POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Bridging Realities: Advancing Human-Avatar Interaction through Sensory Translation Systems in the Physical Metaverse

Tesi di Laurea Magistrale in
Computer Science and Engineering - Ingegneria Informatica

Author: **Maurizio Vetere**

Student ID: 970568
Advisor: Prof. Andrea Bonarini
Co-advisors: Federico Espositi
Academic Year: 2022-23

Abstract

This thesis explores human-robot interaction, focusing on sensory translation to enable humans to embody non-humanoid robot avatars in combined physical and virtual spaces. The core component is a Sensory Translation System, which connects a human Controller to a teleoperated Physical Avatar, in order to study nonverbal communication during a structured activity with an external human Visitor.

The research integrates technologies and methodologies, including hardware, like virtual reality headsets and robots with sensory inputs, and software, for environmental affordance information extraction and human pose translation. This technology bridges the communication gap between the avatar and the human Visitor.

Experiments in shared physical and virtual environments were conducted, using questionnaires for user feedback. The study adapted the Physical Metaverse framework to various robotic platforms and showcased new Human Translations at the X-Cities exhibition and at a Milano Digital Week event.

A key part of the thesis is the "First Contact" experience at X-Cities, demonstrating the potential of this technology in human-robot interaction and expanding the Physical Metaverse. The experience includes two "Mazes": a virtual one, where users embody the Physical Avatar, and a physical one, where they interact with the avatar.

In "The Virtual Maze" users felt a connection with the virtual being, with mixed reactions on understanding its actions and intentions. Some tried communicating, feeling a sense of companionship. In "The Physical Maze" users often perceived the robot as intelligent, developing emotional connections to varying degrees.

A core component that steered and sped up the development is a simulation that is in fact a digital twin of our real system.

This research advances human-robot interaction technology and provides ground for valuable insights into robotics, human-computer interaction, and possibly psychology. It lays the foundation for future experiments within the framework of the Physical Metaverse, exploring theories and applications in sensory translation, human pose translations.

Keywords: human-robot interaction, sensory translation, nonverbal communication, digital twin, Physical Avatar, physical metaverse

Abstract in lingua italiana

Questa tesi esplora l'interazione uomo-robot, concentrandosi sulla traduzione sensoriale per permettere agli umani di impersonare avatar robotici non umanoidi in spazi fisici e virtuali combinati. Il componente principale è un Sistema di Traduzione Sensoriale che connette un operatore umano a un avatar fisico teleoperato, per studiare la comunicazione non verbale durante un'attività strutturata con un visitatore umano esterno.

La ricerca integra tecnologie e metodologie, inclusi hardware, come visori per la realtà virtuale e robot con input sensoriali, e software, per l'estrazione delle informazioni di affordance ambientale e la traduzione della posa umana. Questa tecnologia colma il divario comunicativo tra l'avatar e il visitatore umano.

Sono stati condotti esperimenti in ambienti fisici e virtuali condivisi, utilizzando questionari per il feedback degli utenti. Lo studio ha adattato il framework del Physical Metaverse a varie piattaforme robotiche e presentato nuove traduzioni umane nella mostra X-Cities e a un evento della Milano Digital Week.

Una parte fondamentale della tesi è l'esperienza "First Contact" a X-Cities, che dimostra il potenziale di questa tecnologia nell'interazione uomo-robot e nell'espansione del Physical Metaverse. L'esperienza include due "Mazes": uno virtuale, dove gli utenti impersonano l'avatar fisico, e uno fisico, dove interagiscono con l'avatar.

Nel "Virtual Maze" gli utenti hanno sentito una connessione con l'essere virtuale, con reazioni miste sulla comprensione delle sue azioni e intenzioni. Alcuni hanno provato a comunicare, sentendo un senso di compagnia. Nel "Physical Maze" gli utenti hanno spesso percepito il robot come intelligente, sviluppando connessioni emotive a vari livelli.

Un componente centrale che ha guidato e accelerato lo sviluppo è una simulazione rappresenta un digital twin del nostro sistema reale.

Questa ricerca fa avanzare la tecnologia dell'interazione uomo-robot e fornisce basi per intuizioni preziose nella robotica, nell'interazione uomo-computer e, possibilmente, nella psicologia. Pone le fondamenta per futuri esperimenti all'interno del quadro del Physical Metaverse, esplorando teorie e applicazioni nella traduzione sensoriale e nella traduzione

della posa umana.

Parole chiave: interazione uomