low Energy (BLE) in the main PC. HoloLens then received this data. In particular, for the analysis of the ECG and the respiratory signal, we took a 2 s time window to highlight any changes in physiological signals while the patient was performing short tasks. The Smartex band acquires the raw ECG signal with a frequency of 250 Hz. Data were successively filtered using a zero-phase passband Butterworth filter (cutoff frequencies, 0.1 Hz–20 Hz) and a modified version of the Pan–Tompkins algorithm was implemented to detect the R peaks [83]. The time differences between consecutive R peaks were calculated, obtaining the RR interval time series. For the patient's average heart rate, we considered the mean value of the punctual heart rate values within the 2 s time window. The breath rate was extracted from the raw respiratory signal (acquired at 50 Hz frequency) by removing the mean value and applying a zero-phase bandpass Butterworth filter (cutoff frequencies, 0.1 Hz–0.6 Hz). The peaks in the resulting signal were detected considering the following assumptions: a temporal distance greater than half of the average distance between all peaks and an amplitude greater than half of the average amplitude. As for the ECG signal, the considered breath rate was the average value within the 2 s time window.

Both Kinects are connected via USB to the main PC and are not used for synchronous acquisitions. The first is used during the table-setting task, and the second when the therapist presses the button to evaluate the error between virtual and real objects. We choose which sensor to use by enabling and disabling the USB port to which each Kinect is connected. For the first Kinect, we considered only the joints of the upper half of the body. The 3D coordinates of these joints are then converted from Kinect to a marker coordinate system using the transformation matrix calculated from the calibration. These data are then broadcasted via UDP to HoloLens at a rate of 30 Hz. The second Kinect acquires the RGB image to be provided as input to the computer vision-based algorithm, the output of which, in the form of deviations in the position

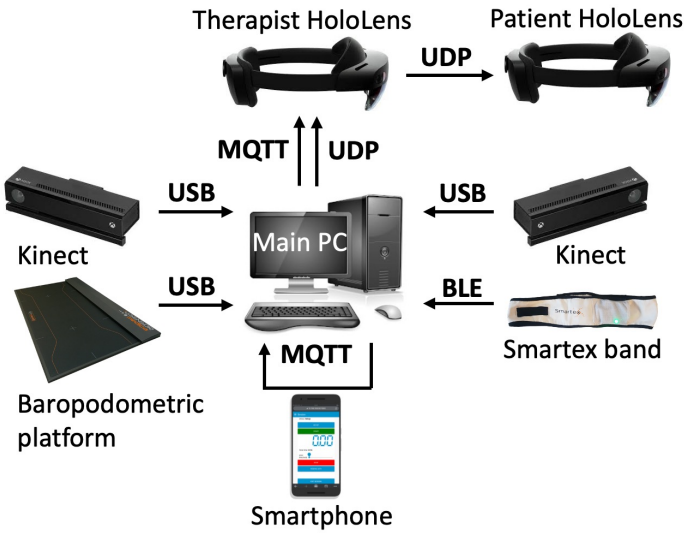


Figure 2.10: Data transmission pipeline.

and attitude of each object, is sent via UDP to HoloLens.

The FreeMed platform also connects to the main PC via USB. Data is collected using the C# program given by the FreeMed company, and broadcasted via UDP to HoloLens. The platform comprises 120 by 50 pressure sensors, with a total of 6000 small sensors. Each sensor returns a value between 0 and 255, with 0 being no pressure and 255 with max. This sensitivity is adjusted during the calibration phase.

Butterworth filters were applied to the Kinect and HoloLens data to reduce noise. A sixth-order Butterworth filter with 3 Hz cutoff frequency was selected for filtering both devices. Figure 2.11 summarizes the data processing flow chart of the main devices.

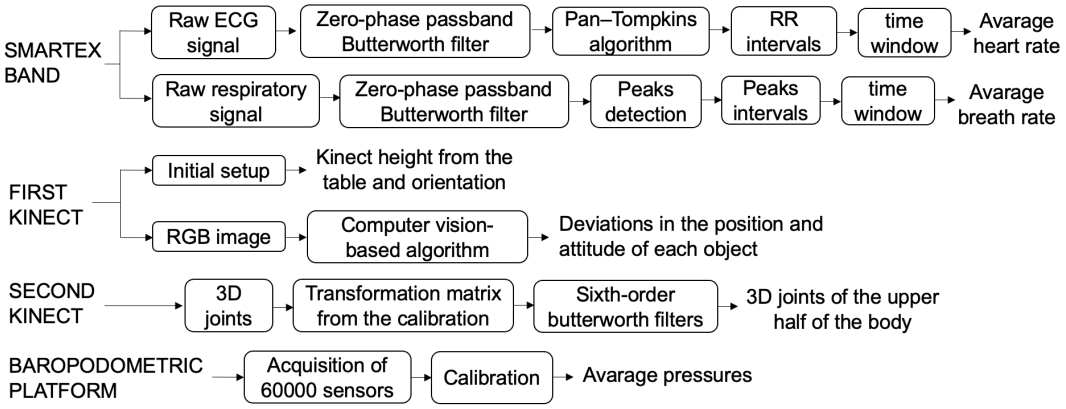


Figure 2.11: Data processing flow chart.

Extrinsic parameters calibration

For Kinect and the therapist's HoloLens to track the patient's kinematics in the same reference system, calibration is required (Figure 2.12). During the set-up phase, a marker with enough detectable feature points was used to derive a transformation matrix from Kinect camera coordinates to marker coordinates. The calibration process is repeated until an acceptable reprojection error is achieved. Vuforia SDK handles all the image target tracking for HoloLens. A spatial anchor is saved in the HoloLens using the same marker used for Kinect calibration. In this way, the Kinect and HoloLens can operate in the same reference system. Once calibrated, the marker can be removed at any time. Additional spatial anchors are saved in the therapist's HoloLens to define the reference systems of the working table and baropodometric platform. On the patient's HoloLens, however, only the spatial anchor related to the working table reference system is saved to operate in the same reference system as the therapist's HoloLens.

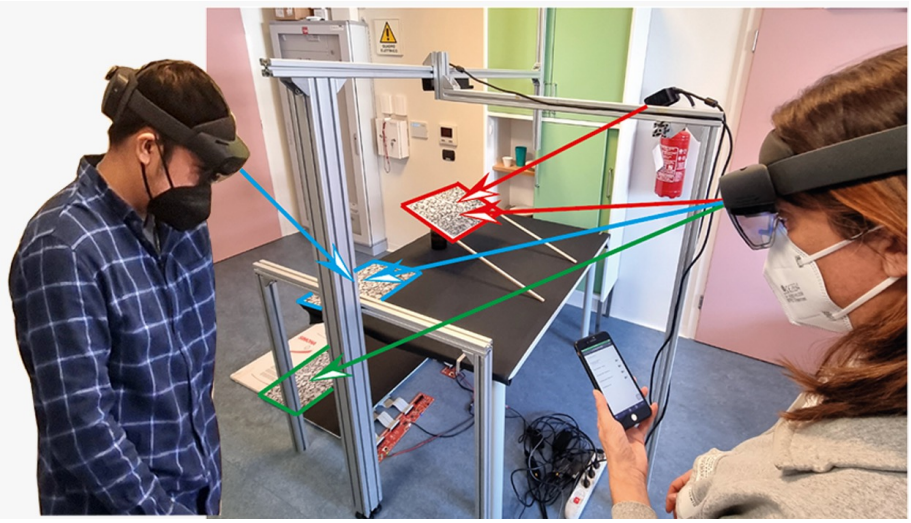
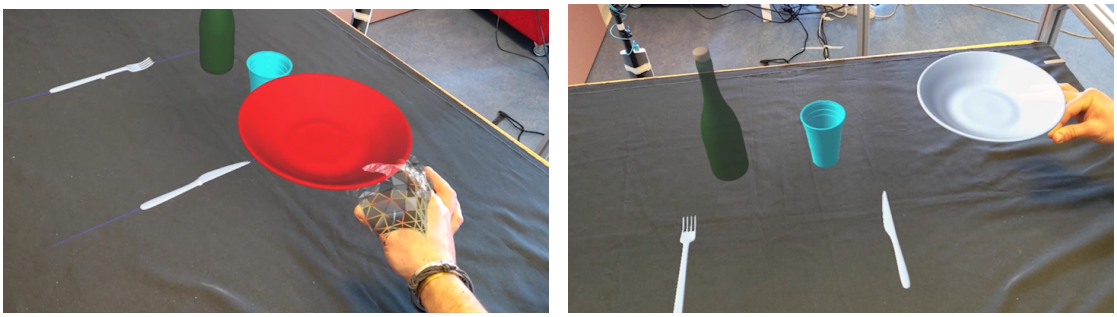


Figure 2.12: Spatial Anchors setting: the red image target is used by the therapist's HoloLens and the Kinect to operate in the same reference system; the blue target is used by the therapist's and the patient's Hololens to have the same reference system of the working plane; the green target is used only by the therapist's HoloLens to localize the baropodometric platform in space.

2.4.1 Evaluation process in SAR

The therapist assessing the patient during the instrumental ADL of setting the table is aided by a SAR scenario that can enhance his clinical assessment in an immersive and engaging way for the patient. The evaluation process involves the following steps:

1. Wearing a head-mounted Microsoft HoloLens 2, the therapist sets the table with virtual objects, Figure 2.13a. A handheld device's graphical interface allows the therapist to select the type and number of objects. Depending on the type of patient being assessed, therapist can adjust the complexity of the setup as needed. During this phase, the patient wearing another HoloLens 2 can view the virtual environment setup from his point of view, Figure 2.13b.



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

Figure 2.13: Example of a SAR environment from the FOV of the (**a**) therapist's HoloLens and (**b**) patient's HoloLens.

2. Once finished, the patient can view the virtual environment previously set up by the therapist and must try to match the virtual objects with the real ones.

3. Once the table setting is completed, the patient is asked to move his hands away from the table to avoid hiding real objects from the camera's view. Then, by pressing a button on the smartphone, the therapist estimates how far the real objects are from the virtual ones based on the position and angle errors that appear in AR next to each object with numbers following the therapist's gaze in Figure 2.14a. Numbers are displayed in different colors (green-yellow-red) according to the tolerance and, therefore, the threshold of error acceptability set by the therapist. If the algorithm does not find a match between a real object and a virtual object because, for example, the patient forgot to add the corresponding real object above the table, the associated virtual object is completely colored red. This indicates that the patient made an error with this virtual object; it is then up to the therapist to assess what kind of error because the algorithm could not return an output.

4. Another panel in AR summarizes the average angles and the average distances between the barycenters of the virtual and real models with the total task execution time, shown in Figure 2.14b.

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

Figure 2.14: (**a**) Example of errors visualization in AR via therapist's HoloLens 2 with (**b**) AR panel in which error averages and total time are summarized.

5. Therapists can decide with a smartphone whether to display additional information about the patient in AR during the exercise session, such as the reconstruction of the patient's kinematics and angles between the limbs, the load distribution of the legs, and his physiological parameters (Figure 2.15).

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

Figure 2.15: Example of information in AR from the therapist's point of view on the (**a**) patient's lower and (**b**) upper body.

6. At the end of each session, the therapist can decide to save all captured data to a text file.

2.4.2 Algorithm for object segmentation, localization & identification

An algorithm was developed in a MATLAB environment to identify and locate real objects of interest placed on a table by a user. Following the processing of an RGB image, this algorithm can detect and identify such objects, as summarized in Figure 2.16.

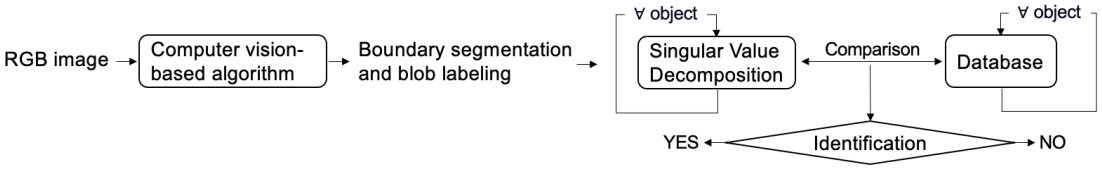


Figure 2.16: Flow chart of algorithm processing data.

It is not a real-time algorithm but is executed only when a snapshot image is taken as input at the time the therapist presses the button to evaluate the error between virtual and real objects. In addition, the captured image only covers the plane of the table, so objects that are outside the camera's FOV are not considered in the object recognition process. Items to identify include polished stainless steel cutlery. A sandblasting process made them opaque and unaffected by the direction of light, overcoming the problem of reflections on their surface that could affect the result of the algorithm.

The algorithm can be divided into the following steps:

Segmentation and localization

First, the RGB image captured by the Kinect fixed on top of the table was captured and processed in the following order:

1. Using a Kinect, grayscale images were acquired of the empty table and the same table covered with real objects.
2. Images were cropped to take into account only the table region of interest (ROI).
3. Each pixel was subtracted from the two previous images following background subtraction, and a threshold was selected to convert the result to a binary image.
4. The resulting mask was applied to the initial RGB image of the table set (Figure 2.17a), and a color-based threshold was applied to remove object shadows from the image (Figure 2.17b).
5. Next, flood-fill operations were performed on the hole pixels of the closed regions [101], as shown in Figure 2.17c.
6. A boundary label was applied to the filtered image [38].
7. Noise was removed by applying a threshold on the minimum number of pixels over the area of each labeled object.
8. The outer boundaries of each object were then traced [84], as shown in Figure 2.17d.
9. Objects were localized by taking the mean of their boundary coordinates and by rotating them using Singular Value Decomposition (SVD).
10. In the end, a mask with each object-centered and aligned was stored.

Identification

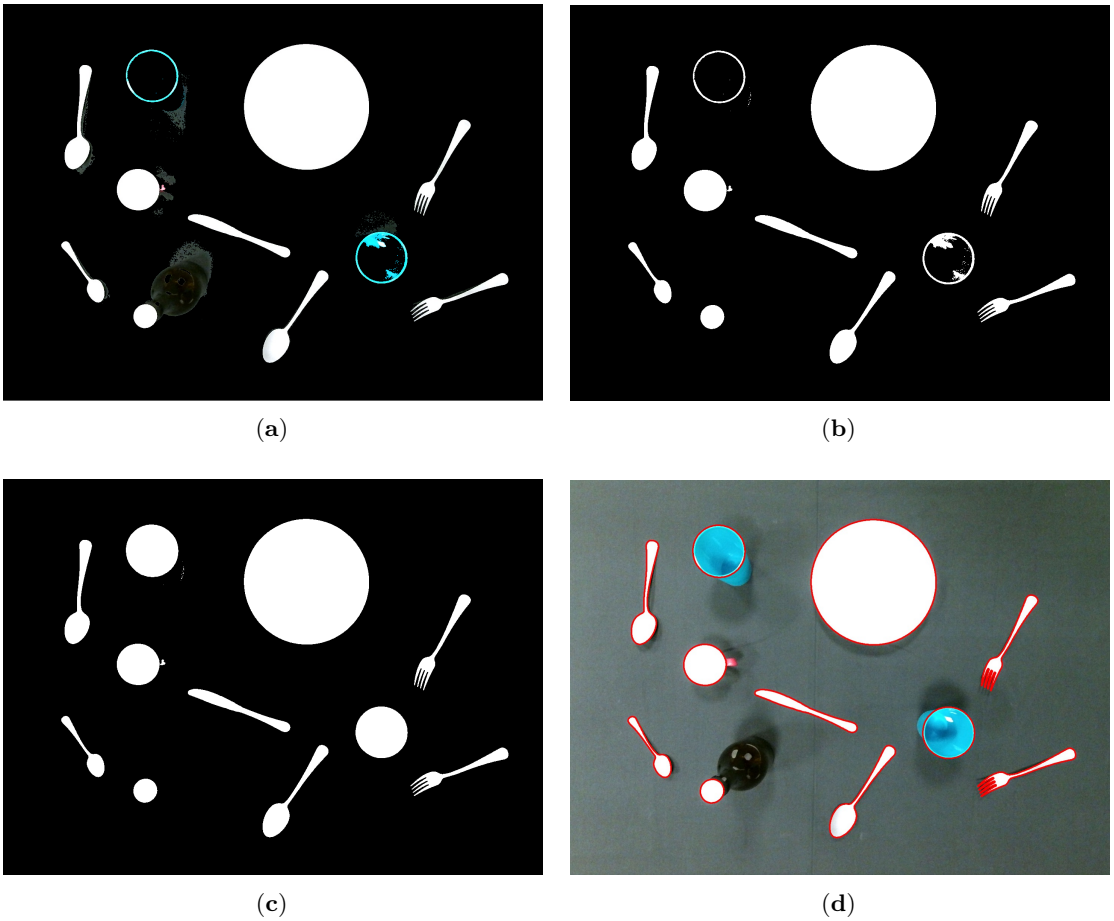
The previous image processing produces a binary image of each object segmented and aligned to the center of the initial image. Objects under consideration were compared to a previously created database using a cost function. The database was created using the same segmentation and realignment method as in the previous subsection, and the final labeling of the objects was carried out manually. Only one image for each object was required to initialise the database that will be referenced during matching (REF_{IM}).

Based on a set threshold, the input image (IN_{IM}) was compared to all objects in the database to identify the best match.

The first step is determining whether the areas between IN_{IM} and REF_{IM} are similar within 30%. If so, the cost function (CF) between them is calculated as follows:

$$CF = \frac{(1 - SC) + (1 - SA) + (1 - SSIM)}{3} \tag{2.1}$$

where SC is the score of similarity related to the object contours. In particular, the contour of IN_{IM} is smoothed with a 2D Gaussian smoothing kernel with a standard deviation whose value changes according to the object's size. Then, the resulting image is converted to a binary image and multiplied by the contour of REF_{IM} to check how many points of the two contours are in common. SA is the score of similarity related to the object areas. It consists of the product of the two binary images of IN_{IM} and REF_{IM} to check how many points of the two areas are in common. $SSIM$ calculates the score related to the structural similarity between the IN_{IM} and REF_{IM}. This score is a multiplicative combination of the three terms, namely the luminance, contrast, and structural term [113]. However, the black background is very predominant with respect to the size of the object when comparing IN_{IM} and REF_{IM} with $SSIM$. Therefore,



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

(**c**) (**d**)

Figure 2.17: (**a**) Resulting of the mask applied to the original RGB image; (**b**) Color-based threshold to remove object shadows; (**c**) Flood-fill; (**d**) Boundary segmentation and blob labeling.

both source images were cropped before comparison to make this score more sensitive to the objects in the images. The two new images have the same size between them, i.e., 50% more than the dimensions of the largest object in IN_{IM} and REF_{IM}, to be sure that the objects are still contained in the cropped images. All scores in Equation (2.1) are normalized. All terms are subtracted from the value 1 because we are looking for the minimum value of CF.

2.4.3 Metric calibration of the working table

After initial camera calibration, metric analyses were performed to assess the implemented algorithm's performance in identifying real objects and estimating their position and orientation.

Camera calibration

During camera calibration, the coordinates of each pixel on the CCD image sensor are compared with their real-world measurements. This is done by taking into account lens distortions, which are the most common monochromatic optical aberrations. At a fixed height of 80 cm, the Kinect

camera captures an image of a planar pattern perpendicular to the table and in its center. The planar pattern consists of 55 Aruco markers located at the vertices of a grid with known positions, Figure 2.18. The geometrical centers and identifiers of the Aruco markers [32] were saved and compared to their locations in the environment.

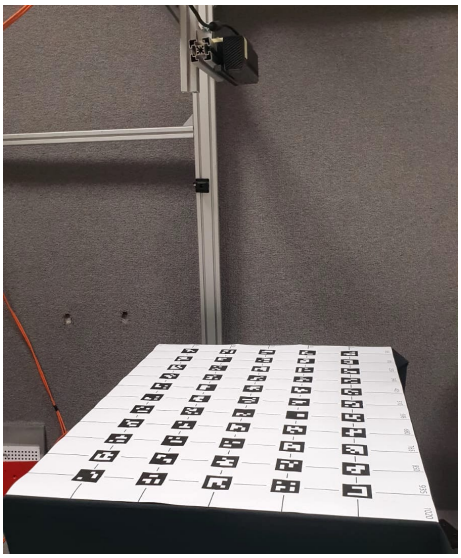


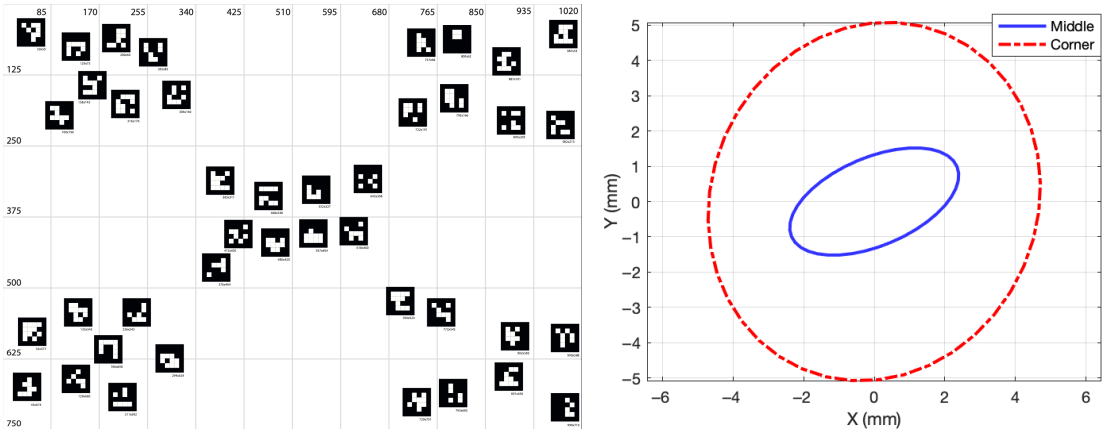
Figure 2.18: Aruco markers calibration plane.

An additional planar Aruco model (Figure 2.19a) was used to evaluate the calibration process and thus the accuracy of a random position on the table plane of dimensions 750 × 1020 mm. Once the set of random Aruco markers in the four corners was taken, the second time, the set of randomly placed Aruco markers in the center was taken, and the corresponding two-dimensional covariance matrices were computed. Covariance matrix results are shown in Figure 2.19b. As expected, the uncertainty ellipse around corners is larger due to the higher camera distortion.

Moreover, the height of the objects used is different. For example, a bottle is much higher than cutlery which is flat on the table. Nevertheless, it has not been necessary to calibrate the camera at different heights because knowing the exact heights of the objects and the camera, with respect to the plane using trigonometric operations, we have always referred to the plane of the table.

Accuracy in the image-pattern-recognition tool

The reference system of the HoloLens 2 worn by the therapist and the user were initially set up by watching a square appear on a predefined pattern using the Vuforia Engine image-pattern-recognition tool and saving its position and orientation over time. It is possible that the two reference systems are not aligned with each other because image marker detection and rendering stability can be affected by several factors. The size of the image marker and the resolution of the head-mounted display (HMD) camera do not affect the final accuracy because the HMD hardware and the image marker are the same for the therapist and the patient. What most affects the final accuracy is the distance and angle between the camera and the image marker. The authors in [87] provided $\leq 2°$ and ≤ 2 mm inclination angle and positional errors, respectively, in 70–75% of cases by using a holographic headset combined with an image-pattern-recognition tool. It could be a problem for the therapist's final assessment. On the other hand, except for



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

Figure 2.19: (**a**) Aruco markers plane for accuracy checking; (**b**) Ellipses of uncertainty in position (95% confidence level with $k = 2.4478$ [99]).

a small error of different visualizations of the virtual objects in the SAR environment, it does not suffer of differences in position and attitude of each object between the real one and its virtual model. Everything is evaluated on board the patient's HMD with its reference system. Therefore, for our metric purpose of therapist assessment of patient exercise, the accuracy of the image pattern recognition tool is irrelevant between the therapist's HMD and the user's HMD.

Algorithm accuracy for object segmentation, localization & identification

We performed rotation tests with a knife to evaluate the performance of the developed computer vision-based algorithm. In particular, Figure 2.20 shows a cropped RGB acquisition image of tests conducted using a manual rotation motion platform (Standa 126865) covered with black to facilitate the background subtraction and filtering process.

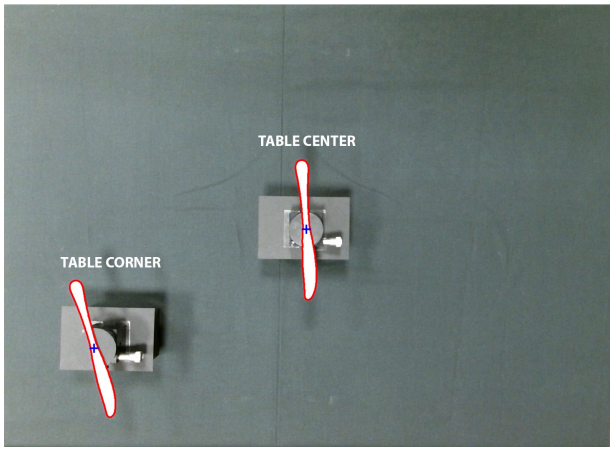
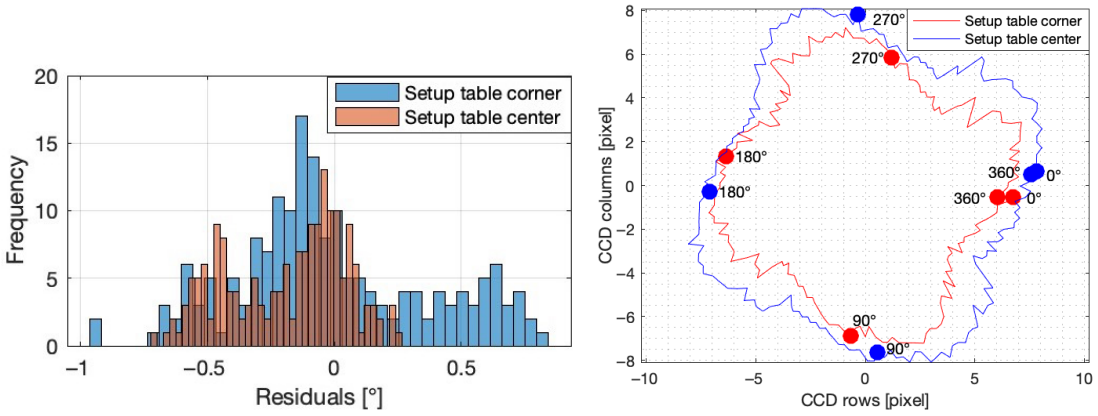


Figure 2.20: Cropped RGB image acquired for rotation tests in the two setups: in the center of the table and near a corner of the table.



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

Figure 2.21: (**a**) Histograms of residuals and (**b**) object center positions during rotations in the two setups.

Figure 2.20 shows both setups: one for tests conducted in the center of the table and one for tests near a corner.

For the rotation tests, 180 acquisitions were performed for each of the two setups from 0 to 360° with a step size of 2°. The decision to carry out these tests on both the center and sides of the acquired images was to estimate better the algorithm's performance over the entire table surface. The differences between the obtained rotations from the SVD algorithm and the one from the rotation motion platform taken as ground truth are shown as histograms of residuals for the two different setups in Figure 2.21a. The histogram spread for the setup at the center of the table is smaller than for the setting near the corner due to the higher camera distortion. However, the residual in estimating rotations for the localization algorithm in general over the entire table surface is less than 1°. During the same rotation tests in the two setups, the object center positions at each step were calculated as the mean of its boundary coordinates. The results of the object centers at each step are shown in Figure 2.21b.

2.4.4 Preliminary User study

An experimental test campaign approved by the ethics committee was also carried out with patients and healthy testers, as shown in Figure 2.22. This preliminary user study aims to assess the statistical significance of the data collected. The parameters analyzed are:

- errors in object placement;
- execution time;
- hand speed;
- breath rate;
- heart beat;
- pressure distribution.

Figure 2.22: User study with four random testers among the eight participants.

In particular, errors in object placement refer to the median error in position and angle between real and virtual objects above the table; execution time quantifies the time between the moment the tester starts to pick up the first real object and the moment he finishes arranging all the objects on the table; hand speed is obtained from the acquired kinematics of the tester. For statistical analysis, we consider its maximum value. When a tester had a problem in either joint, he was forced to perform the entire test using only that one; physiological parameters, such as breath rate and heart beat, are calculated with respect to variations from their basal values; *pressure distribution* of each foot is analyzed with Warren Sarle's bimodality coefficient (BC) [88]. BC lies within a range of 0 to 1, where values greater than 0.555 indicate bimodal or multimodal data distributions.

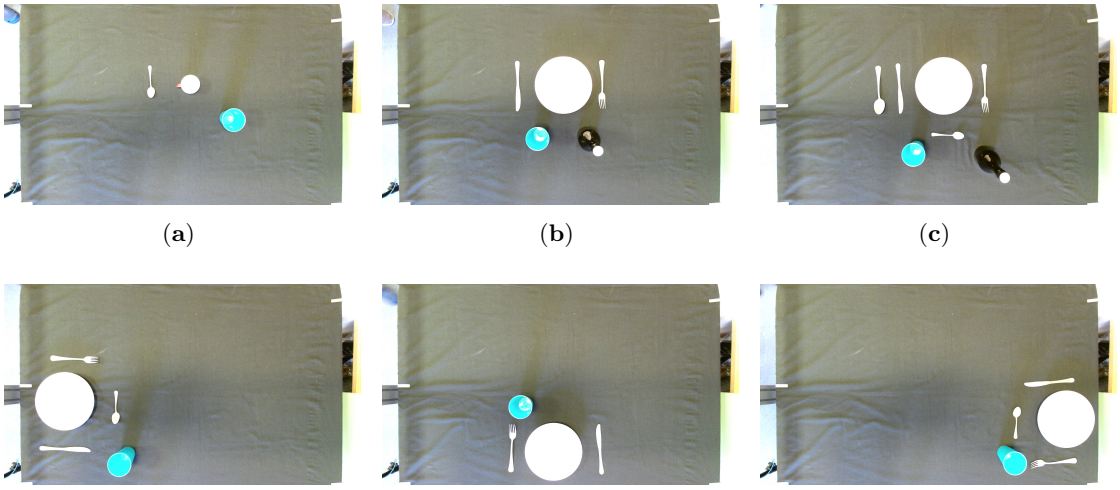
Eight subjects participated in the tests voluntarily after signing a consent form. These were divided into two groups: three were patients, and five were healthy users. The selected patients, with ages between 19 and 69 years, including one female, have different pathologies:

- User 1, C5 incomplete tetraplegia, the major deficit in the left hand.
- User 2, cerebellar ataxia, balance, and stability problems.
- User 3, tetraparesis from Guillain–Barré outcomes, upper limb manipulation deficit.

None of them reported having experience with AR technologies such as HoloLens. Instead of the healthy people, three out of five had already used HoloLens; they are all between ages 20 and 35, including one female.

In the first step, the therapist was trained to set up the table with virtual objects, start the test, and decide whether or not to display some of the available parameters in real-time. The therapist can select six standard table configurations from a smartphone, as shown in Figure 2.23, to give more standardization to the data collected during testing.

The testing protocol was organized as follows:



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

(**d**) (**e**) (**f**)

Figure 2.23: Different table setting configurations: (**a**–**c**) from a simple set-up to a complex one in the center of the table, and (**d**–**f**) from different angles.

1. After completing a consent form, the tester receives an initial explanation of the task.
2. Before starting, basal values of physiological data, such as heartbeat and breath rate, were estimated by acquiring data for 5 min.
3. The therapist starts the tests in sequence: each tester must set the table in any configuration provided by the therapist.

All testers repeated the protocol for three consecutive days. Given the familiarity with the standard table-setting task and the ease of superimposing the real objects with the virtual ones, no initial training was necessary for the tester. We collected all the data acquired on different days in two populations: the one defined by healthy users and the one described by patients. The two-sample t-test [54] is used to compare whether the average difference for each selected parameter between the two populations is significant or if it is due to random effects. Before the t-test, we applied an initial variance test to check whether the two data samples were from populations with equal variances. In case of a negative outcome, it is replaced with Welch's formulas. The results accept the null hypothesis at the 5% significance only for breath rate and heart beat parameters. It means that there is no significant difference between patients and healthy testers for these two parameters. It can be attributed to the simplicity of testers' tasks and the test's short duration. In fact, it goes from an average duration value for healthy testers of 27 s to one of 59 s for patients. The difference in the other parameters allows the two populations to be distinguished. Figure 2.24 shows the boxcharts of errors in object placement and execution time.

The difference in mean execution time between healthy testers and patients shown in Figure 2.24c is more significant than that related to errors in object placement (Figure 2.24a and Figure 2.24b). In many human-performed tasks, the more precisely the assignment is to be accomplished, the slower it is. Fitts' law [122] reveals the correlation between speed and accuracy regarding human muscle movement. In our case, unlike a healthy person, for whom good results can be obtained in less time, i.e., with more speed, for a patient, even with more time, acceptable results can be obtained in terms of errors in object placement. No time limits were imposed

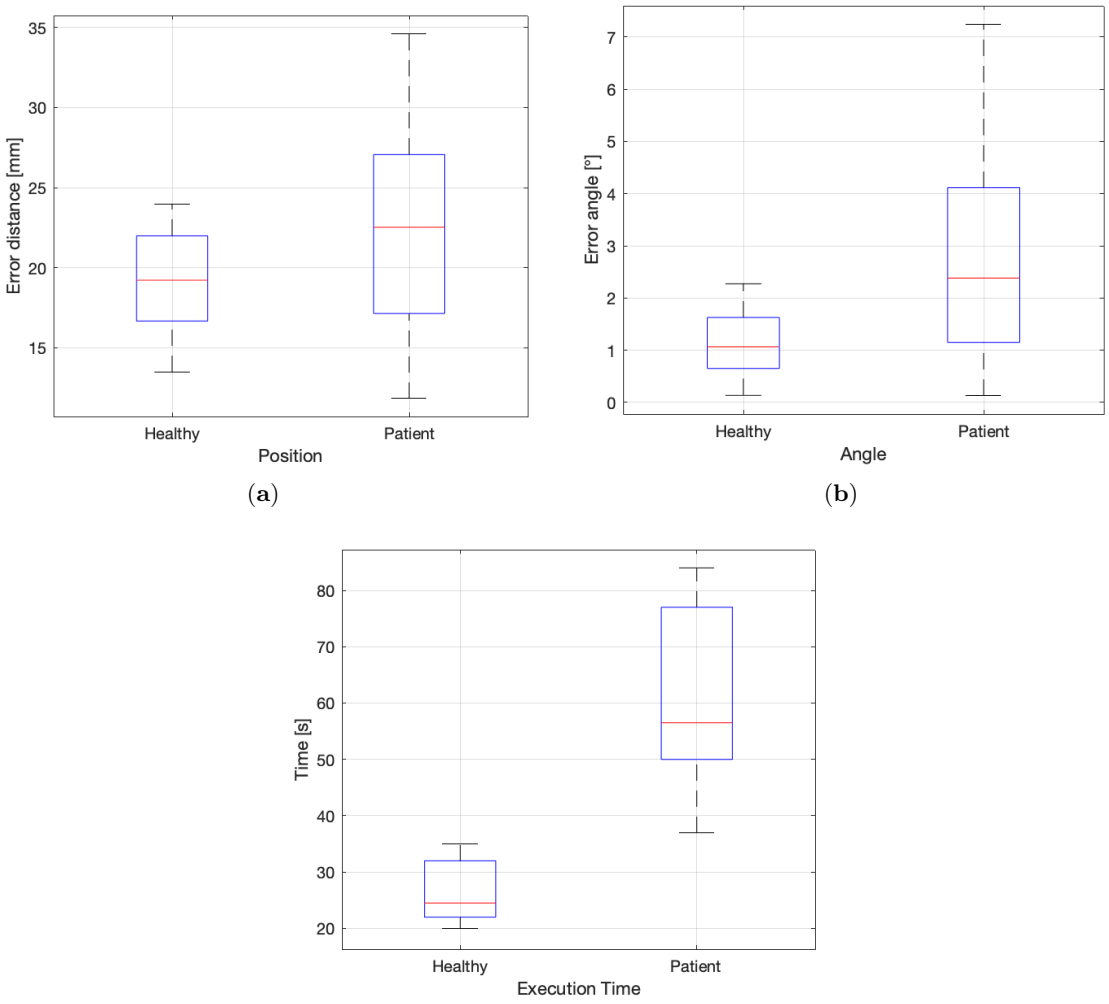


Figure 2.24: Boxcharts of the median error in (**a**) position (p-value = 0.001), (**b**) angle (p-value = 2.4×10^{-7}) and (**c**) execution time (p-value = 4.8×10^{-5}).

during the test, but therapists only told patients to place the objects in the correct position. The above follows Fitts' law trade-off between speed and accuracy: to try to keep accuracy low, the maximum speeds and, therefore, the execution times between patients and healthy testers change. Figure 2.25 shows an example of the speed results at 6 Hz of setting the table in the configuration of Figure 2.23d.

Longer execution times for patients result in lower maximum speed. In fact, for the healthy tester in the example, the maximum speed is higher, and five-speed abrupt changes can be identified, each corresponding to the five objects in the selected configuration. As each object is grabbed from shelves, the speed remains high for the healthy subject, almost without slowing down during the grab control phase. The time to place the object in the correct position is also low and corresponds to the low-speed moments. For the patients, on the other hand, there are many more and smoother variations at low speeds, indicating continuous grabbing and releasing

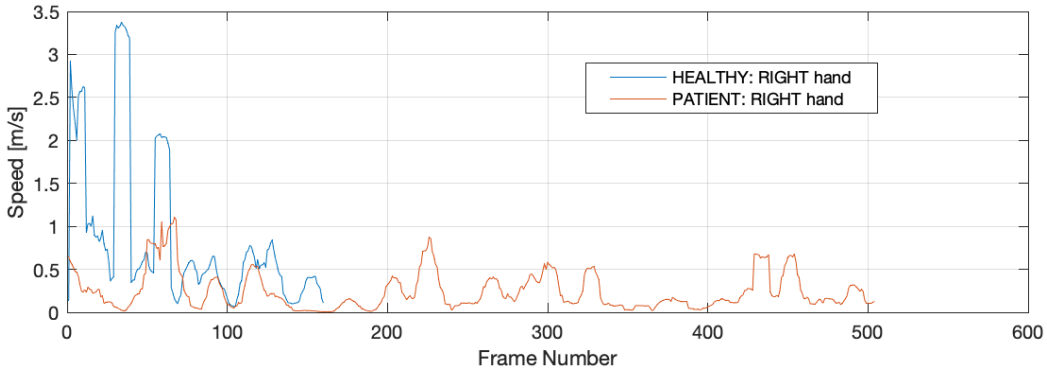


Figure 2.25: Example of speed comparison on the same test between a patient and a healthy tester.

of objects without clean manipulation during the control phase in the final positioning of objects and grabbing them from the shelves.

For the same speed example, we show the result of the pressure distribution, as shown in Figure 2.26.

The healthy tester usually puts all his weight on the leg on the side where he extends the arm he is using. On the other hand, for patients with stability problems, this is not true. The trend of the healthy subject, especially when he sets the table toward the lateral sides, follows a bimodal trend that can be identified with Warren Sarle's bimodality coefficient. Applying the BC to the data in Figure 2.26, we obtain the result in Figure 2.27 where, as might be expected, the BC is greater for the healthy tester for both feet.

In addition, this testing campaign also defined the acceptability threshold of each parameter for patients. We used the results of healthy testers as acceptability thresholds, so, for example, an error of 18 mm for object position (Figure 2.24a) and 1° for its angle (Figure 2.24b) resulted acceptable for patients. Errors may be due to how the HMD glasses were worn or how the virtual images were displayed in AR.

2.4.5 Offline interface

One of the advantages of the designed AR framework is the possibility for therapists to have additional information available to assess patients in real-time and in the correct location near patients. However, they can save all the data collected during testing for further analysis. An interface was designed in MATLAB to read and visualize this information collected once the patient's name, day, and test number were selected. It allows therapists to have an overview of the entire test performed by patients, with the possibility of analyzing multiple parameters synchronized with each other, stopping or moving in time at will. In addition, offline analysis allows patients' performance to be compared even between tests performed at a distance of time. An example of how the offline post-processing interface for each tester looks is shown in Figure 2.28.

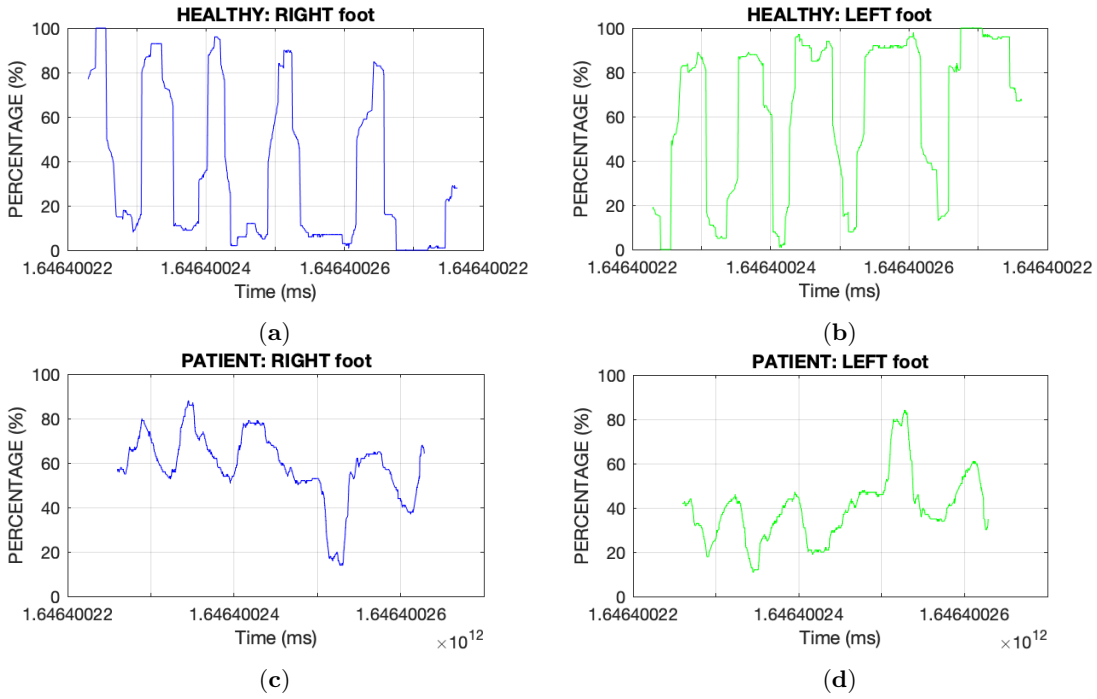


Figure 2.26: Example of pressure distribution on the same test between (**a**,**b**) a healthy tester and (**c**,**d**) a patient with right and left foot, respectively.

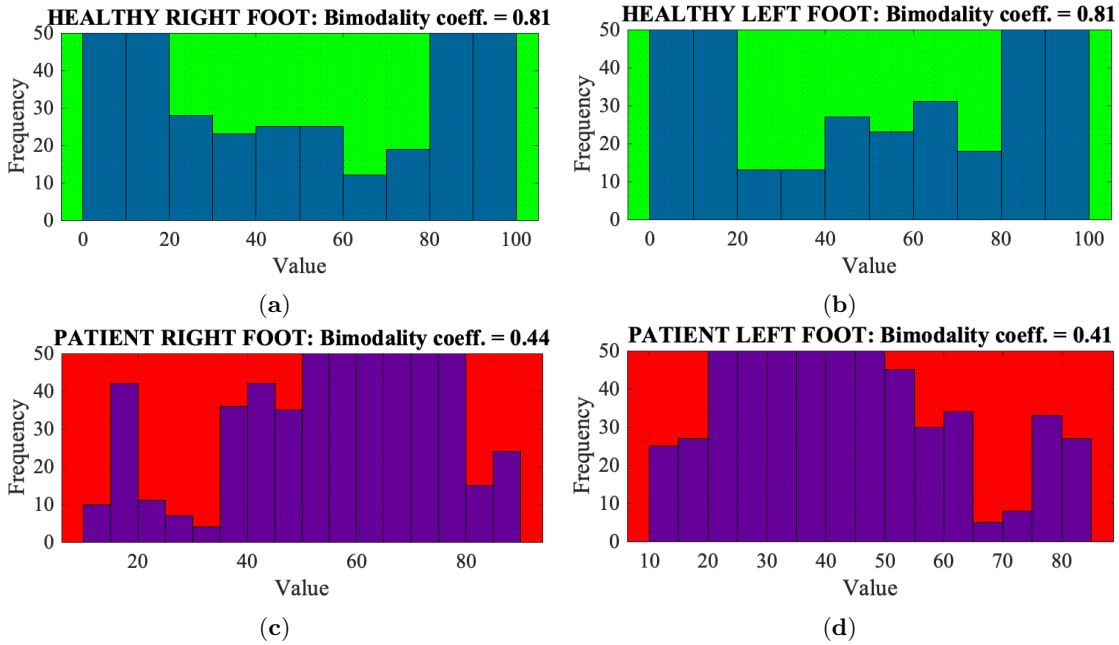
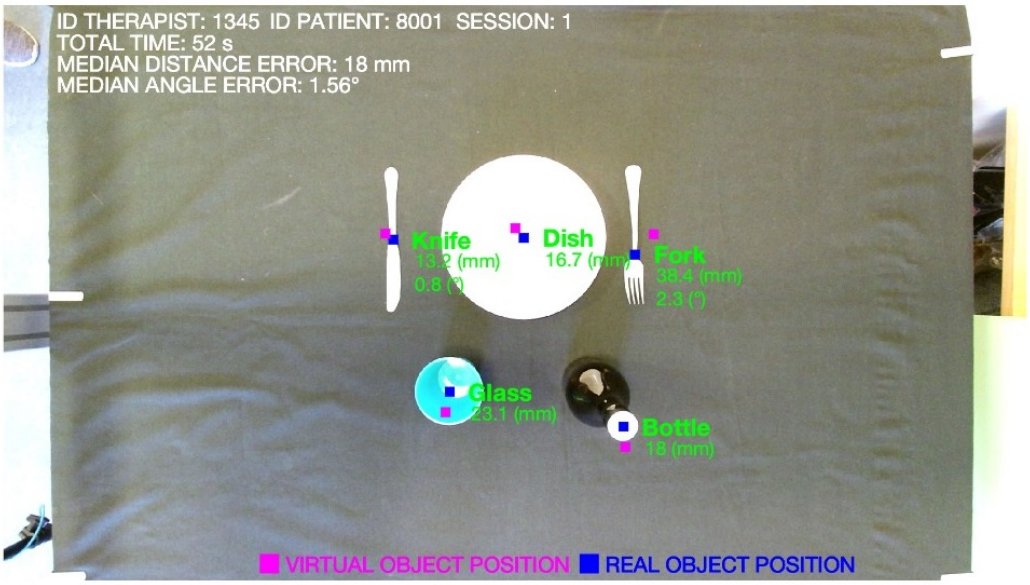
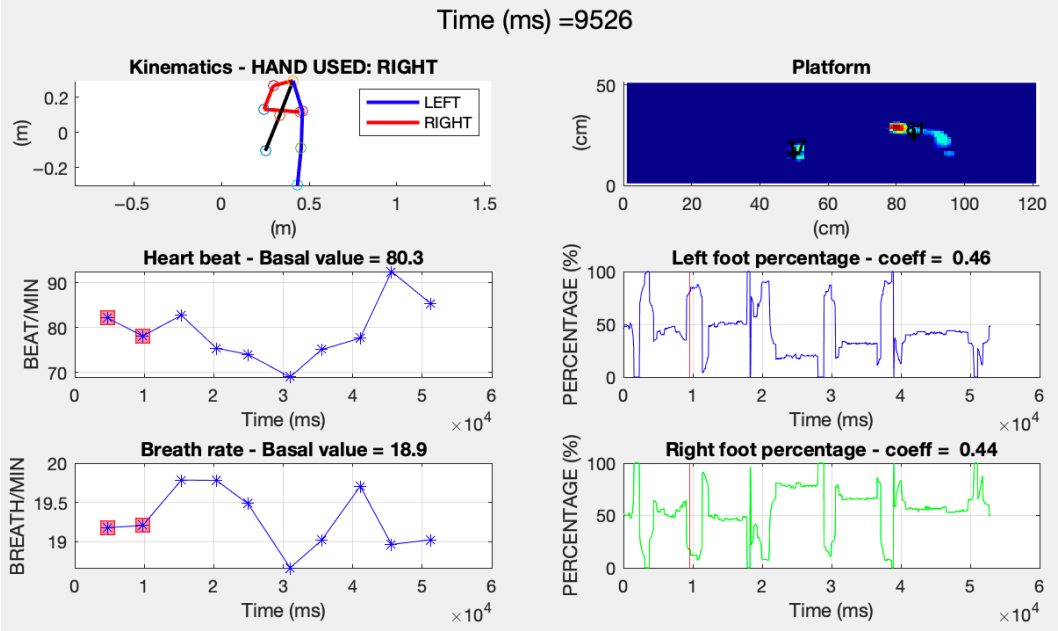


Figure 2.27: Example of Warren Sarle's bimodality coefficient of the same (**a**,**b**) healthy tester and (**c**,**d**) patient from Figure 2.26 data.



(**a**)

Figure 2.28: *Cont.*



(**b**)

Figure 2.28: (**a**) Image with therapist and tester data, all errors in object placement and time of execution; (**b**) all other tester parameters are summarized in this second panel.

Chapter 3

AV in educational settings

Technological advancements have transformed the education landscape in recent years, ushering in a new era of immersive and interactive learning experiences. Among these innovations, MR has emerged as a groundbreaking tool, bridging the gap between the physical and digital worlds to create engaging educational environments [25]. MR offers a unique and promising approach to enhancing teaching and learning processes across diverse disciplines by merging virtual and real-world elements to create a SAV that enhances users' perception, extendable to all levels of the PAL, Figure 3.1.

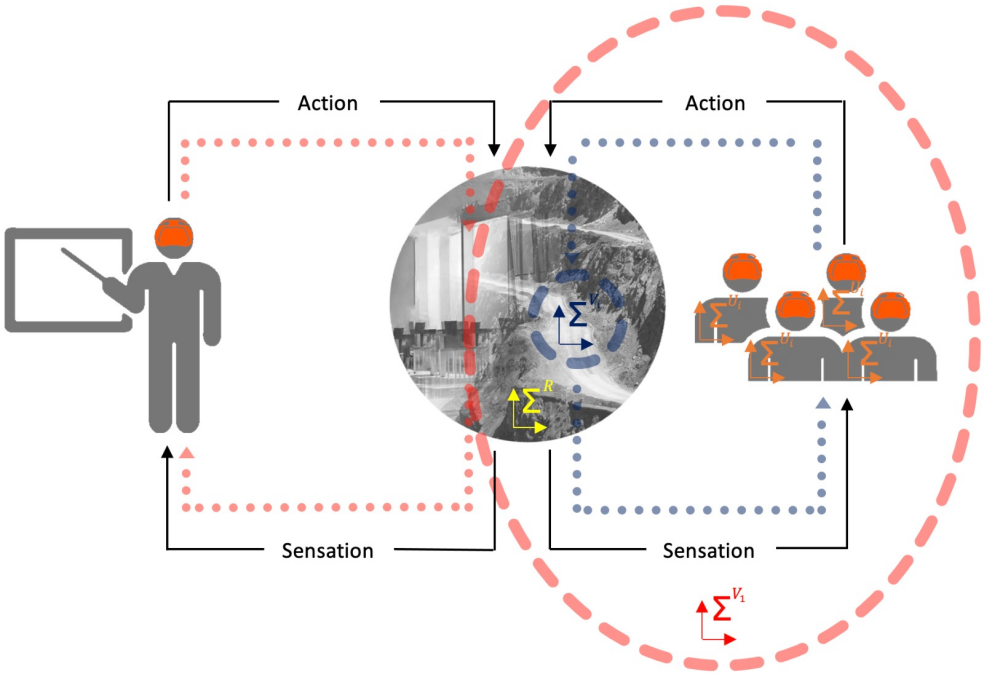


Figure 3.1: The third level of the perception-action loop between teacher, i.e. supervisor, and students in a shared augmented virtuality framework. The external environment R is represented in this context by, for example, a virtual mine.

As educators and learners embrace this cutting-edge technology, questions arise about its

effectiveness, impact on student engagement and understanding, and its ability to address the challenges faced in traditional teaching methodologies.

Here are some key aspects of MR in education:

- Enhanced engagement: MR captivates students' attention and interest by creating interactive and immersive learning experiences. It transforms abstract concepts into tangible visualizations, making complex subjects more engaging and enjoyable.

- Active learning: MR encourages active participation and hands-on experiences, allowing students to explore and manipulate virtual objects in real-time. This interactive approach fosters a deeper understanding and retention of knowledge.

- Personalized learning: MR can adapt to individual learning styles and preferences. Students can interact with content at their own pace and revisit challenging topics until they grasp the concepts thoroughly.

- Real-world application: MR enables students to bridge the gap between theory and practice by simulating real-world scenarios. This practical experience enhances students' problem-solving skills and prepares them for real-life challenges, overcoming the limited resources of practical education classes.

- Collaborative learning: MR can facilitate collaborative learning experiences where students can work together in a shared virtual space. It encourages teamwork, communication, and peer-to-peer learning.

- Access to remote or dangerous environments: MR allows students to safely explore distant or hazardous locations. For example, they can visit historical sites, travel through the human body, or simulate science experiments without physical constraints. In addition, they can "visit" places, such as production plants, power plants, mine sites, and plants for which special permits are mandatory.

- Multi-disciplinary applications: MR is not limited to specific subjects. It can be applied across various disciplines, including science, engineering, arts, history, and more, tailoring experiences to different educational needs.

- Professional training and skill development: MR can be used in vocational education and training, providing learners with realistic simulations of job-related tasks and scenarios. This approach is precious in fields where hands-on experience is crucial.

- Accessibility and inclusivity: MR can accommodate various learning needs, making education more accessible for students with disabilities or learning difficulties.

- Continuous innovation: As MR technology advances, the educational potential will continue to grow, opening up new possibilities for creative and dynamic learning experiences.

In educational settings that use innovative technologies to enhance the experience of students and teachers, MiroLab researchers of the University of Trento were involved in the MiReBooks EIT RawMaterials project [22].

3.1 MiReBooks

The MiReBooks project aims to redefine higher education in mining in Europe by developing an innovative series of interactive manuals on mining, leveraging virtual and augmented reality technology, Figure 3.2. It stands for Mixed Reality Handbooks for Mining Education [49].

This initiative aims to address the challenges prevalent in mining education by synergizing traditional paper-based teaching materials with MR elements, resulting in comprehensive and pedagogically cohesive MR manuals for seamless integration in the classroom.

Figure 3.2: Example of a typical lesson on mining in Augmented Virtuality. © MiReBooks via MiReBooks website

In addition, the project's innovative approach has potential for application in various academic disciplines. With the implementation of MiReBooks, the teaching landscape is poised to transform, enabling teachers to significantly improve student engagement, provide a wealth of enriched content, and open up new opportunities to improve comprehension.

Implementations of MiReBooks open a wide range of examples of industrial mining environments for students to explore, leading to a deep understanding of the industry context. This comprehensive immersion provides graduates with digital native skills, enabling them to influence and significantly shape the industry's future. With MR at its core, MiReBooks promises to optimize the learning experience, drive operational efficiency and foster innovation.

In MiReBooks-assisted classes, students will use specialized smartphone applications to access augmented illustrations embedded in textbooks. These illustrations activate additional information, providing valuable insights. In addition, students can wear virtual reality goggles, which transport them into immersive virtual mining environments or 3D filmed sequences of real mining processes. This combination of technologies enhances the learning process and paves the way

for better operational practices and innovative approaches.

Within the MiReBooks project, TU Bergakademie Freiberg's task in Germany was preparing a test lecture on continuous mining methods using AV. AV technologies have generated new challenges to make the experience as immersive and realistic as possible, Figure 3.3. The first challenge involves adding depth perception to 360° images [64] to obtain a final photorealistic 3D model of the environment, while the second includes estimating the poses of objects from 360° videos to give the possibility to interact with their 3D virtual models [121]. These challenges will be explored in detail in the following sections of this chapter.

Figure 3.3: Summary of challenges overcame to allow virtual lessons on mining. First challenge: obtain a photorealistic 3D model of the mine environment; Second challenge: estimating truck poses from 360° videos.

3.2 Photorealistic 3D model

Media acquired by 360° cameras (also known as omnidirectional, spherical, or panoramic) is becoming increasingly important to many applications. Compared to conventional cameras, images taken by 360° cameras offer a larger FOV, which is why they are traditionally useful for applications that derive their state from environmental information. Examples include robot localization, navigation, and visual servoing [10]. However, omnidirectional cameras have recently also become an essential tool for content creation in AV applications because spherical photographs and videos can provide high realism. For example, applications for real estate agents already make use of omnidirectional images and video data within AV headsets to improve the realism of virtual customer inspections and research domains span widely from 360° tourism [36] to education in 360° classrooms [48]. AV applications using omnidirectional media allow users to change the view within the boundaries of a 360° image captured at a specific Point of Interest

(POI). Thus, AV users are commonly restricted to head rotations only, while translations require transitioning into a 360° image captured at a different POI [67]. Thus, motion parallax is missing in AV applications, which use omnidirectional data. Furthermore, view transitions are limited to where omnidirectional images or videos exist. These shortcomings limit the benefit of omnidirectional media in AV. For example, the missing 3D information restricts the usage of advanced exploration techniques [107, 106] and the missing motion parallax can cause visual discomfort [110]. The proposed work combines omnidirectional photorealistic image data with the corresponding 3D representation to overcome these limitations. Since 3D reconstructions commonly suffer from poor color representations, a projective texture mapping of omnidirectional images is applied. This approach supports photorealistic image fidelity at the POIs and motion parallax at viewpoints nearby. To enable projective texture mapping of 360° image data, the presented approach involves omnidirectional camera pose estimation that automatically identifies the position and orientation of the 360° camera relative to the 3D representation of the environment. In order to contextualize the work, an overview of related works is provided, followed by a description of the methodologies employed for omnidirectional camera pose estimation and projective texture mapping. Finally, the system is subjected to an evaluation, and potential pathways for future research are discussed.

3.2.1 Related work

Camera pose detection has always been a key problem in computer vision. For example, Makadia et al. [69] proposed a useful method for aligning large rotations with potential impact on 3D shape alignment to estimate the rotation directly from images defined on the sphere and without correspondence. Unfortunately, this approach is quite resistant only to small translations of the camera [68]. Another work [57] addresses the problem of camera pose recovery from spherical panoramas using pairwise essential matrices. In this case, the exact position of each panorama was an important step to ensure the consistency of visual information about a database of georeferenced images. Here the pose recovery works with a two-stage algorithm for rotations and after for translations with a bad result if the camera starting pose is far from the correct one. The problems mentioned above have been overcome by the proposed method in this thesis because it also works for significant variations of translation and rotations. Also, Levin et al. present in [60] a method to compute camera pose from a sequence of spherical images using an essential matrix for initial pairwise geometry. Differently from the proposed work and the work of [57], they also use a rough estimate of the camera path as an additional system input to calculate camera positions. An example of generating a texture map of a 3D model with 2D high-quality images is given in [58]. In particular, it is a specific application in the e-commerce presentation of shoes. It consists of a texture mapping technique that comprises several phases: mesh partitioning, mesh parameterization and packing, texture transferring, and texture correction and optimization. In particular, in the texture transferring step, each mesh is allocated to a front image, and all meshes that use the same front image are put in a group. Finally, the pixels from the front image corresponding to the 3D mesh are extracted. Differently, the proposed method uses only a spherical image to recreate the high-resolution 3D model by projecting each pixel of the image from the correct camera pose previously found. The obtained results are faster and better if the user's FOV rotates without large displacements concerning the camera pose. A similar approach but for another application related to realizing surveying tasks in architectural, archaeological, and cultural landscape conservation is provided by Abmayr et al. [1]. They developed a laser scanner that offers high-accuracy measurements of object surfaces, combined with a panoramic color camera, to achieve precise and accurate monitoring of the actual environment by employing colored point clouds. The camera rotates according to the same tripod as the laser scanner. Many

similarities with the method described in the present thesis can be found. The main difference resides in using a single 360° camera instead of a rotating unit and using an automatic pose estimation method instead of using the same tripod for the laser scanner and camera during the acquisition process. The proposed method in this thesis is faster, and the 3D model reconstruction can be more complete because it does not need to be at a fixed distance from the camera during the scanning process. This aspect becomes more important if it is necessary to reconstruct a high-resolution model with different cameras from unknown positions. Finally, an interesting study was provided by Teo et al. [108], where, in the context of remote collaboration, helpers shared 360° live videos or 3D virtual reconstructions of their surroundings from different places to work together with local workers. The results showed that participants preferred having both 360° and 3D modes, as it provides variation in controls and features from different perspectives. The method proposed in this thesis combines a 360° live video and 3D virtual reconstruction to combine their advantages without switching between them.

3.2.2 Method

This section explains the localization algorithm to estimate the camera pose (i.e., its positions and orientations in the environment) and the method used to project the texture mapping on a 3D representation of the environment. These two tasks are the basis of the proposed solution, which aims to achieve a photorealistic 3D model suitable for VR experiences.

Camera pose estimation

A good alignment between the virtual environment and the captured image is fundamental for the final texture projection covered in the next chapter. For example, this step is necessary when an operator needs to place the camera in a predefined position and orientation. Some human errors may be made during this operation, and a method to find an accurate camera pose is necessary. Moreover, a slight angle or minor position error can compromise the final result for considerable distances. The large-scale automatic camera pose identification algorithm has been implemented in Matlab 2019b using a ZMQ communication protocol between Matlab and Unity 3D. Particle Swarm Optimization (PSO) was used. The procedure of the camera pose estimation is shown in Figure 3.4.

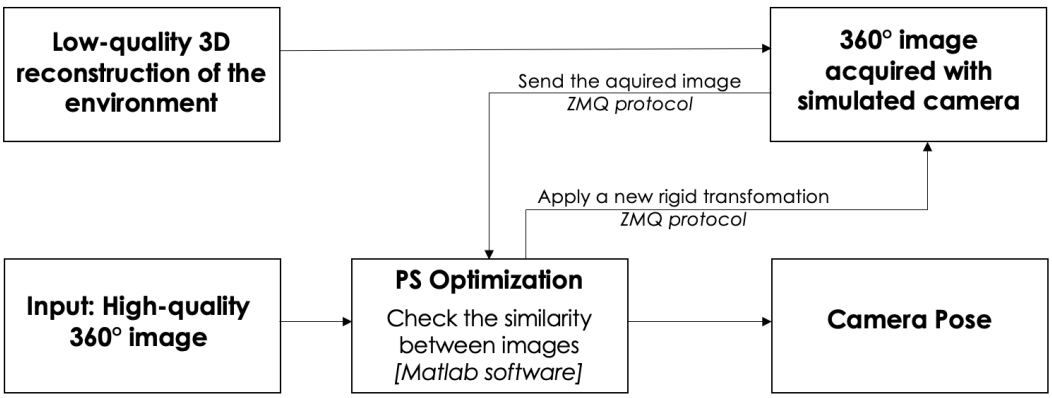


Figure 3.4: Schematic diagram of the camera pose detection algorithm.

Starting from the reconstructed 3D model with its low-quality texture but with depth infor-

mation of the environment and given as input a high-quality equirectangular photorealistic image taken by an omnidirectional camera, the localization algorithm finds the pose that gives a 360° image taken with a simulated camera that is as similar as possible to the input one. Specifically:

i. A new camera position is set for each iteration of the PSO algorithm.

ii. The equirectangular image corresponding to the set camera pose at the previous step is acquired.

iii. The algorithm checks the similarity between the new image and the input one that has to be used as a new texture for the 3D mesh; the parameters to be optimized are the translation and the Euler angles to be applied to the 3D model to generate an equirectangular image that matches the one in the input. The cost function for comparing the two equirectangular images uses the following quantities:

- The structural similarity ($SSIM$) index of the equirectangular images.
- The mean-squared error (MSE) between the two equirectangular images.
- $SSIM$ of the approximation coefficients ($SSIM_A$) of level 1 of the wavelet decomposition.
- $SSIM$ of the horizontal detail coefficients ($SSIM_H$) of level 1 of the wavelet decomposition;
- $SSIM$ of the vertical detail coefficients ($SSIM_V$) of level 1 of the wavelet decomposition;
- $SSIM$ of the diagonal detail coefficients ($SSIM_D$) of level 1 of the wavelet decomposition.

The final cost function C obtained by adding the quantities mentioned above is:

$$C = SSIM + MSE + SSIM_A + SSIM_H + SSIM_V + SSIM_D. \tag{3.1}$$

The MSE represents the cumulative squared error between two images $x(i,j)$ and $y(i,j)$:

$$MSE(x,y) = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}[x(m,n) - y(m,n)]^2, \tag{3.2}$$

where M and N are the number of rows and columns of x and y.

$SSIM$ is used for measuring the similarity between two images x and y [113]. The $SSIM$ Index quality assessment index is based on the computation of three terms, namely the luminance term l, the contrast term c, and the structural term s. The overall index is a multiplicative combination of the three terms:

$$SSIM(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma}, \tag{3.3}$$

where:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{3.4}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{3.5}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}. \tag{3.6}$$

μ_x,μ_y,σ_x,σ_x and σ_{xy} are the local means, standard deviations, and cross-covariance for images x and y. C_1,C_2,C_3 are constants to avoid instability for image regions where the local mean or standard deviation is close to zero. Choosing α=β=γ=1 and C_3=$\frac{C_2}{2}$, the index simplifies to:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_x\sigma_y + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{3.7}$$

iv. the PSO optimization runs until convergence, giving as output the best camera pose (translation and Euler angles) that makes the two images as similar as possible.

Texture projection

This chapter describes the method to apply high-quality texture mapping. Essentially, a merge of the high-quality360° image with the 3D mesh is performed. Firstly, the 3D Cartesian coordinates and colors of each 360° image's pixel were obtained by projecting the equirectangular image on the surface of a unitary radius sphere. Given an equirectangular image with N rows and M columns, each image's pixel in 2D Cartesian coordinates (n, m) was transformed in spherical coordinates, computing the corresponding azimuth a and elevation e, setting the radius R equal to 1. The equations used for the conversion are:

$$a = -(\frac{m}{M} - 0.5) \cdot 2\pi \tag{3.8}$$

$$e = -(\frac{n}{N} - 0.5) \cdot \pi \tag{3.9}$$

$$R = 1. \tag{3.10}$$

Finally, the 3D Cartesian coordinates are obtained to be visualized in Matlab software like a 3D point cloud. The mapping from spherical coordinates to 3D Cartesian coordinates is:

$$x = R \cdot cos(e) \cdot cos(a) \tag{3.11}$$

$$y = R \cdot cos(e) \cdot sin(a) \tag{3.12}$$

$$z = R \cdot sin(e) \tag{3.13}$$

This "spherical" point cloud was imported inside Unity and placed with the position and orientation found in the previous pose estimation step chapter.

The Raycasting technique was used: through the Ray class, it is possible to emit or "cast" rays in a 3D environment and control the resulting collisions. The rays used in Raycasting are invisible lines with the center of the image sphere as the origin and are oriented in each pixel's direction. The key point is that these invisible lines or rays cast into the scene can return information about GameObjects that the rays have hit.

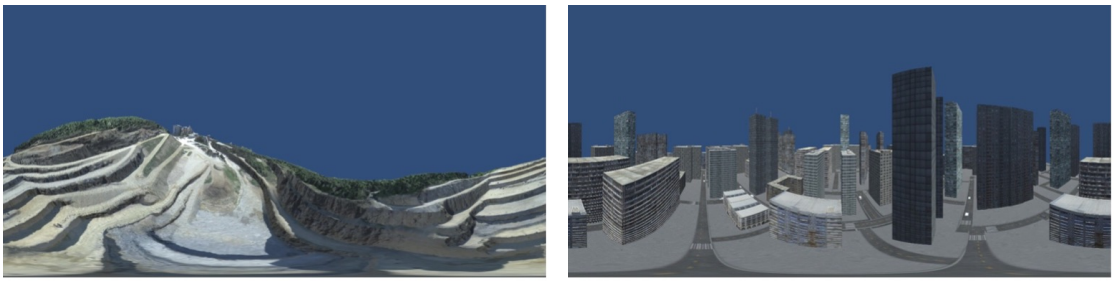
Attached to the environment's mesh as GameObject in Unity is a Mesh Collider to register a hit with the ray. When a ray intersects or "hits" a GameObject, the event is referred to as a RaycastHit. This hit provides details about the GameObject and where it was hit, including a reference to the GameObject's Transform, the length of the ray when it hits something, and the point where the hit happened.

Once the collision of each pixel is detected, their new position is saved with color properties. Lastly, the new point cloud was used to reconstruct a high-quality photorealistic

texture, using the Screened Poisson Surface Reconstruction algorithm [52] implemented in Meshlab [20]. This algorithm is particularly useful when the model to reconstruct is very big, with fine details to be preserved. The reconstruction of the 3D model was done by setting the Reconstruction Depth parameter (i.e., the maximum depth of the octree used to make the reconstruction) to 13. The default value of Meshlab for this parameter is 8; it was increased because, in general, the higher this value is, the more time will be needed for reconstitution, and the more details will be preserved [52]. It was kept at 13 because, after 14, it is not possible to see a real change in the final result. The Minimum Number of Samples was set to 1.5, and the Interpolation Weight to 4 as the default values of Meshlab. Since the Poisson algorithm tends to "close" the reconstructed mesh, the triangles whose area was above a certain threshold were deleted to preserve the original form of the reconstructed environment.

3.2.3 Evaluation

For the validation of the camera pose localization algorithm and the high-quality texture mapping projection, a Wavefront 3D Object File (OBJ file extension) of two 3D high-quality virtual outdoor environments, one for a mine and one for a city, were imported into Unity 3D platform. An original script was also written to simulate a 360° camera. The 360° capture technique is based on Google's Omni-directional Stereo (ODS) technology using Cubemap rendering [34]. After the Cubemap is generated, it is possible to convert this Cubemap to an equirectangular map which is a projection format used by 360° video players. After placing the simulated camera at a specific pose inside the scene of a specific scenario, a high-quality equirectangular image was acquired, Figure 3.5. These will be the input images whose pose has to be detected by the developed algorithm.

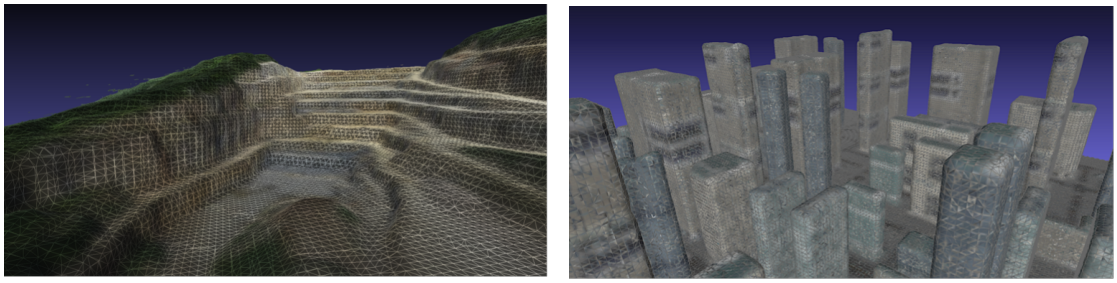


(**a**) (**b**)

Figure 3.5: High-quality equirectangular images whose detection poses must be identified for a mine (**a**) and city (**b**) environments.

To simulate the acquisition of the environment through a 3D scanner, a point cloud for each analyzed environment was extracted from the 3D high-quality models using the Cloud Compare software [33]. These point clouds were downsampled to simulate a 3D model with less detail than the input model, and new reconstructions were performed in MeshLab [20] to obtain new low-quality 3D models, Figure 3.6. New scenes were then recreated in Unity with the downsampled 3D models.

Figure 3.7 shows the schematic diagram of our camera pose detection algorithm proposed in Figure 3.4 applied to the specific example of the mine environment. The input omnidirectional image has a resolution of 4096 X 2048 pixels. However, to improve the calculation time speed, the comparison between images is done by downsampling them to 256 X 128 pixels for both the



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

Figure 3.6: The 3D downsampled models used by the localization algorithm for a mine (**a**) and city (**b**) environments.

analyzed environments. The bounding box dimensions of the scenario with the mine are 113 m x 169 m x 37 m for the x, y, z coordinates, respectively. Instead, the dimensions of the city environment are 440 m x 100 m x 435 m.

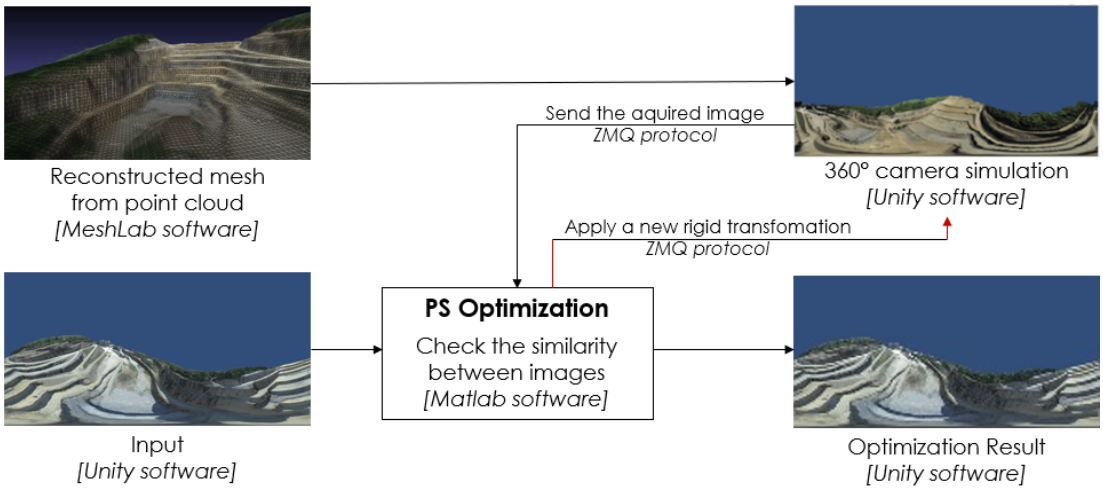


Figure 3.7: Example of the camera pose detection algorithm flow for the mine environment.

The same analysis was done for both environments using the same approach and shifting the camera pose by the same values. Table 3.1 shows the position and orientation for ten random trials. The initial starting position was set to the origin $(0, 0, 0)$ with null rotations for each trial. The research limits were set to ± 20.00 m for translations and $\pm 80.00°$ for rotations.

By default, Unity applies the following rotation order: Extrinsic Rotation around the z axis (γ), then around the x axis (α), and finally around the y axis (β). The average time spent by the PSO algorithm is around 20 min. The tests were run on a PC with an Intel i7-9700KF processor and 64.0 GB of RAM.

For each of the ten trials of Table 3.1, the PSO algorithm has been run, changing five times the numbers of generations, i.e., 200, 250, 300, 350, 400, keeping the number of particles fixed to 100, and five times changing the number of particles, i.e., 60, 70, 80, 90, 100, keeping the number of generation fixed to 400. The number of generations and particles was changed to force the algorithm to increase variability.

Trial	x [m]	y [m]	z [m]	α [°]	β [°]	γ [°]
1	-4.00	10.00	15.00	10.00	15.00	18.00
2	5.00	-2.00	5.00	10.00	-60.00	1.00
3	-8.00	5.00	-6.00	30.00	45.00	15.00
4	2.00	-7.00	15.00	-10.00	-45.00	-20.00
5	10.00	10.00	10.00	20.00	-15.00	5.00
6	0.00	15.00	8.00	25.00	-15.00	6.00
7	-5.00	2.00	-5.00	-10.00	60.00	-1.00
8	-1.00	-2.00	-3.00	-4.00	-5.00	-6.00
9	-15.00	10.00	10.00	40.00	70.00	40.00
10	-19.00	19.00	-19.00	2.00	80.00	-5.00

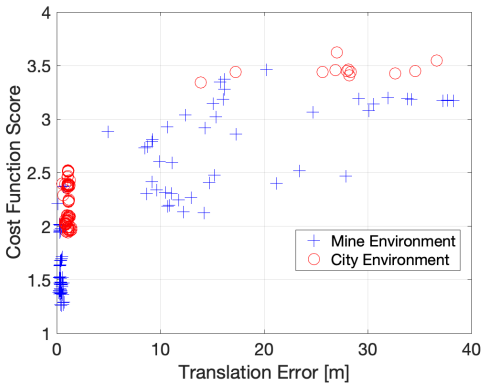
Table 3.1: Camera poses chosen for 10 trials (ground truth).

To compute the pose detection error, the translation and the rotation part were separated. The translation error is computed by performing the Euclidean distance between the camera position found by the PSO algorithm and the ground truth. For what concerns the rotations, firstly, the rotations found by the optimization process and the ground truth were decomposed in axis and angle notation. Consequentially, the error, in the case of rotation, has two terms: the error in the axis orientation with respect to the ground truth and the amount of rotation around such axis.

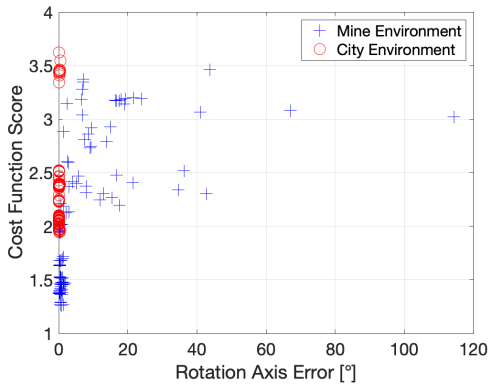
Figure 3.8 shows the cost function score for the various error components explained above (Equation (3.1)), while Figure 3.9 shows the three possible couple combinations of the error components with respect to the final score optimization value.

As can be noticed, sometimes, a higher cost function score at the end of the optimization does not mean an incorrect pose was found. This fact is probably due to the mesh reconstruction process. Indeed, after this process, portions of the environment could be less accurate compared to the real model. For this reason, considering different camera poses, the meaning of the final reached score values is not absolute or easily comparable.

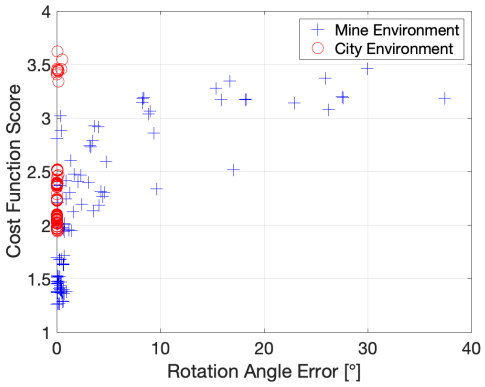
This generates the need to quantify the accuracy of the camera localization measurement within a scene. Despite the uncertainty concerning the accuracy in the pose found by the algorithm with respect to the final cost function score, Figure 3.8 and Figure 3.9 show that, for this particular environment, a score below 1.6 ensures that an accurate result has been obtained. In particular, a score below 1.6 means that, for the trial performed, the error in translation is below 0.7 m, the difference in the amount of rotation is below 1°, and the difference in the rotation axis orientation is below 2°. The same errors correspond to a cost function score of 2 for the city environment. The score is higher because the city environment is a scenario with much more detail than a mine. Many of these details, through initial downsampling, are lost, and the initial reconstructed mesh is much less detailed, as seen in Figure 3b. The final score, therefore, which measures the similarity between the input high-quality equirectangular image and that obtained from this low-quality model, turns out to be higher. However, the errors, especially those related to rotations (Figure 3.8b and Figure 3.8c), are lower for the city environment even at high levels of the cost function score because the environment is different. Because of this relationship of the cost function threshold from the level of detail of the reconstructed 3D model, there is a need for further analysis to investigate possible acceptance criteria and multidimensional models capable of finding a correlation between the different terms of the cost function and the uncertainty in translation and rotation. For example, Figure 3.10 shows that MSE could be a possible discriminant factor for accuracy. Indeed, in this case, the accurate solutions are all centered



(**a**) Cost function score vs translation error.



(**b**) Cost function score vs axis orientation error.



(**c**) Cost function score vs rotation angle error.

Figure 3.8: 2D plots of the cost function score vs the errors in translation (**a**), axis orientation (**b**), and rotation angle (**c**).

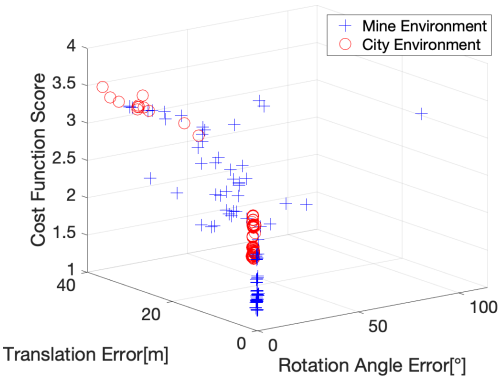
around the 0.005 value for both examinated environments.

Once the camera poses were found for each environment, this information is used to set the 360° image projected on the surface of a unitary radius sphere in the correct position and orientation, Figure 3.11a. After that, using the Raycasting technique, the 3D mesh, Figure 3.11b. is hit by 360° image pixels, Figure 3.11c.

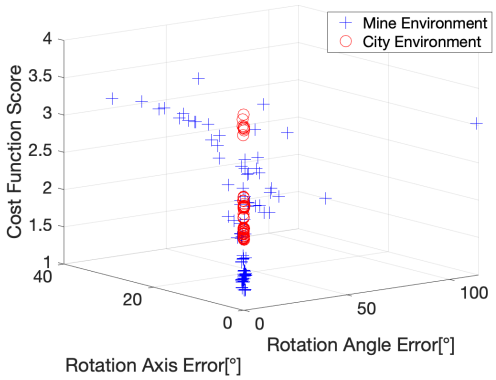
The final reconstructions of the high-quality 3D models using the Screened Poisson Surface Reconstruction algorithm implemented in Meshlab are shown in Figure 3.12a and Figure 3.12b for the mine and city environments, respectively.

3.2.4 NeRF for high-quality 3D rendering

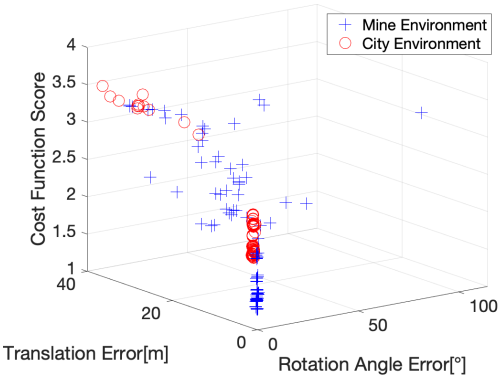
The previous approach presents a method developed to obtain a high-quality textured mesh by combining a raw 3D mesh model of the environment and 360° images. It supports head rotations around all three axes. While this enables immersive experiences, the missing translations may cause several perceptual issues [109], limiting explorations to the pre-defined viewpoint. To overcome this problem, the same method can be applied with more than one camera. However, for the final texture reconstruction, it is necessary to define a discriminating parameter to choose



(**a**) Cost function score vs translation error vs rotation angle error.



(**b**) Cost function score vs axis orientation error vs rotation angle error.



(**c**) Cost function score vs translation error vs axis orientation error.

Figure 3.9: 3D plots of the cost function score and the errors in translation, rotation angle, and axis orientation.

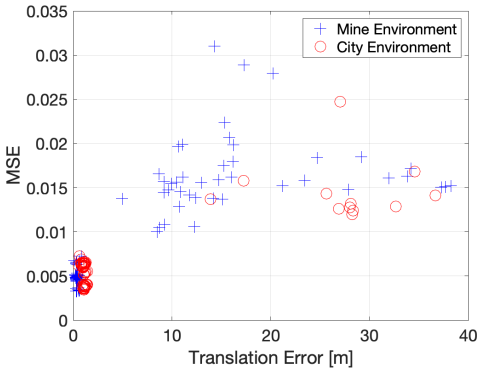
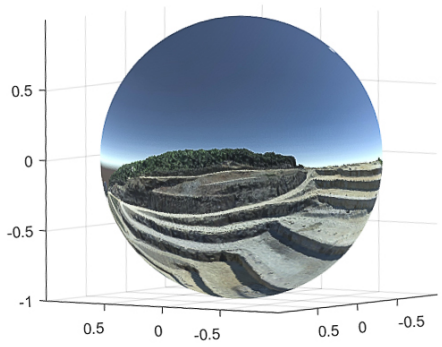
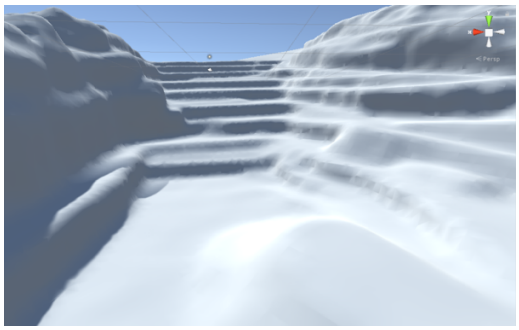


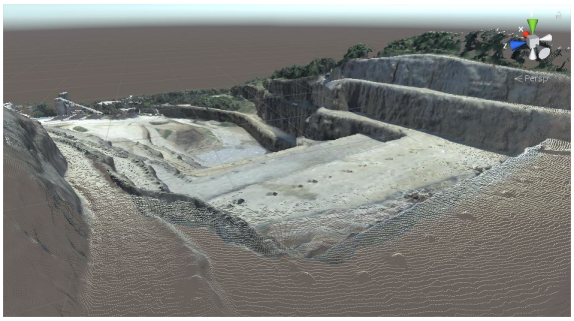
Figure 3.10: MSE score vs translation error.



(**a**) 360° image placed on the surface of a unitary sphere (Matlab Software).

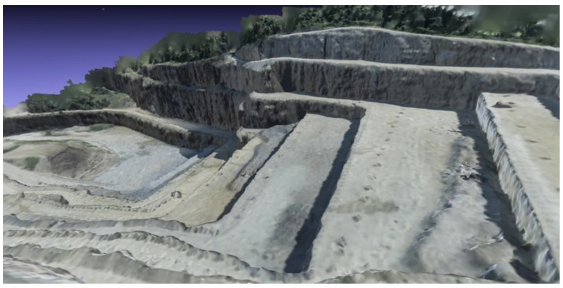


(**b**) Raw 3D mesh (Unity Software).

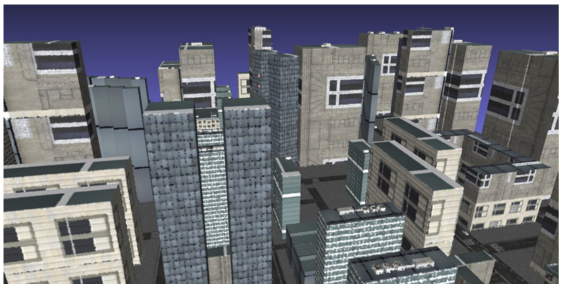


(**c**) Point cloud obtained projecting the pixels of the 360° image on the raw 3D mesh (Unity Software).

Figure 3.11: The pixels of the 360° image of the mine environment are projected on a sphere surface (**a**), which is put in the correct camera pose found by our algorithm inside the raw 3D mesh (**b**). The pixels are then projected using the ray cast technique on the raw mesh, obtaining a new dense point cloud (**c**).



(**a**)



(**b**)

Figure 3.12: Final results after the 3D reconstruction for the mine (**a**) and the city (**b**) environments.

which pixels to use from one or the other camera. This new parameter can be useful if the FOV of one camera is better for some areas of the mesh than another, allowing for a better texture reconstruction result.

In order to avoid this iterative process, in recent years, a novel technique NeRF (Neural Radiance Fields) [72] based on deep learning architecture directly generates high-quality 3D renderings from a collection of 2D images or videos, Figure 3.13.

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

Figure 3.13: NeRF input as a set of calibrated images (**a**) and output a 3D scene representation (**b**). © from [73].

NeRF represents an object with a neural network that outputs colour and density for each point in 3D space. Colour and density values are accumulated along rays, one ray for each pixel in a 2D image, Figure 3.14.

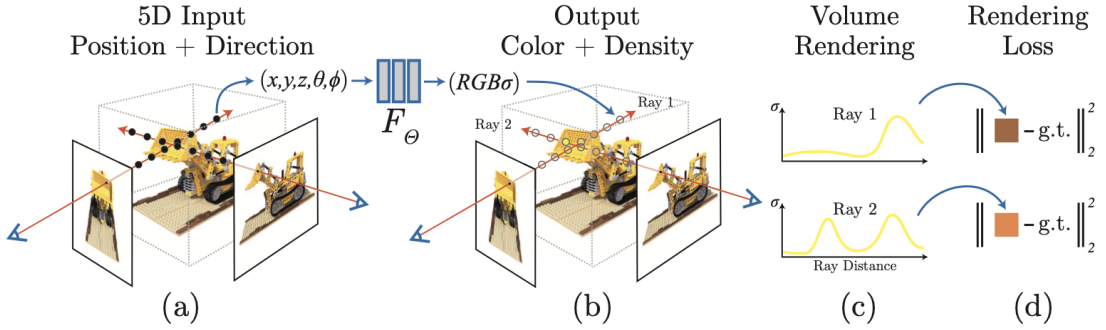


Figure 3.14: An overview of NeRF scene representation and differentiable rendering procedure. © from [73].

In particular, NeRF algorithm aims to learn a function f that can render novel views of the scene from any viewpoint. This function takes as input a 5D coordinate (x, y, z, θ, ϕ), where (x, y, z) is the spatial location and (θ, ϕ) is the viewing direction. The output of f is a 4D vector (r, g, b, σ), where (r, g, b) is the RGB colour and σ is the volume density at that point.

The NeRF algorithm trains the neural network by minimizing a rendering loss, which measures the difference between rendered and input images. The rendering process is based on volume rendering, which simulates how light travels through a medium and interacts with it. It uses gradient descent to update the weights of f such that the rendered images match the input images as closely as possible. Hierarchical sampling, positional encoding, and fine-tuning are used in NeRF algorithm to improve the performance and quality of the final rendering.

Due to its high rendering speed and good results, it is becoming very popular in computer graphics and computer vision. While photogrammetry [71] is suitable for static scenes, NeRF excels in capturing dynamic and highly detailed scenes. In photogrammetry, the images are processed to identify corresponding points, from which the spatial relationships are reconstructed using triangulation algorithms. NeRF, on the other hand, leverages deep learning to model the volumetric scene representation directly, enabling the synthesis of novel views with intricate details and realistic lighting effects. It can also handle occlusions, transparency, reflections, and other challenging effects that are difficult for traditional 3D reconstruction methods.

It is constantly evolving, and new NeRF-based algorithms are emerging [75, 93, 9, 78] to improve the performance of NeRF in terms of speed and quality of final 3D rendering under every possible condition. The training time has significantly decreased from the initial version of NeRF in 2020 [72], which required approximately 12 hours, to the latest InstantNGP in 2022 [75], which takes about five minutes.

These properties make NeRF suitable for applications such as novel view synthesis, AV, and AR, where the goal is to generate realistic and immersive 3D renderings of the scene from different perspectives.

3.3 3D Interaction from 2D images

The second challenge aims to enhance student learning with a more immersive and engaging experience by enabling the interaction with 3D virtual models of objects from 360° videos, Figure 3.15.

To achieve this goal, a general approach, which can be used for any object and shape, was developed to allow the transition to the objects' 3D model from their current pose. In particular, a method was designed to estimate objects' 6 Degrees of Freedom (DoF) pose in equidistant 2D images, making 3D interaction possible. 6DoF estimation is one of the main challenging research topics in computer vision [39, 111, 43].

The developed pipeline has two main steps: vehicle segmentation from the image background and estimation of the vehicle pose. Deep learning methods were used to perform the first task, and for the latter, the same Unity simulator seen in Section 3.2 was used to generate the equirectangular synthetic images used for comparison.

3.3.1 Related work

6DoF pose estimation using RGB images involves different fields such as bin picking problems [5], robot manipulation [11], autonomous vehicles [92], and MR applications [50].

Usually, to accomplish this task, deep learning methods are used. One of the main approaches to 6DoF pose estimation, as described in [117], is to decouple the translation and the rotation estimation. The translation is estimated by localizing the object's center in the image and predicting its distance from the camera. After that, the rotation is estimated by regressing to a quaternion representation. A 6DoF Object detection system with two stages is also proposed in [103]. A single Shot Multibox Detector (SSD) [62] extracts the object bounding boxes, and

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

(**c**) (**d**)

Figure 3.15: Transitioning from 2D video to the 3D virtual object: (**a**) 2D video. (**b**) Object replacement after detection and localization. (**c**) Object rotating in front of the user's viewpoint. (**d**) Digital information contextualized with the vehicle model. The corresponding videos can be found here.

an Augmented AutoEncoder (AAE) estimates the object rotation. Like the previous approach, DCS-PoseNet [119] uses a two-step process to estimate 6DoF from 2D object bounding boxes. First, the framework segments the object from the cropped image, then predicts 6DoF pose using DSC-PoseNet, which employs a differential renderer.

Some solutions try to regress rotation and translation simultaneously. For example, 6D-VNet [116] uses an end-to-end deep learning network to estimate the 6DoF pose of vehicles. The network extends the Mask R-CNN object detector, takes its intermediate outputs, and further regresses for rotation and translation of the object in 3D space.

Other approaches instead try to solve a Perspective-n-Point problem [63]. For example, the pose estimation method Pix2Pose [85] proposes a deep learning network to supplement a 2D detection pipeline to enable pose estimation. It regresses pixel-wise 3D coordinates from images using texture-less 3D models. The pixel-wise prediction is used to form 2D-3D correspondences. Finally, the PnP algorithm can be applied.

In [14], the authors propose an extension of the EfficientDet architecture [104] used for 2D object detection to predict the rotation and the translation of the object in the 3D space.

Most current works describe the problem statement and solution for regular RGB images. The application of the algorithms of these works to equirectangular images is tricky. The main reason is that equirectangular images present severe distortion, and there is a lack of training data related to these images. To the best of our knowledge, some works try to perform 2D object detection in equirectangular images [118, 123], but none performs an estimate of the 6DoF pose.

The pipeline presented in this thesis solves the problem of 6DoF pose estimation for objects in equirectangular images. Additionally, while other methods primarily rely on deep learning models to perform the task, the proposed one uses deep learning only for segmentation, which is

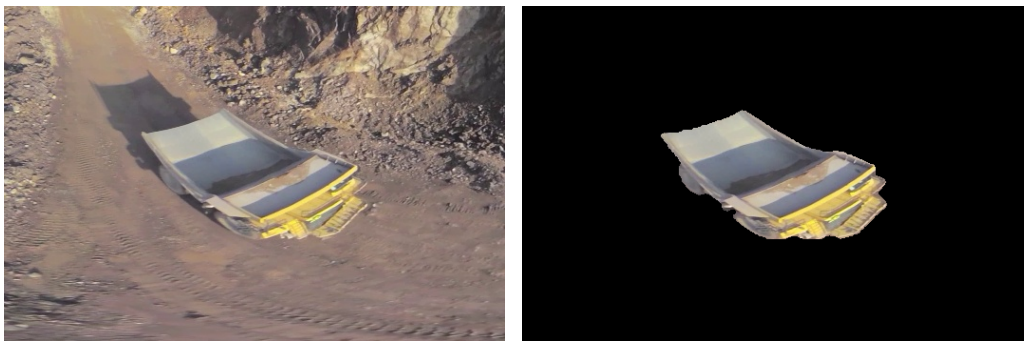
the first step. It then uses an optimization technique for pose estimation that does not require a trained network and can be applied to any object without effort. Indeed, a benefit of the proposed method is that there is no need to create a training dataset for the pose estimation, avoiding a task that can be quite time-consuming and difficult in terms of the acquisition of the ground truth pose, scalability, and full coverage of possible poses [91].

3.3.2 Algorithm description

The developed algorithm can be subdivided into two main steps:

- vehicle segmentation from the equirectangular image;
- vehicle pose detection with respect to the camera reference frame (6DoF).

A convolutional model for real-time instance segmentation, Yolact++ [13], was used to accomplish the first task. Yolact++ proved to be accurate for the segmentation of trucks, also in equirectangular images, in which the distortion is significant. However, Yolact++ was trained with 500 equirectangular images to make the vehicle segmentation more robust by manually labeling trucks in frames taken from 360° videos of open-pit mining operations. Figure 3.16 shows the result of the trained Yolact++ model.



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

Figure 3.16: The result of the trained convolutional model Yolact++ for an equirectangular image of a truck. (**a**) An example of an input image showing a truck in a mining environment. (**b**) Result of the segmentation in which the truck is correctly segmented.

The only requirement for the vehicle pose detection is to have an accurate CAD model even without textures of the item whose pose must be estimated. For the tests described in this section, the CAD model of a Komatsu HD785 truck was used inside Unity. The same 360° camera simulator implemented in Section 3.2 was used to capture equirectangular images of the CAD model. Using the output given by Yolact++, the vehicle pose detection was performed. The algorithm developed to accomplish this task is schematized in Figure 3.17.

The data exchange between Unity and Matlab was activated via a connection based on the ZMQ protocol, following the same approach as in Section 3.2. First the truck was randomly placed inside Unity. Then a picture of the scene was taken using the 360° simulated camera. This picture is called synthetic image. In Matlab, this synthetic image and the output of Yolact++, i.e., the segmented image of the real world, are compared for similarity with a score. A Particle Swarm (PS) optimizer is responsible for finding the optimal solution. At each new iteration, the

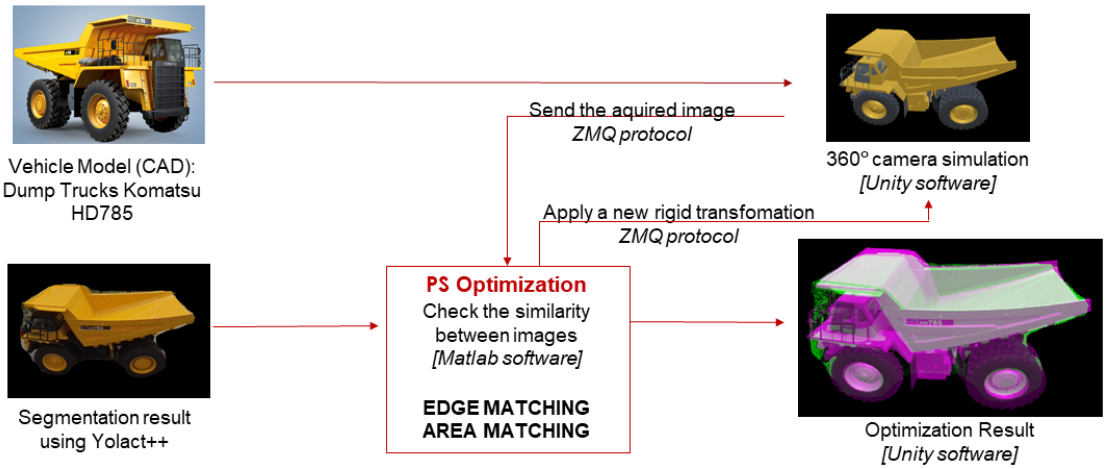


Figure 3.17: Scheme of the pose detection algorithm.

algorithm sends a new pose to Unity. The 360° simulated camera takes a new picture, and the previous steps are repeated until convergence or the maximum number of iterations are reached.

Hereafter, the expression of the Cost Function (CF):

$$CF = S_E + S_A + S_C + S_V, \tag{3.14}$$

where its terms depend on:

- edges (S_E);
- area (S_A);
- difference in the centroids of the edges (S_C);
- difference in the eigen vectors of the edges (S_V);

The following subsections explain the various term of the cost function in detail. The real-world and synthetic images in Fig. 3.18 are taken as an example to show the computations made for the different terms.

Edges

The first term of the cost function is relative to edges. The "Canny" algorithm [26] was used to compute the images of the edges of the real-world (E_r) and synthetic image (E_s). Since a perfect correspondence between the CAD model and the actual vehicle is impossible, the edges of E_s were smoothed by applying a Gaussian filter with a standard deviation of 0.5. This last image was called E_{sg}. The two images are then multiplied pixel by pixel, computing E_m as:

$$E_m = E_r \cdot E_{sg}. \tag{3.15}$$

Fig. 3.19 shows the images involved in the computation.

The score term is computed as follows:

$$S_E = 1.0 - \frac{n_m}{n_s}, \tag{3.16}$$

where n_m and n_s are the number of pixels of E_m and E_s that are greater than zero.

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

Figure 3.18: Real-world and synthetic images are examples to illustrate the different terms of the cost function: (**a**) real-world image and (**b**) synthetic image.

Areas

The corresponding binary images (BW_r and BW_s) are computed from the real and synthetic worlds. A_s is the name of the area of BW_s, which is the number of pixels whose value is greater than 0. A dilated version of BW_s, called BW_{sd}, is also computed using a disk with a diameter of 7 pixels as the morphological structuring element. Let's indicate with BW_d the difference between BW_{sd} and BW_s:

$$BW_d = BW_{sd} - BW_s. \tag{3.17}$$

A_d is the area of BW_d.

Now, it is possible to compute the images M_a and M_d, i.e. the result of the pixel-wise multiplication between BW_r and BW_s, and between BW_r and BW_d:

$$M_{rs} = BW_r \cdot BW_s, \tag{3.18}$$

$$M_{rd} = BW_r \cdot BW_d. \tag{3.19}$$

Figure 3.20 shows the images involved in the computations.

The score relative to the areas is computed with the following equation:

$$S_A = 1.0 - A_{rs}/A_s + A_{rd}/A_d, \tag{3.20}$$

where A_{rs} and A_{rd} are the corresponding areas of M_{rs} and M_{rd}.

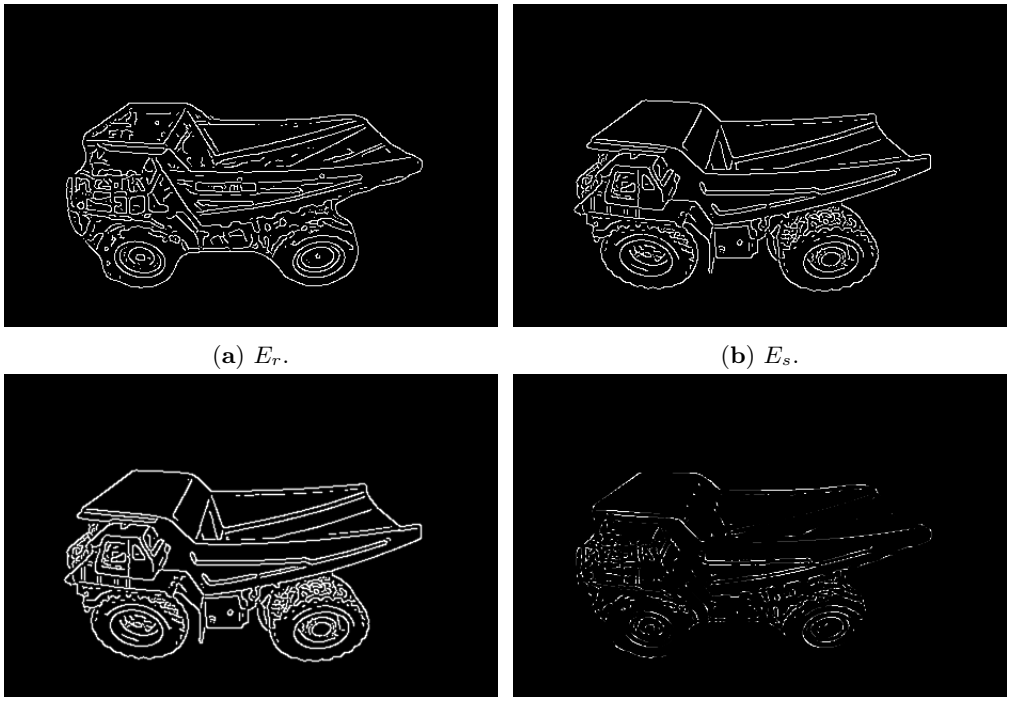
Difference in the centroids of the edges

This part of the cost function is in charge of computing the difference between the centroids of the images E_r and E_m. The formula to compute the centroids of the images is the following:

$$x_c = \frac{\sum_{i=1}^{N} I(x_i, y_i) \cdot x_i}{\sum_{i=1}^{N} I(x_i, y_i)}, \tag{3.21}$$

$$y_c = \frac{\sum_{i=1}^{N} I(x_i, y_i) \cdot y_i}{\sum_{i=1}^{N} I(x_i, y_i)}, \tag{3.22}$$

where (x_c, y_c) are the coordinates of the centroid of the image I, N is the number of pixels whose value is greater than 0, (x_i, y_i) are the general coordinates of the pixel i, and $I(x_i, y_i)$ is the grey value of the pixel in position (x_i, y_i).



(**a**) E_r. (**b**) E_s.

(**c**) E_{sg}. (**d**) E_m.

Figure 3.19: The images involved in the computation of the cost function term relative to edges. (**a**) E_r, edges of the real-world image. (**b**) E_s, edges of the synthetic image. (**c**) E_{sg}, edges of the synthetic image after the Gaussian filter. (**d**) E_m, pixel-wise multiplication between E_r and E_{sg}.

The cost function term is computed as:

$$S_C = \frac{\sqrt{(x_{cr} - x_{cm})^2 + (y_{cr} - y_{cm})^2}}{\sqrt{R^2 + C^2}}, \tag{3.23}$$

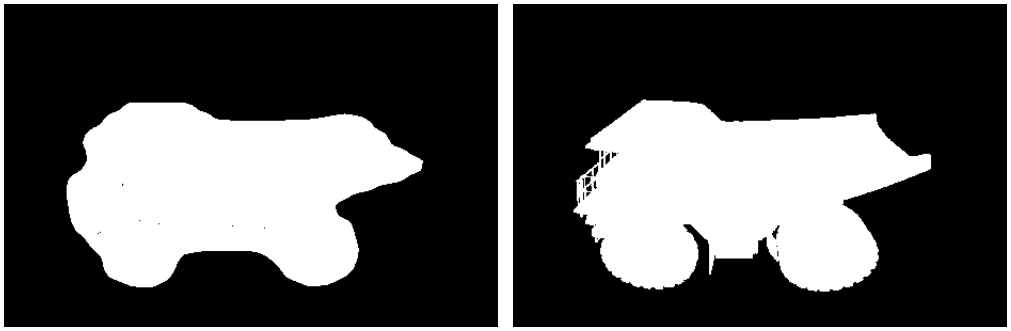
where (x_{cr}, y_{cr}) and (x_{cm}, y_{cm}) are the coordinates of the centroids of E_r and E_m, and R and C are the number of rows and columns of E_r.

Difference in the eigen vectors of the edges

The last term of the cost function can be explained as a constraint for the edge matching to be uniform on all the parts of the edge images, i.e. E_r and E_{sg}. To reach this aim, the image coordinates (x_i, y_i) of the pixels whose value is greater than 0 is arranged in a matrix of dimension $N \times 2$, where N is the number of pixels whose value is greater than 0. Then the covariance matrices C_r and C_m of this matrix are computed for E_r and E_m. The eigenvectors are computed for both once C_r and C_m are obtained. Let's call $\vec{v_r}$ and $\vec{v_m}$ the two eigenvectors that corresponds to the highest eigenvalue for C_r and C_m, Figure 3.21.

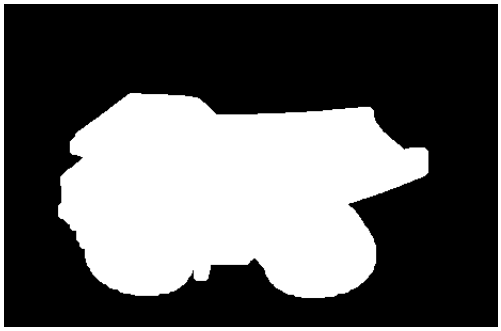
The cost function term is the dot product between $\vec{v_r}$ and $\vec{v_m}$:

$$S_V = 1.0 - \vec{v_r} \cdot \vec{v_m}. \tag{3.24}$$

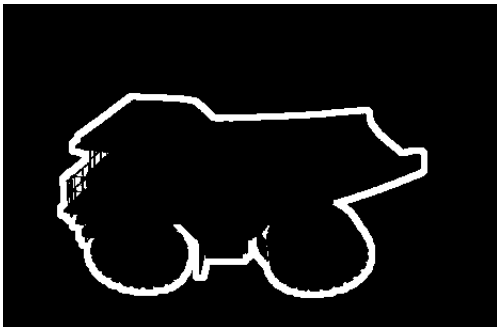


(**a**) BW_r.

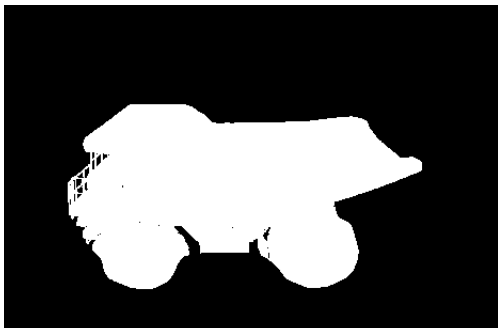
(**b**) BW_s.



(**c**) BW_{sd}.



(**d**) BW_d.



(**e**) M_{rs}.

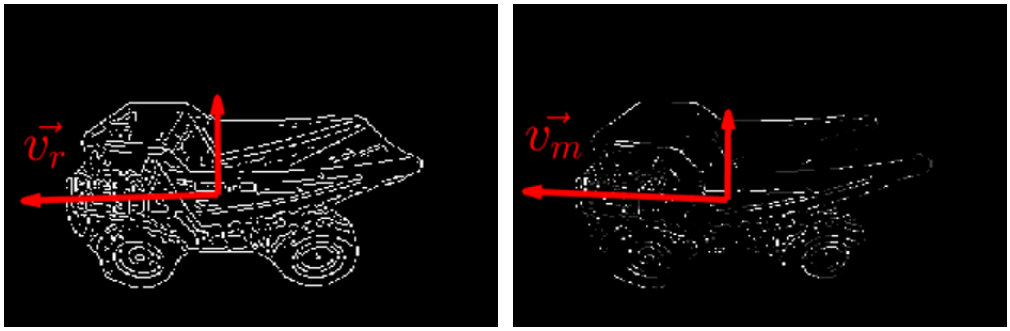
(**f**) M_{rd}.

Figure 3.20: The images involved in the computation of the cost function term relative to the areas. (**a**) BW_r, a binary image of the real-world image. (**b**) BW_s, a binary image of the synthetic image. (**c**) BW_{sd}, synthetic binary image dilated. (**d**) BW_d, result of the subtraction between BW_{sd} and BW_s. (**e**) M_{rs}, result of the multiplication between BW_r and BW_s. (**f**) M_{rd}, result of the multiplication between BW_r and BW_d.

3.3.3 Results

Figure 3.22 shows the experimental setup to test the developed algorithm. A 360° camera, such as the Insta360 ONE X, is placed on a rotary stage, which is placed on a translation stage. The camera frames a miniature model of a Komatsu HD785 truck.

Ten images were acquired by translating the translation stage of 8 cm each new acquisition, and eleven images by rotating the camera of 5° each new acquisition, Figure 3.23.



(**a**) Eigenvector representation of E_r. (**b**) Eigenvector representation of E_m.

Figure 3.21: E_m and E_r with their respective eigenvectors centered in the centroids of the two images. (**a**) E_r and its eigenvectors. (**b**) E_m and its eigenvectors.

Figure 3.22: Experimental setup to test the developed algorithm. A 360° camera is placed on a rotary and a translation stage. The camera frames the miniature model of a truck.

Concerning the parameters used for the PS optimization, the swarm size was set to 150 and the maximum number of iterations to 75. The research range was set to ±20° for rotations and to ±20 cm for translations. The initial pose conditions were set randomly from nominal values within the imposed research ranges. The algorithm ran on an Intel(R) Core(TM) i7-9700KF CPU. The mean computational time to find the optimum was about 20 min.

Table 3.2 and Figure 3.24 show the results obtained for the imposed rotations.

Table 3.3 and Figure 3.25 show the results obtained for the imposed translations.

Figure 3.26 shows an example of the results obtained; in this case the camera was rotated by 15° with respect to the initial orientation.

3.3.4 Discussion

Results show that the developed algorithm achieved good results for both translations and rotations. In particular, the maximum difference in the rotation estimation was 3.2° for the nominal

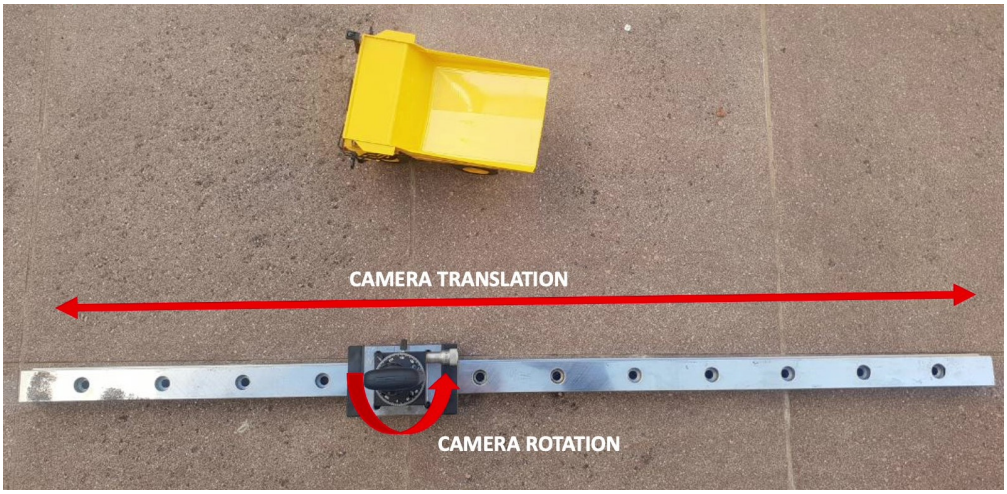


Figure 3.23: Scheme of the rotations and translations imposed to the camera.

Nominal angle [°]	**Measured angle** [°]	**Difference** [°]
5.0	6.4	1.4
10.0	10.1	0.1
15.0	17.3	2.3
20.0	19.7	-0.3
25.0	26.3	1.3
30.0	31.0	1.0
35.0	37.0	2.0
40.0	41.9	1.9
45.0	48.2	3.2
50.0	52.1	2.1

Table 3.2: Results obtained by the algorithm applying a rotation of 5° at each step.

Nominal translation [cm]	**Measured translation** [cm]	**Difference** [cm]
8.0	8.4	0.4
16.0	16.0	0.0
24.0	24.9	0.9
32.0	32.9	0.9
40.0	40.8	0.8
48.0	49.0	1.0
56.0	57.1	1.1
64.0	67.9	3.9
72.0	76.0	4.0

Table 3.3: Results obtained by the algorithm applying a translation of 8 cm at each step.

rotation of 45°, Table 3.2. Figure 3.27 shows the optimization result for this case.

The mean difference for the rotation is 1.5°, while the standard deviation is 1.0°.

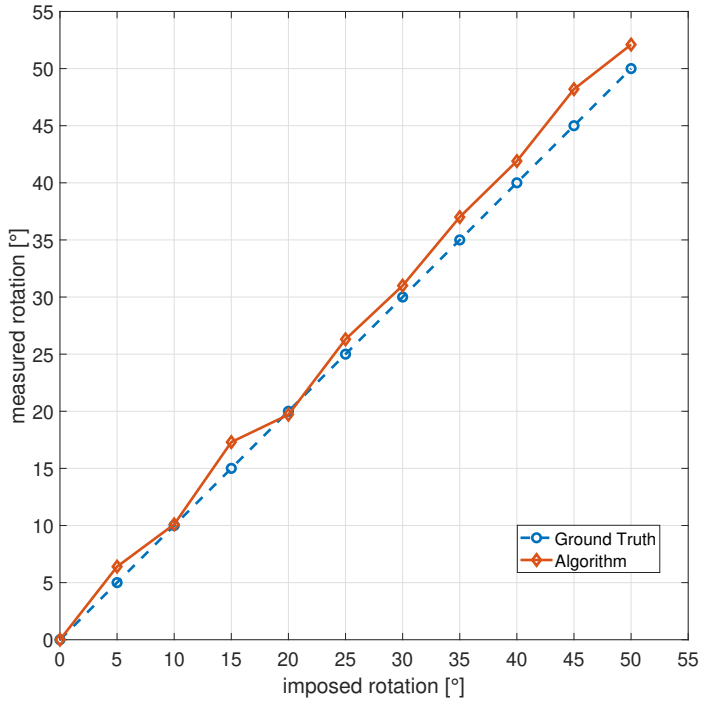


Figure 3.24: Comparison of the imposed rotations with the measured ones.

The maximum difference in the translation estimation was instead 4 cm for the nominal translation of 72 cm, Table 3.3. As shown in Figure 3.25, the difference increases at the increase of the translation amount. Probably, looking at Figure 3.28, this is due to how the vehicle appears in the equirectangular image. In this case, the vehicle appears quite far, and small translations cannot be appreciated from the image point of view. Indeed, in this case, at least visually, the difference between the real world and the synthetic image does not seem relevant, Figure 3.28.

The mean difference for the translation is 1.4 cm, while the standard deviation is 1.5 cm. The computational time of 20 min makes the proposed algorithm applicable only offline. However, in the case of a video, once the pose is estimated in the first frame, the search field for the next frame is minimal because the vehicle will be in a pose very near to one of the previous frames. It will speed up the elaboration and pose detection.

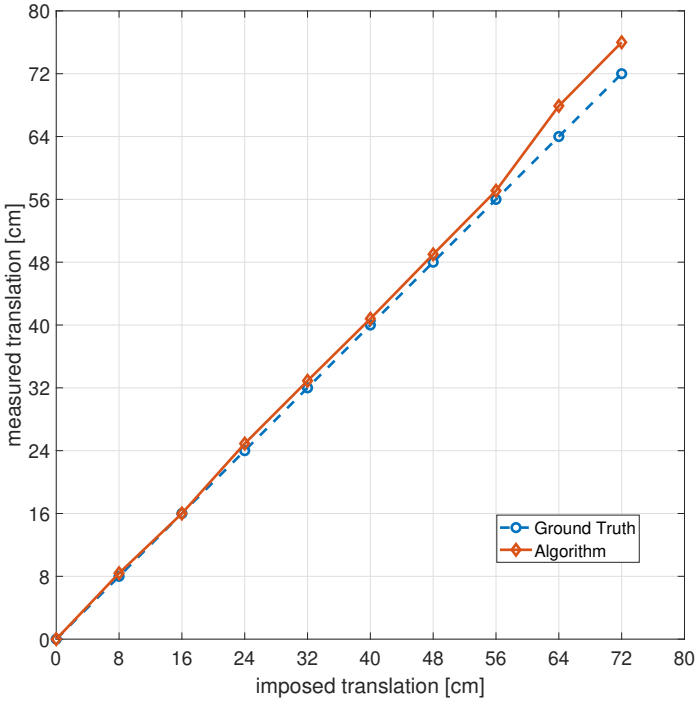
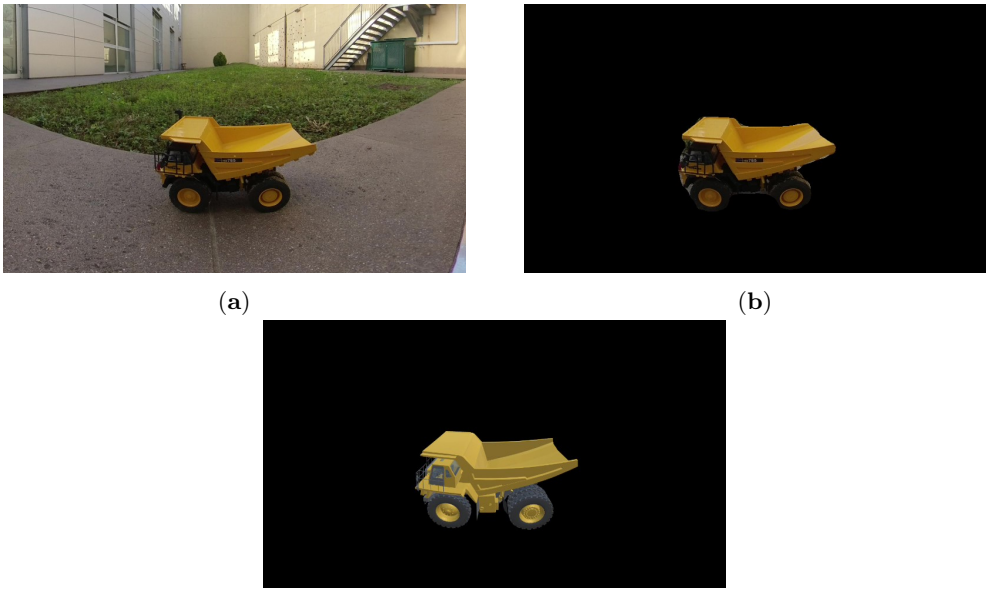


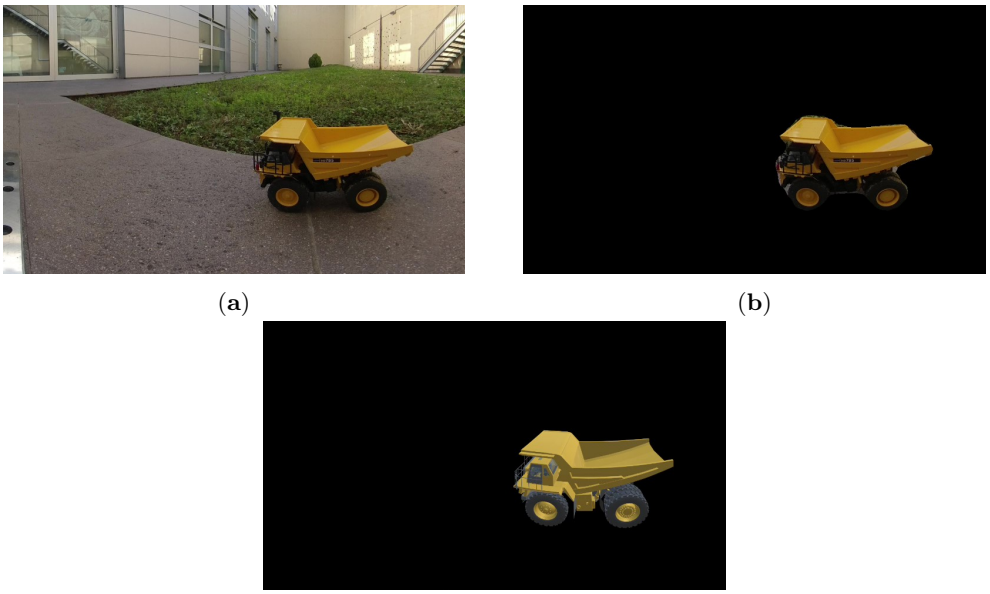
Figure 3.25: Comparison of the imposed translations with the measured ones.



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

(**c**)

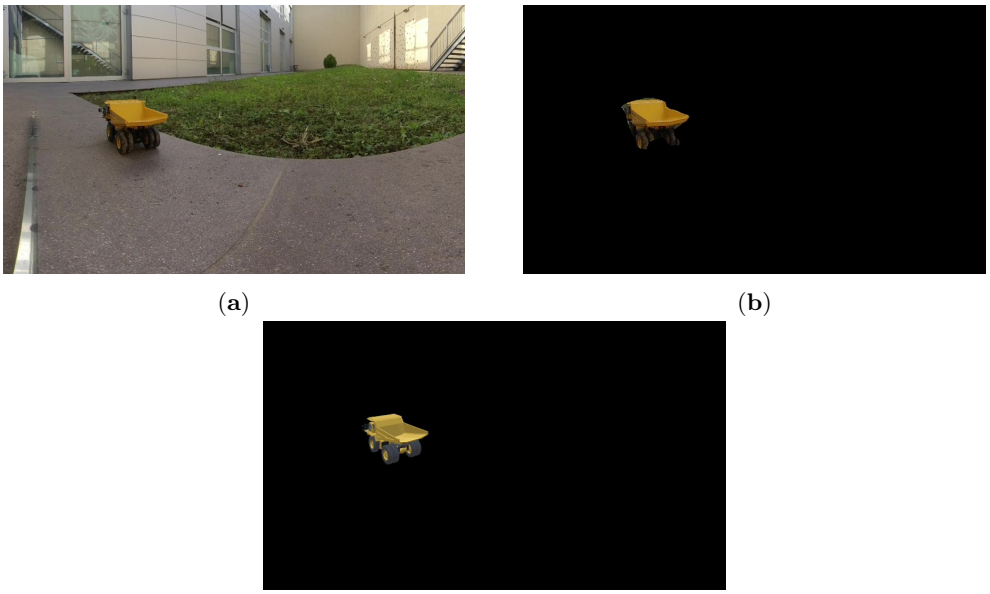
Figure 3.26: An example of the optimization result where the camera was rotated by 15°. (**a**) A portion of the input equirectangular image taken by the 360° camera. (**b**) Result of the segmentation. (**c**) Optimization result in which the CAD is rendered in the final pose found by the optimization algorithm.



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

(**c**)

Figure 3.27: Optimization result where the camera was rotated by 45°. (**a**) Input equirectangular image taken by the 360° camera. (**b**) Result of the segmentation. (**c**) Optimization result of CAD model.



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

(**c**)

Figure 3.28: Optimization result where the camera was translated by 72 cm. (**a**) Input equirectangular image taken by the 360° camera. (**b**) Result of the segmentation. (**c**) Optimization result of CAD model.

Chapter 4

AR in industry

In the last decades, technological growth in industry has mainly focused on automation rather than its fusion with human work. It has generated a clash between human and robot workers, where the latter, especially those carrying low-level tasks, are worried about being replaced by the former.

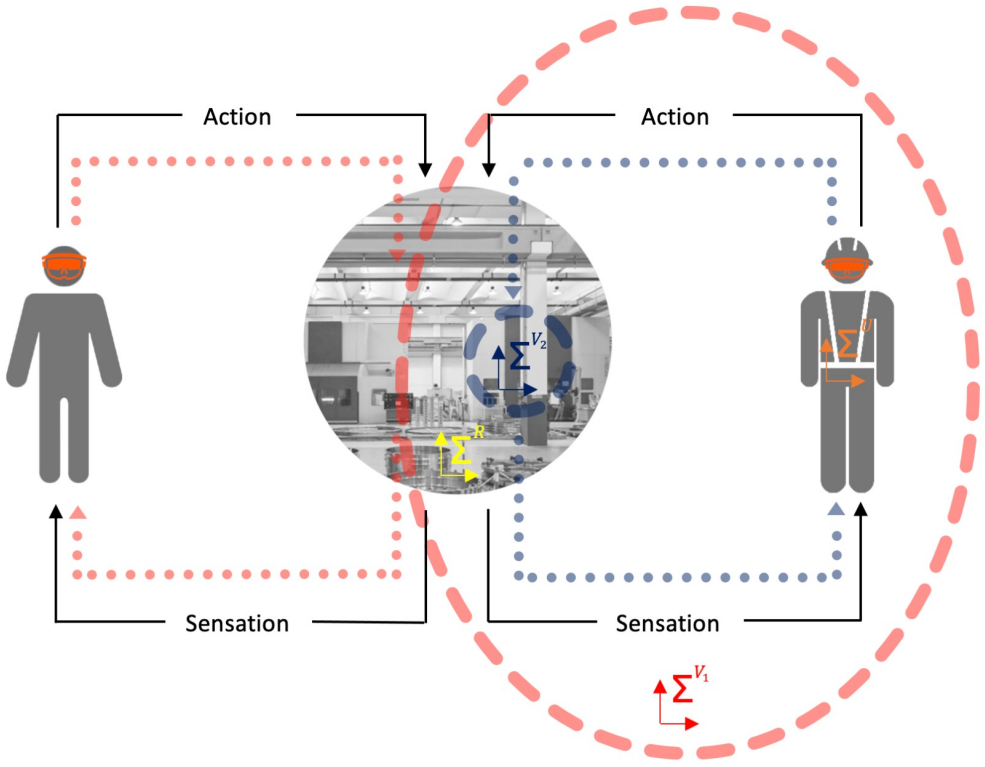


Figure 4.1: The third level of the perception-action loop between a supervisor and an operator in a shared augmented reality framework. The real environment R is represented in this context by, for example, a production environment.

Industry 4.0 aims to overcome this situation. Some innovative technologies, such as AR,

provided an alternative approach to the production environment and created an intelligent factory [82]. Many companies already use this technology because it improves process performance, enhances users' perception at all levels of PAL (Figure 4.1) and reduces costs in different business areas.

An example in the automotive industry is Volkswagen group, which uses an AR system to help employees navigate its huge factories for maintenance, inventory, inspections, and other activities. Another industrial example is the aircraft manufacturer Boeing which managed to reduce the assembly time for the wiring of its planes. In this case, technicians use voice commands, keep their hands free, and get help from a remote expert who sees exactly what they see.

The operator can return to compete or collaborate with robots thanks to AR technology. The operator will not be replaced but will receive a useful tool to increase his senses and cognitive abilities. It leads to lower staff costs by reducing the time of execution, the number of recurring errors, or accidental damages of the components due to incorrect execution of procedures.

Many articles in the literature show the advantages of using AR instructions in improving working procedures [105, 114]. In addition, by increasing speed and accuracy, AR reduces mental effort [41].

4.1 Smart Gate

At MiroLab, an interactive AR demo for industrial setting on loaded pallet shape measurement and checking by three simulated ToF cameras was developed and presented to IEEE MetroXRAINE 2022.

After the pallet enters the scanning area, the user wearing a HoloLens 2 headset, through the visualization of point clouds from the simulated cameras, can understand how the system works and is shown of the load size with respect to the pallet, highlighting the out-of-shape areas. The user can then move the above box, which has physical properties, to another position with his hands, and through a smartphone interface, a new scan will be launched, and the dimensions of the load will be checked in this new configuration. An initial calibration using a 2D Vuforia marker is necessary to save the reference system to fix the demo's position.

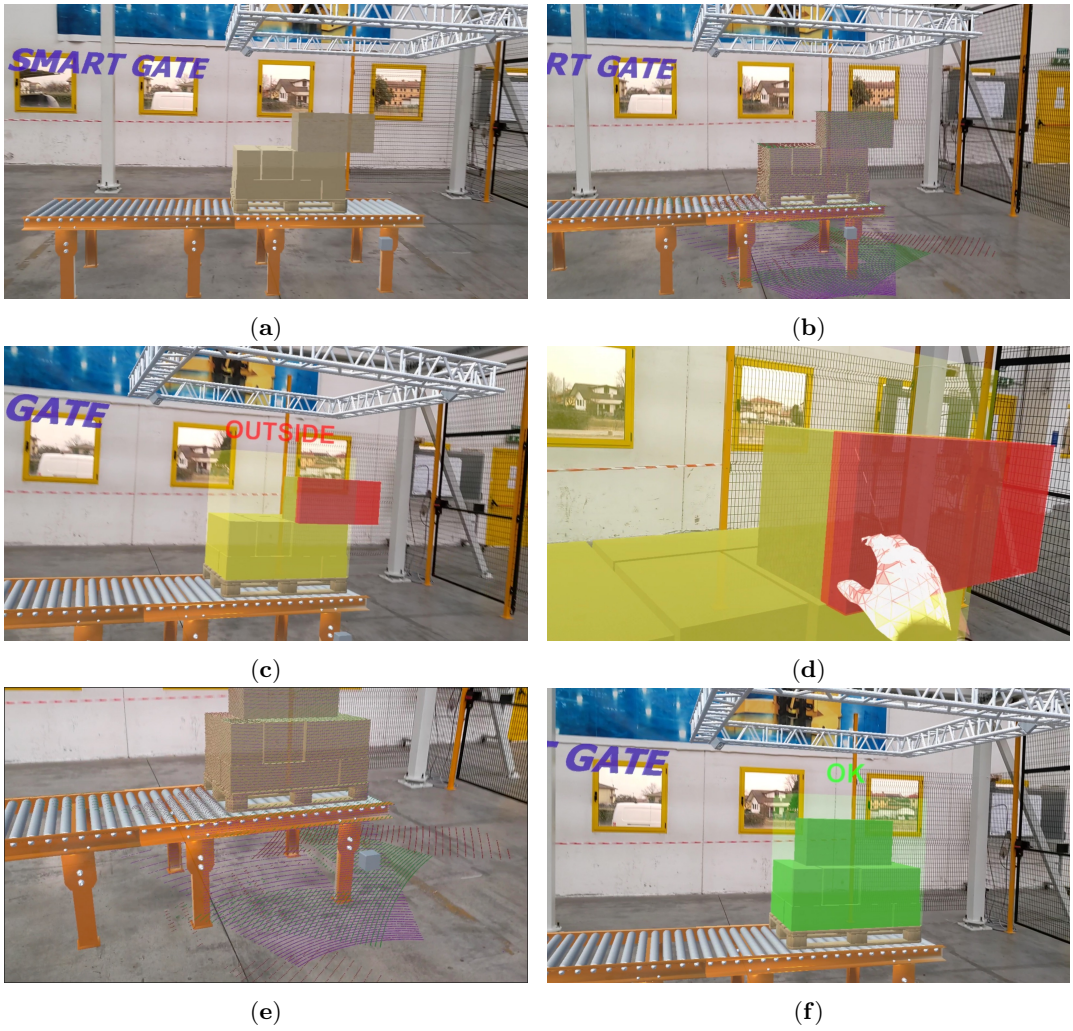
The immersive experience of the demo is further enhanced by the high-level functions integrated into the virtual environment. These functions include collision detection and physical properties such as gravity, 3D sound rendering for realistic audio feedback, and network capabilities for smartphone interaction within the same local area network through a client-server protocol such as MQTT. Collision detection and physics allow users to interact with virtual objects realistically, contributing to a more immersive experience.

In addition, this demo is a valuable tool for training and education. The possibility of visualizing the points intersecting with the load from the simulated cameras provides a better understanding of how the loaded pallet shape measurement and checking works. Furthermore, incorporating advanced functionalities, such as physical object manipulation in a virtual environment with physical properties, enhances overall immersion and increases engagement.

This demo was designed in AR rather than other MR spectrum technologies, such as VR because AR offers the user a higher level of proprioception while maintaining a connection with the surrounding environment. In fact, AR allows visualizing the best 3D camera poses in terms of load reconstruction directly in the area where the real scanning system will be installed.

4.1.1 Demo process

This demo presents a virtual replica of a measuring system able to reconstruct and check the shape of loaded pallets through three simulated ToF cameras. The proposed work offers an



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Figure 4.2: Frames captured during the demo session using HoloLens.

engaging AR equipped with high-level functions to provide an immersive experience for the user wearing the HoloLens headset. The system leverages simulated camera point clouds, allowing users to obtain measurements of load dimensions relative to the pallet.

The demo begins when the pallet enters the designated scanning area, Fig. 4.2a. Simulated cameras acquire point clouds by intersecting objects using the raycasting method, Fig. 4.2b. An algorithm then evaluates whether there are points outside the pallet shape. The results of the load size analysis are shown in AR to the user by highlighting any out-of-shape areas, Figure 4.2c.

The demo incorporates physical object manipulation through hand gestures using the HoloLens' three-dimensional hand-tracking feature to offer users greater flexibility, Figure 4.2d. This interactive process enhances user engagement, and the final visualization allows an understanding of pallet dimensions.

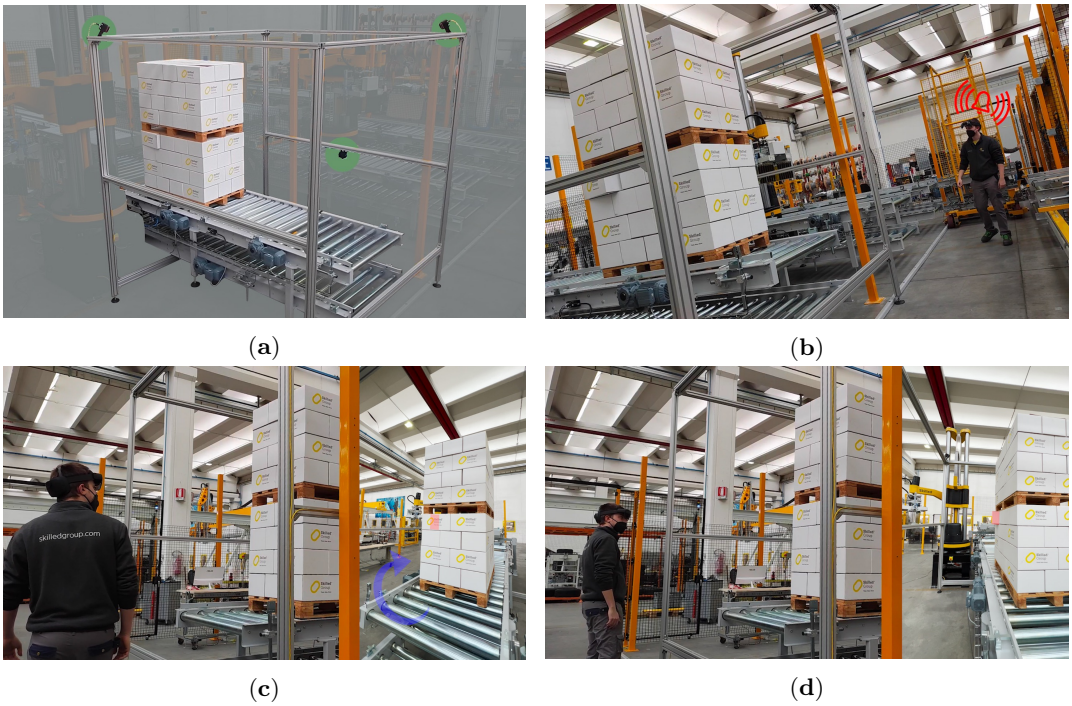
Users can position a box above the load to another location. A new scan is initiated in the updated configuration through a smartphone-based control interface developed using Node-RED,

Figure 4.2e. Consequently, the system rechecks the measurement of load size with respect to the pallet, Fig. 4.2f.

Initial system calibration is essential to place the measurement system in the desired position. This calibration process is performed using a Vuforia 2D marker, which establishes a reference system with respect to the physical environment against which it is possible to move using the developed smartphone interface. When finished, a World Anchor is used to fix the reference system of the virtual scene in the real world.

4.1.2 Industrial application

A real application follows this demo in industrial and logistics settings, where in this case, only the results of the scanning area are shown to operators in AR, Figure 4.3a. If some boxes are outside the pallet shape, the operator is alerted, Figure 4.3b, and guided to the desired location through AR cues, Figure 4.3c and Figure 4.3d.



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

(**c**) (**d**)

Figure 4.3: Real industrial application.

4.2 Grinding in aviation

An additional developed AR application in the industrial sector involves aiding operators during the industrial grinding process to respect working tolerances with repeatability comparable to an automatic system. For such a purpose, an AR headset was used to support the operator in respecting the desired working parameters. The proposed framework was developed and tested in collaboration with Trentino-based company Fly SpA company, which operates in the aeronautics and aerospace industry for Rolls-Royce company. The grinding process involves titanium welded

components, specifically the root welding of titanium alloy for Boeing aircraft engine turbine blades casing, Figure 4.4 .

Figure 4.4: Example Boeing aircraft engine turbine blades casing. © Fly SpA website

The proposed system consists of two main elements: a hardware and software infrastructure for data acquisition and a visual interface for HoloLens.

4.2.1 Acquisition System

An acquisition system was designed to acquire the desired working parameters, Figure 4.5. It had to replicate the real grinding process as much as possible on the welded turbine blades simulating it on welded samples.

The selected process parameters are the tool's possible inclinations: the feed rate, the tool pressure (which controls the cutting speed), and the load applied to the tool. To measure the tool inclination, elevation, azimuth, and feed rate, an HTC Vive Tracker was used, Figure 4.6.

A motion-tracking accessory was attached to the tool and tracked in 3D through the HTC Lighthouse system. The tool's vertical load was measured using a load cell placed under the sample and fixed with a vise. A National instruments board, NI USB-6210, was taken to read the load cell. The frequency used to sample these parameters is 47 Hz. A further parameter sent to the HoloLens during the process was the temperature reached by the sample. This information was obtained using a FLIR A615 thermal camera placed nearby the working area with a sampling rate of 50 Hz. Figure 4.7 reports the schematic structure of the measurement infrastructure and the information that the operator can receive through HoloLens. The software acquisition interface was designed in the Qt framework. Communication with the laptop interface and HoloLens was structured using ZeroMQ and MQTT libraries, respectively.

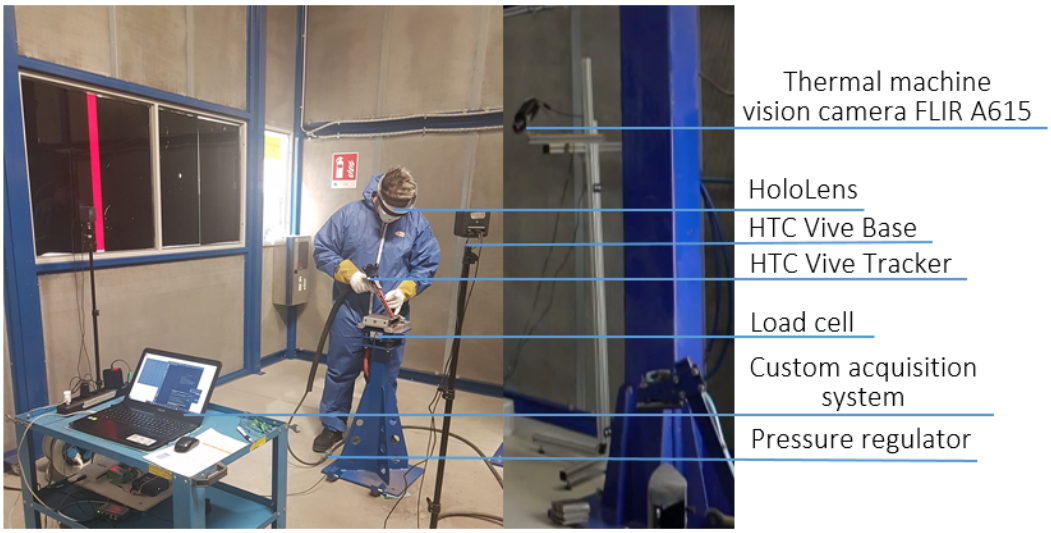


Figure 4.5: A picture of the experimental acquisition system. Here are highlighted the exploited sensors and interfaces.

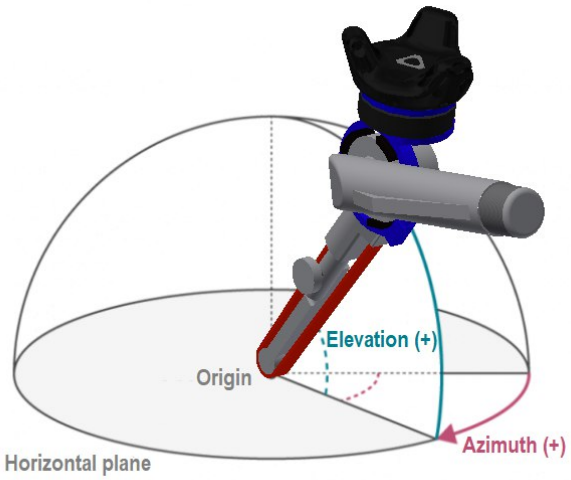


Figure 4.6: The spherical geometry considered for the characterization of the tool position. Both elevation and azimuth were computed from the quaternion measured from HTV Vive Tracker.

4.2.2 AR Interface

For the human being, the understanding of the behaviour of a machine can be learned through adequate interfaces and the repetitive use of the system. In order to find the best AR interface and thus maximize the potential of this technology, a test campaign was run with five grinding operators who have yet to experience such a technology. Four user interfaces, shown in Figure 4.8, were designed with the Unity platform. The operator can see through all of them, where the background is black.

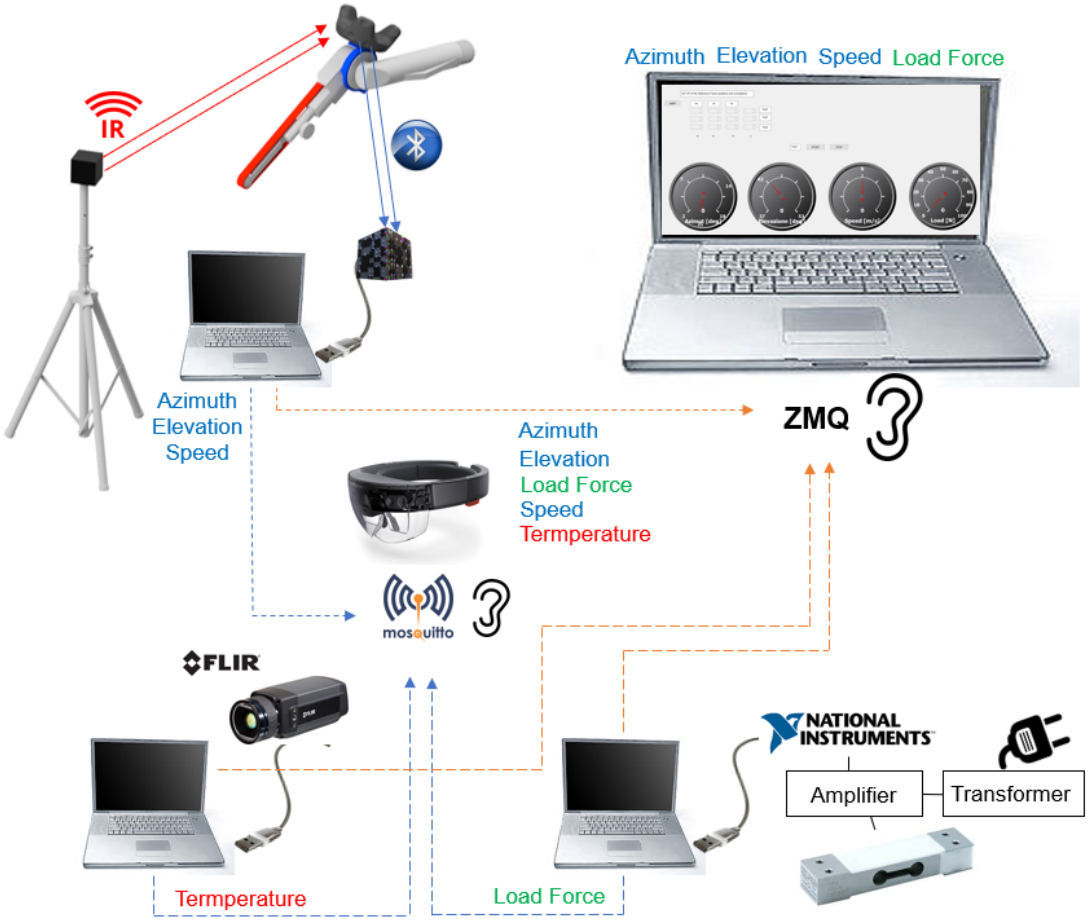


Figure 4.7: The schematic representation of the connections among the elements in the system. For each connection is reported the transmitted data type.

The first interface has a digital control for angles (arrows) and force (thumb emoticon). The second one has an analogue control, in dimensions and transparency, for angles (external box) and force (thumb emoticon). The third has an analogue control, in dimensions and transparency, for force (arrows) and the analogue position for the angles (2d ball). The last one has an analogue control, in transparency and dimensions, for angles (arrows) and force (circle emoticon).

Different interfaces were tested on different days by each operator. In this way, they could not influence each other or use the experience learned with the previous interface. After initial training on how the system works, ten tests were performed for each operator.

In order to evaluate the performance of each operator with the different interfaces, a statistical study was achieved by analyzing the parameters' standard deviations at each test. Furthermore, for each interface, at the end of testing session, each operator reported his opinion through a questionnaire in terms of:

- Understanding of the system
- Usability

- The mental activity required

- Further implementations, comments or observations

Based on the statistical analysis of the results and the judgment expressed by the operators, the best interface was identified as the fourth interface.

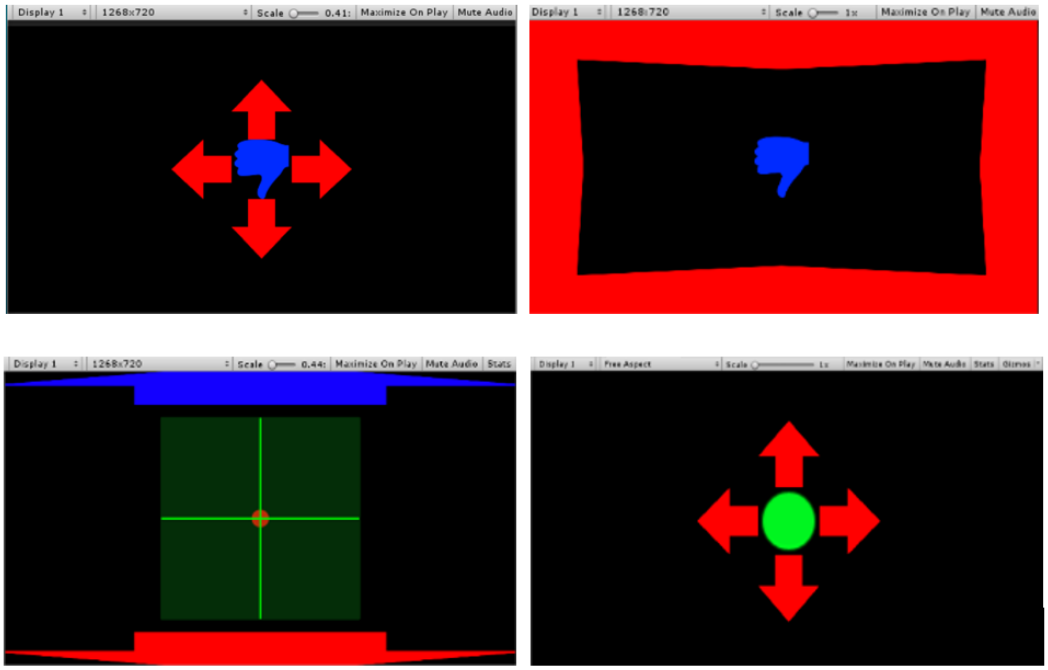


Figure 4.8: The four HoloLens interfaces describe with animations the tool's inclinations (TI) and its vertical force (F). The top left one describes through a digital control TI with arrows and F with an inch. The top right instead defines through an analogue control, in transparency and dimensions, TI with an external box and F with an inch. The down left shows TI with the centring of a bubble and F with analogue arrows. Through an analog control, the last one characterizes TI with arrows and F with a coloured circle.

An audio system has been implemented to limit the overload of the AR interface from a visual point of view. By using a different sensory input channel, the operator can respect a process parameter, such as the tool's feed speed, with good results.

Initially, the operator would accelerate or decelerate based on what appeared on the smart glasses, so he could not respect a constant speed. After some training and with audio, having a fixed allocated time for completing the process, the operator could finally maintain a constant speed given the short sample length, 150 mm.

During the process, the temperature is monitored in real-time, and only if the threshold is reached the video of the thermal heating is shown to the operator superimposed to the sample, Figure 4.9. The transparency level of this image can be adjusted thanks to vocal commands, while its positioning is identified using the Vuforia Engine package in the virtual Unity environment.

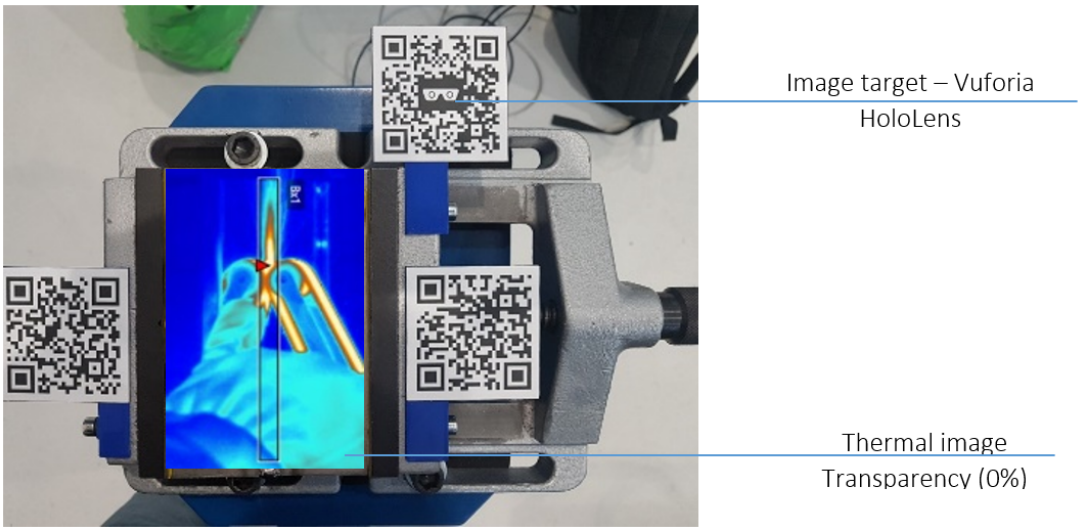


Figure 4.9: AR implementation for thermal heat: in the picture are visible the Vuforia markers exploited for generating the AR image, the blue one superposed to the specimen, here colours were mapped to specified heat levels.

4.2.3 Results

The best interface found was exploited in an experimental campaign design to study the effect of the technology on the grinding process. The worst operative case was considered in the test: high azimuth and elevation angles, high vertical force and high feed rate. For high angles without AR assistance, the operator did not have physical references on keeping those parameters constant: holding the tool parallel to the sample, which corresponds to a low elevation angle, would be easier. The parameter values selected for the test were 16° for the azimuth, 45° for the elevation and 30 N for the vertical force.

The same test was run with and without the HoloLens. At the beginning of both tests, the operator stated that the tool was in the correct configuration. As for the force parameter, it was not possible to initially help the operator because he switched on the tool when it was not yet in contact with the working piece; otherwise, the rotation of the abrasive would leave marks on the sample. Then, he was told to keep the initial parameters as constant as possible and to finish the test in seven seconds.

The charts in Figure 4.10 show process parameters results achieved with and without HoloLens, respectively. When the operator starts without HoloLens, guiding him back inside the correct ranges is no longer possible if he does not respect them. This case can be seen in Figure 4.10 (red line), where the error is not corrected. In the opposite case, Figure 4.10 (blue line), the HoloLens helped the operator and corrected his behaviour.

In all plots, a 2nd-order lowpass Butterworth filter is applied with a cut-off frequency of 1 Hz to minimize the effects of vibration from the grinding operation.

Force trends in Figure 4.10c do not start from zero because the operator used a finger as a guide on the sample from the beginning of the test.

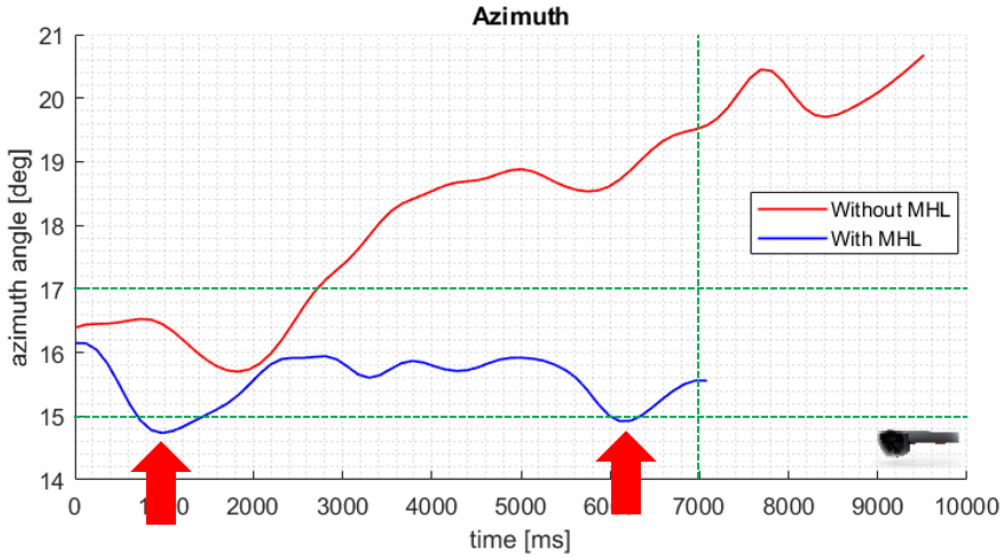
The choice to use ranges instead of continuous corrections was made to avoid mentally overloading the operator, which could compromise and damage the quality of his work.

When trying to reduce the ranges, the operator's performance arrived at a certain threshold

and did not improve because he could no longer follow the corrections suggested by the system. The ranges selected for the tests with HoloLens were $\pm 1°$ for the angles, $\pm 5\,\mathrm{N}$ for the vertical force and $\pm 0.5\,\mathrm{s}$ for the time taken to perform the test.

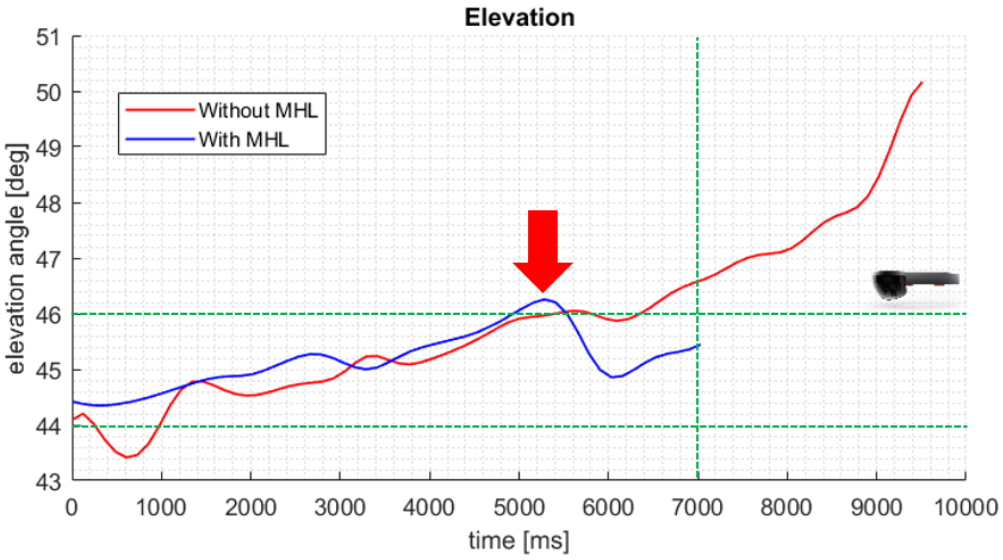
The material of the sample was titanium. It could have a problem at 250 °C, so an alarm threshold was set at 200 °C for safety. However, such a value is never reached since the temperature never exceeds the 160 °C. If the threshold alarm is lowered to 150 °C when reached, the operator sees on the sample its thermal heat. In this case, the software then generates a new corrective parameters configuration on HoloLens to reduce the heating of the sample, such as applying a lower pressure.

The experimental evidence verified that human capability can be enhanced through AR technology. Companies that operate in the field of assembly, navigation, training and maintenance activities [98, 40] usually use AR, but this is an example of how the same technology can be applied to mechanical industrial operations supporting the operator's work directly online, resulting in improved quality. The experimental results showed how, thanks to the visual and audio support, the operator kept constant within the process tolerance limits and the parameters that affect the grinding operation for the whole working time. Also, the adjustable thresholds enabled more versatile management of the processes, resulting in a more controlled process outcome.

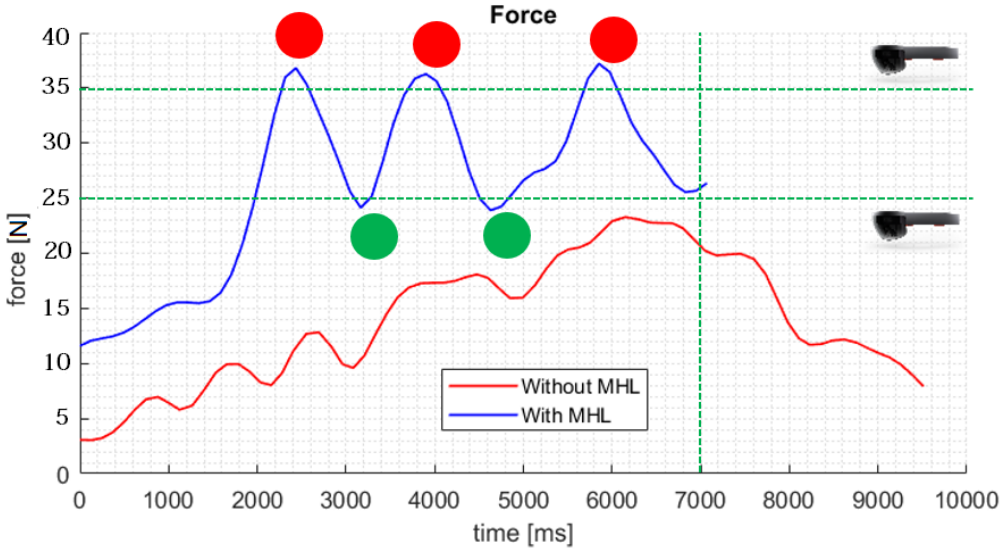


(**a**)

Figure 4.10: *Cont.*



(**b**)



(**c**)

Figure 4.10: Parameters acquisition with (blue) and without (red) the use of HoloLens (MHL). In green are the thresholds for the considered parameter in the test: (**a**) Azimuth, (**b**) Elevation and (**c**) Force parameter, respectively. The icon highlights the input provided to the worker through HoloLens.

Conclusions and future work

This dissertation, which focuses on applications and methods designed to restore human centrality and enhance their capabilities within the PAL, demonstrated the profound impact of human-centered technologies in reshaping the landscape of healthcare, education, and industry.

In particular, AR technologies combined with innovative measurement systems continue evolving and finding new healthcare applications. They have the potential to revolutionize medical practices, improve patient and therapist outcomes, and enhance the overall healthcare experience. This dissertation introduces several frameworks designed for this purpose. In particular, an innovative AR multidimensional framework was developed for the ADL scenario of setting up the table in a shared AR environment where both therapist and patient have access to the same space-aligned VR. The therapist enhances the assessment of a patient's daily life activity, and through their interaction, it was possible to increase patient engagement and therapist involvement. In this demo, all levels of PAL are enhanced in terms of both collaboration and supervision. The co-design of this prototype was realized in collaboration with clinical experts of the Villa Rosa Rehabilitation Hospital in Pergine Valsugana (Italy). The calibrated setup ensures an uncertainty in object localization of 5 mm with a confidence level of 95% and residual values due to estimated object rotations of less than 1°. The designed framework was evaluated with a user study involving patients and healthy testers. It allowed the selection of significant parameters, their acceptance thresholds, and the goodness of the proposed method. The proposed framework was developed for the specific ADL of setting the table. However, it can also be applied in other AAL scenarios for the metrological assessment of impaired or frail users and to optimize the living environment. It can be applied in OT to evaluate treatment/training effectiveness in the clinical setting objectively. In a future test campaign, the prototype will be used on other patients in parallel with their treatment and training to restore their autonomy with proper evaluation in the AUSILIA infrastructure.

In addition, MR is likely to become an increasingly valuable and integral component of the modern educational landscape because of more engaging learning. In order to make the experience more immersive and realistic by enhancing users' sensations within the PAL, two challenges were overcome in this dissertation: obtaining a photorealistic 3D model and estimating the pose of an object to enable 3D interaction from 2D equirectangular images.

The first challenge was overcome by combining photorealistic with 3D environment representations using a 360° high-quality image and a 3D model of an environment with low-quality. At the core of the proposed system was developed an approach for automatic large-scale 360° camera pose estimation within a 3D environment and a method for projective texture mapping spherical images. The camera pose estimator developed works for significant differences in rotation and displacement and works without the need to start from a known point of view. The positions and orientations of the camera were estimated with a translation error below 0.7 m, and below 1° and

2° for the difference in the amount of rotation, and the difference in the rotation axis orientation, respectively. These results were obtained for both virtual environments analyzed at full size and with search limits of ±20.00 m for translations and ±80.00° for rotations using an MSE of 0.005 as a possible discriminant factor for accuracy. While this work was validated using a 360° camera simulation in virtual scenes, its capabilities can also be tested on real scenes. In such situations, the light conditions could be very different between the model and the equirectangular image, so the luminance must be carefully considered. Furthermore, the presented approach is valid until the view of the user rotates without large displacements from the camera's initial position because not all the mesh areas are covered after the pixel projection. To overcome this problem, the same method can be applied with more than one camera, but in the case of the texture's final reconstruction, there is no discriminating parameter that allows us to choose which pixels to use from one or another camera for the final reconstruction. This choice can be useful if the field of view of one camera is better for some mesh areas than another to obtain a better result and can be implemented in future work. As discussed in this dissertation, this issue can be easily overcome by a novel technique, NeRF, that, starting from a collection of 2D images or videos, directly generates high-quality 3D renderings. Finally, in the optimization camera pose process, a further study can be done to find a correlation between the different terms of the cost function and the uncertainty in translation and rotation by investigating other possible acceptance criteria through a multidimensional analysis.

The second challenge was overcome with an innovative method designed to estimate the 6DoF pose of vehicles in equirectangular images. This method relies on deep learning methods only for the object segmentation, while the pose is estimated through a cost function optimization. Only the CAD model of the object is needed for this step, even without textures, for the nature of the cost function used. This makes the proposed method quite flexible to be applied to any object and lighting conditions due to the lack of color-affected terms in the comparison for pose estimation. The algorithm results were tested through an experimental setup, comparing them to measured rotations and translations applied to the camera in the real world. A maximum difference of 3.2° was obtained from the ground truth data for rotations, and 4 cm for translations over a research range of ±20° and ±20 cm, respectively. Future works can improve the computational time and reduce the pose detection error.

Finally, an interactive AR demo for industrial settings was designed by integrating high-level functions into the virtual environment. It has received positive feedback as a useful tool for training and education of the proposed measurement system and for determining the optimal camera positions in the area where the scanning system will actually be installed.

The second designed application was to test the effectiveness of AR technology applied to an industrial operation as the manual grinding process. Operators' working capabilities and skills were analyzed by comparing performances and the processes' outcomes while performing the same activities with and without Microsoft HoloLens. The experimental evidence verified that human capability can be increased through AR technology. Five operators with no previous experience of such a technology tested and evaluated four AR interfaces. After an initial selection achieved through a trial session, the best one was exploited in an experimental campaign design to study the effect of the technology on the target grinding operation. The initial study presented to select the best interface has allowed to obtain excellent results for the operators who had no previous experience with this technology. Experimental results showed that, with visual and audio support, the operator could keep the parameters affecting the grinding operation constant within the tolerance limits of the process and throughout the working time. In addition, adjustable thresholds allowed more versatile process management with a more controlled outcome. This application proves how an operator equipped with the right technology returns to

be the heart of smart factories. He has fewer chances of making mistakes and completes actions in less time, thanks to the enhancement of his PAL through the amplification of his senses and the increase of available information. Economic benefits are achieved through reduced runtime errors, often reflected in a lower cost and better work.

Regardless of the field of application, each actor can improve his PAL from an augmented visualization to a collaborative or supervised framework, according to the desired level of perception.

In the coming years, challenges such as the initial cost of technology, content development and technical barriers will be overcome through research contributions and increasing adoption.

The combination of immersive technologies and optimized frameworks prospects a future in which technology serves humans, fostering a harmonious relationship with them and resulting in greater human well-being and progress.

References

[1] T Abmayr et al. "Realistic 3D reconstruction–combining laserscan data with RGB color information". In: *Proceedings of ISPRS Internation Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 35.Part B (2004), pp. 198–203.

[2] Clark C Abt. *Serious games*. University press of America, 1987.

[3] Karthik Adapa et al. "Augmented reality in patient education and health literacy: a scoping review protocol". In: *BMJ open* 10.9 (2020), e038416.

[4] Sylvie Allouche et al. *Inquiring into human enhancement: interdisciplinary and international perspectives*. Springer, 2015.

[5] Marcos Alonso, Alberto Izaguirre, and Manuel Graña. "Current research trends in robot grasping and bin picking". In: *The 13th International Conference on Soft Computing Models in Industrial and Environmental Applications*. Springer. 2018, pp. 367–376.

[6] Ryoma Aoyama et al. "The utility of augmented reality in spinal decompression surgery using CT/MRI fusion image". In: *Cureus* 13.9 (2021).

[7] John Asmuth and Michael L Littman. "Learning is planning: near Bayes-optimal reinforcement learning via Monte-Carlo tree search". In: *arXiv preprint arXiv:1202.3699* (2012).

[8] Osnat Atun-Einy and Michal Kafri. "Implementation of motor learning principles in physical therapy practice: survey of physical therapists' perceptions and reported implementation". In: *Physiotherapy theory and practice* 35.7 (2019), pp. 633–644.

[9] Jonathan T Barron et al. "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 5855–5864.

[10] Ryad Benosman, S Kang, and Olivier Faugeras. *Panoramic vision*. Springer-Verlag New York, Berlin, Heidelberg, 2000.

[11] Aude Billard and Danica Kragic. "Trends and challenges in robot manipulation". In: *Science* 364.6446 (2019), eaat8414.

[12] Mark Billinghurst and Hirokazu Kato. "Collaborative mixed reality". In: *Proceedings of the first international symposium on mixed reality*. 1999, pp. 261–284.

[13] Daniel Bolya et al. "YOLACT++: Better Real-time Instance Segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[14] Yannick Bukschat and Marcus Vetter. *EfficientPose: An efficient, accurate and scalable end-to-end 6D multi object pose estimation approach*. 2020. arXiv: `2011.04307 [cs.CV]`.

[15] Isidro III Butaslac et al. “Application of Participatory Design Methodology in AR: Developing Prototypes for Two Context Scenarios”. In: *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE. 2022, pp. 244–248.

[16] Isidro III Butaslac et al. “The feasibility of augmented reality as a support tool for motor rehabilitation”. In: *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*. Springer. 2020, pp. 165–173.

[17] Isidro M. Butaslac et al. “Systematic Review of Augmented Reality Training Systems”. In: *IEEE Transactions on Visualization and Computer Graphics* (2022), pp. 1–20. DOI: 10.1109/TVCG.2022.3201120.

[18] Christopher Cao and Robert J Cerfolio. “Virtual or augmented reality to enhance surgical education and surgical planning”. In: *Thoracic surgery clinics* 29.3 (2019), pp. 329–337.

[19] Gill Chard. “An investigation into the use of the Assessment of Motor and Process Skills (AMPS) in clinical practice”. In: *British Journal of Occupational Therapy* 63.10 (2000), pp. 481–488.

[20] Paolo Cignoni et al. “Meshlab: an open-source mesh processing tool.” In: *Eurographics Italian chapter conference*. Vol. 2008. Salerno, Italy. 2008, pp. 129–136.

[21] Mike Cooley. “Human-centered design”. In: *Information design* (2000), pp. 59–81.

[22] Lea Daling et al. “Mixed Reality Books: Applying Augmented and Virtual Reality in Mining Engineering Education”. In: *Augmented Reality in Education*. Springer, 2020, pp. 185–195.

[23] Mariolino De Cecco et al. “Augmented reality to enhance the clinician’s observation during assessment of daily living activities”. In: *Augmented Reality, Virtual Reality, and Computer Graphics: 4th International Conference, AVR 2017, Ugento, Italy, June 12-15, 2017, Proceedings, Part II 4*. Springer. 2017, pp. 3–21.

[24] Mariolino De Cecco et al. “Sharing Augmented Reality between a Patient and a Clinician for Assessment and Rehabilitation in Daily Living Activities”. In: *Information* 14.4 (2023), p. 204.

[25] Christopher J Dede, Jeffrey Jacobson, and John Richards. *Introduction: Virtual, augmented, and mixed realities in education*. Springer, 2017.

[26] Lijun Ding and Ardeshir Goshtasby. “On the Canny edge detector”. In: *Pattern Recognition* 34.3 (2001), pp. 721–725.

[27] Anne G Fisher and Abbey Marterella. “Powerful practice: A model for authentic occupational therapy”. In: *(No Title)* (2019).

[28] Alberto Fornaser et al. “Augmented virtualized observation of hidden physical quantities in occupational therapy”. In: *2018 International Conference on Cyberworlds (CW)*. IEEE. 2018, pp. 423–426.

[29] Andreas G Franke, Robert Northoff, and Elisabeth Hildt. “The case of pharmacological neuroenhancement: medical, judicial and ethical aspects from a german perspective”. In: *Pharmacopsychiatry* 48.07 (2015), pp. 256–264.

[30] Joaquın M Fuster. “Upper processing stages of the perception–action cycle”. In: *Trends in cognitive sciences* 8.4 (2004), pp. 143–145.

[31] Nicola Garau et al. “A multimodal framework for the evaluation of patients’ weaknesses, supporting the design of customised AAL solutions”. In: *Expert Systems with Applications* 202 (2022), p. 117172.

[32] Sergio Garrido-Jurado et al. “Automatic generation and detection of highly reliable fiducial markers under occlusion”. In: *Pattern Recognition* 47.6 (2014), pp. 2280–2292.

[33] Daniel Girardeau-Montaut. “CloudCompare”. In: *France: EDF R&D Telecom ParisTech* 11 (2016).

[34] Google Inc. *Rendering Omni-directional Stereo Content*. Accessed: 26-07-2023.

[35] Andrea Grisenti et al. “Technological Infrastructure Supports New Paradigm of Care for Healthy Aging: The Living Lab Ausilia”. In: *Ambient Assisted Living: Italian Forum 2019 10*. Springer. 2021, pp. 85–99.

[36] Jaakko Hakulinen et al. “Omnidirectional video in museums–authentic, immersive and entertaining”. In: *International Conference on Advances in Computer Entertainment*. Springer. 2017, pp. 567–587.

[37] Amanda M Hall et al. “The influence of the therapist-patient relationship on treatment outcome in physical rehabilitation: a systematic review”. In: *Physical therapy* 90.8 (2010), pp. 1099–1110.

[38] Robert M Haralock and Linda G Shapiro. *Computer and robot vision*. Addison-Wesley Longman Publishing Co., Inc., 1991.

[39] Zaixing He et al. “6D Pose Estimation of Objects: Recent Technologies and Challenges”. In: *Applied Sciences* 11.1 (2021), p. 228.

[40] Steven J Henderson and Steven Feiner. “Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret”. In: *2009 8th IEEE International Symposium on Mixed and Augmented Reality*. IEEE. 2009, pp. 135–144.

[41] Steven J Henderson and Steven K Feiner. “Augmented reality in the psychomotor phase of a procedural task”. In: *2011 10th IEEE international symposium on mixed and augmented reality*. IEEE. 2011, pp. 191–200.

[42] Wolfgang Hoenig et al. “Mixed reality for robotics”. In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2015, pp. 5382–5387.

[43] Sabera Hoque et al. “A Comprehensive Review on 3D Object Detection and 6D Pose Estimation with Deep Learning”. In: *IEEE Access* (2021).

[44] Shiu-Wan Hung, Che-Wei Chang, and Yu-Chen Ma. “A new reality: Exploring continuance intention to use mobile augmented reality for entertainment purposes”. In: *Technology in Society* 67 (2021), p. 101757.

[45] Shannon E Jarrott, Hye Ran Kwack, and Diane Relf. “An observational assessment of a dementia-specific horticultural therapy program”. In: *HortTechnology* 12.3 (2002), pp. 403–410.

[46] Jason Jerald. *The VR book: Human-centered design for virtual reality*. Morgan & Claypool, 2015.

[47] M Carmen Juan et al. “Using augmented reality to treat phobias”. In: *IEEE computer graphics and applications* 25.6 (2005), pp. 31–37.

[48] Denis Kalkofen et al. “Tools for teaching mining students in virtual reality based on 360 video experiences”. In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE. 2020, pp. 455–459.

[49] Denis Kalkofen et al. “Tools for Teaching Mining Students in Virtual Reality based on 360°Video Experiences”. In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 2020, pp. 455–459. DOI: 10.1109/VRW50115.2020.00096.

[50] Hyo Jeong Kang, Jung-hye Shin, and Kevin Ponto. “A Comparative Analysis of 3D User Interaction: How to Move Virtual Objects in Mixed Reality”. In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 2020, pp. 275–284.

[51] Sidney Katz. “Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living.” In: *Journal of the American Geriatrics Society* 31.12 (1983), pp. 721–727.

[52] Michael Kazhdan and Hugues Hoppe. “Screened poisson surface reconstruction”. In: *ACM Transactions on Graphics (ToG)* 32.3 (2013), pp. 1–13.

[53] Maryam Khademi et al. “Comparing “pick and place” task in spatial augmented reality versus non-immersive virtual reality for rehabilitation setting”. In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2013, pp. 4613–4616.

[54] Tae Kyun Kim. “T test as a parametric statistic”. In: *Korean journal of anesthesiology* 68.6 (2015), pp. 540–546.

[55] Kiyoshi Kiyokawa, Haruo Takemura, and Naokazu Yokoya. “A collaboration support technique by integrating a shared virtual reality and a shared augmented reality”. In: *IEEE SMC’99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 99CH37028)*. Vol. 6. IEEE. 1999, pp. 48–53.

[56] Mykel J Kochenderfer. *Decision making under uncertainty: theory and application*. MIT press, 2015.

[57] Robert Laganiere and Florian Kangni. “Orientation and pose estimation of panoramic imagery”. In: *Mach Graph Vis* 19.3 (2010), pp. 339–363.

[58] Jiing-Yih Lai et al. “A high-resolution texture mapping technique for 3D textured model”. In: *Applied Sciences* 8.11 (2018), p. 2228.

[59] A-Young Lee et al. “Determining the effects of a horticultural therapy program for improving the upper limb function and balance ability of stroke patients”. In: *HortScience* 53.1 (2018), pp. 110–119.

[60] Anat Levin and Richard Szeliski. “Visual odometry and map correlation”. In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* Vol. 1. IEEE. 2004, pp. I–I.

[61] Michael Lederman Littman. *Algorithms for sequential decision-making*. Brown University, 1996.

[62] Wei Liu et al. “Ssd: Single shot multibox detector”. In: *European conference on computer vision*. Springer. 2016, pp. 21–37.

[63] Xiao Xin Lu. “A review of solutions for perspective-n-point problem in camera pose estimation”. In: *Journal of Physics: Conference Series*. Vol. 1087. IOP Publishing. 2018, p. 052009.

[64] Alessandro Luchetti et al. “Omnidirectional camera pose estimation and projective texture mapping for photorealistic 3D virtual reality experiences”. In: *Acta IMEKO* 11.2 (2022).

[65] Alessandro Luchetti et al. “Stepping over Obstacles with Augmented Reality based on Visual Exproprioception”. In: *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE. 2020, pp. 96–101.

[66] Alessandro Luchetti et al. “The human being at the center of smart factories thanks to augmented reality”. In: *2019 IEEE 5th International forum on Research and Technology for Society and Industry (RTSI)*. IEEE. 2019, pp. 51–56.

[67] Andrew MacQuarrie and Anthony Steed. “The effect of transition type in multi-view 360 media”. In: *IEEE transactions on visualization and computer graphics* 24.4 (2018), pp. 1564–1573.

[68] Ameesh Makadia and Kostas Daniilidis. “Direct 3d-rotation estimation from spherical images via a generalized shift theorem”. In: *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.* Vol. 2. IEEE. 2003, pp. II–217.

[69] Ameesh Makadia and Kostas Daniilidis. “Rotation recovery from spherical images without correspondences”. In: *IEEE transactions on pattern analysis and machine intelligence* 28.7 (2006), pp. 1170–1175.

[70] Keiichi Matsuda. *HYPER-REALITY*. 2016. URL: `https://www.youtube.com/watch?v=mpbWQbkl8_g#t=20m15s`.

[71] Edward M Mikhail, James S Bethel, and J Chris McGlone. *Introduction to modern photogrammetry*. John Wiley & Sons, 2001.

[72] Ben Mildenhall et al. “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis”. In: 2020.

[73] Ben Mildenhall et al. “Nerf: Representing scenes as neural radiance fields for view synthesis”. In: *Communications of the ACM* 65.1 (2021), pp. 99–106.

[74] Paul Milgram and Fumio Kishino. “A taxonomy of mixed reality visual displays”. In: *IEICE TRANSACTIONS on Information and Systems* 77.12 (1994), pp. 1321–1329.

[75] Thomas Müller et al. “Instant neural graphics primitives with a multiresolution hash encoding”. In: *ACM Transactions on Graphics (ToG)* 41.4 (2022), pp. 1–15.

[76] Yuki Nakamura et al. “Supporting daily living activities using behavior logs and Augmented Reality”. In: *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*. IEEE. 2013, pp. 658–663.

[77] Joshua Newnham. *Microsoft HoloLens By Example*. Packt Publishing Ltd, 2017.

[78] Michael Niemeyer et al. “Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5480–5490.

[79] Trond Nilsen, Steven Linton, and Julian Looser. “Motivations for augmented reality gaming”. In: *Proceedings of FUSE* 4 (2004), pp. 86–93.

[80] World Health Organization et al. *The need to scale up rehabilitation*. Tech. rep. World Health Organization, 2017.

[81] Karen Otte et al. “Accuracy and reliability of the kinect version 2 for clinical measurement of motor function”. In: *PloS one* 11.11 (2016), e0166532.

[82] Volker Paelke. “Augmented reality in the smart factory: Supporting workers in an industry 4.0. environment”. In: *Proceedings of the 2014 IEEE emerging technology and factory automation (ETFA)*. IEEE. 2014, pp. 1–4.

[83] Jiapu Pan and Willis J Tompkins. "A real-time QRS detection algorithm". In: *IEEE transactions on biomedical engineering* 3 (1985), pp. 230–236.

[84] Haris Papasaika-Hanusch. "Digital image processing using matlab". In: *Institute of Geodesy and Photogrammetry, ETH Zurich* 63 (1967).

[85] Kiru Park, Timothy Patten, and Markus Vincze. "Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 7668–7677.

[86] Loren E Peitso and James Bret Michael. "The promise of interactive shared augmented reality". In: *Computer* 53.1 (2020), pp. 45–52.

[87] Laura Pérez-Pachón et al. "Effect of marker position and size on the registration accuracy of HoloLens in a non-clinical setting with implications for high-precision surgical tasks". In: *International journal of computer assisted radiology and surgery* 16 (2021), pp. 955–966.

[88] Roland Pfister et al. "Good things peak in pairs: a note on the bimodality coefficient". In: *Frontiers in psychology* 4 (2013), p. 700.

[89] Thammathip Piumsomboon et al. "Empathic mixed reality: Sharing what you feel and interacting with what you see". In: *2017 International Symposium on Ubiquitous Virtual Reality (ISUVR)*. IEEE. 2017, pp. 38–41.

[90] Shuwen Qiu et al. "Human-robot interaction in a shared augmented reality workspace". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 11413–11418.

[91] Jason Rambach et al. "Learning 6dof object poses from synthetic single channel images". In: *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE. 2018, pp. 164–169.

[92] Amir Rasouli and John K Tsotsos. "Autonomous vehicles that interact with pedestrians: A survey of theory and practice". In: *IEEE transactions on intelligent transportation systems* 21.3 (2019), pp. 900–918.

[93] Christian Reiser et al. "Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 14335–14345.

[94] Alessandro Ricci et al. "The mirror world: Preparing for mixed-reality living". In: *IEEE pervasive computing* 14.2 (2015), pp. 60–63.

[95] Tom Rodden, John A Mariani, and Gordon Blair. "Supporting cooperative applications". In: *Computer Supported Cooperative Work (CSCW)* 1 (1992), pp. 41–67.

[96] Nina Rohrbach et al. "An augmented reality approach for ADL support in Alzheimer's disease: a crossover trial". In: *Journal of neuroengineering and rehabilitation* 16 (2019), pp. 1–11.

[97] Dieter Schmalstieg and Tobias Hollerer. *Augmented reality: principles and practice*. Addison-Wesley Professional, 2016.

[98] Bernd Schwald and Blandine De Laval. "An augmented reality system for training and assistance to maintenance in the industrial context". In: (2003).

[99] Randall C Smith and Peter Cheeseman. "On the representation and estimation of spatial uncertainty". In: *The international journal of Robotics Research* 5.4 (1986), pp. 56–68.

[100] Ingrid Söderback, Marianne Söderström, and Elisabeth Schälander. "Horticultural therapy: the 'healing garden'and gardening in rehabilitation measures at Danderyd Hospital Rehabilitation Clinic, Sweden". In: *Pediatric rehabilitation* 7.4 (2004), pp. 245–260.

[101] Pierre Soille et al. *Morphological image analysis: principles and applications*. Vol. 2. 3. Springer, 1999.

[102] Michele Stocco et al. "Augmented reality to enhance the clinical eye: The improvement of adl evaluation by mean of a sensors based observation". In: *Virtual Reality and Augmented Reality: 16th EuroVR International Conference, EuroVR 2019, Tallinn, Estonia, October 23–25, 2019, Proceedings 16*. Springer. 2019, pp. 291–296.

[103] Martin Sundermeyer et al. "Augmented autoencoders: Implicit 3d orientation learning for 6d object detection". In: *International Journal of Computer Vision* 128.3 (2020), pp. 714–729.

[104] M Tan and Q EfficientNet Le. "Rethinking model scaling for convolutional neural networks. arXiv 2019". In: *arXiv preprint arXiv:1905.11946* (2020).

[105] Arthur Tang et al. "Comparative effectiveness of augmented reality in object assembly". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2003, pp. 73–80.

[106] Markus Tatzgern et al. "Exploring real world points of interest: Design and evaluation of object-centric exploration techniques for augmented reality". In: *Pervasive and mobile computing* 18 (2015), pp. 55–70.

[107] Markus Tatzgern et al. "Transitional augmented reality navigation for live captured scenes". In: *2014 IEEE Virtual Reality (VR)*. IEEE. 2014, pp. 21–26.

[108] Theophilus Teo et al. "Mixed reality remote collaboration combining 360 video and 3d reconstruction". In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–14.

[109] Jayant Thatte and Bernd Girod. "Towards Perceptual Evaluation of Six Degrees of Freedom Virtual Reality Rendering from Stacked OmniStereo Representation". In: *Electronic Imaging* 2018.5 (2018), pp. 352-1–352-6. ISSN: 2470-1173.

[110] Jayant Thatte and Bernd Girod. "Towards perceptual evaluation of six degrees of freedom virtual reality rendering from stacked omnistereo representation". In: *Electronic Imaging* 30 (2018), pp. 1–6.

[111] Jonathan Tremblay et al. "Deep object pose estimation for semantic robotic grasping of household objects". In: *arXiv preprint arXiv:1809.10790* (2018).

[112] Shiyao Wang et al. "Augmented reality as a telemedicine platform for remote procedural training". In: *Sensors* 17.10 (2017), p. 2294.

[113] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[114] Stefan Wiedenmaier et al. "Augmented reality (AR) for assembly processes design and experimental evaluation". In: *International journal of Human-Computer interaction* 16.3 (2003), pp. 497–514.

[115] Dennis Wolf et al. "care: An augmented reality support system for dementia patients". In: *Adjunct Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 2018, pp. 42–44.

[116] Di Wu et al. “6D-VNet: End-To-End 6DoF Vehicle Pose Estimation From Monocular RGB Images”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2019, pp. 1238–1247. DOI: 10.1109/CVPRW.2019.00163.

[117] Yu Xiang et al. “PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes”. In: *CoRR* abs/1711.00199 (2017). arXiv: 1711.00199. URL: http://arxiv.org/abs/1711.00199.

[118] Wenyan Yang et al. “Object detection in equirectangular panorama”. In: *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE. 2018, pp. 2190–2195.

[119] Zongxin Yang, Xin Yu, and Yi Yang. “Dsc-posenet: Learning 6dof object pose estimation via dual-scale consistency”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 3907–3916.

[120] Matteo Zanetti et al. “Integrated Measurements of Stress, Motion Capture and Environmental Parameters for Ambient Assisted Living Scenarios”. In: (2019).

[121] Matteo Zanetti et al. “Object Pose Detection to Enable 3D Interaction from 2D Equirectangular Images in Mixed Reality Educational Settings”. In: *Applied Sciences* 12.11 (2022), p. 5309.

[122] Shumin Zhai, Jing Kong, and Xiangshi Ren. “Speed–accuracy tradeoff in Fitts’ law tasks—on the equivalency of actual and nominal pointing precision”. In: *International journal of human-computer studies* 61.6 (2004), pp. 823–856.

[123] Pengyu Zhao et al. “Spherical criteria for fast and accurate 360 object detection”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 2020, pp. 12959–12966.

[124] Egui Zhu et al. “Augmented reality in healthcare education: an integrative review”. In: *PeerJ* 2 (2014), e469.

List of Figures

List of Tables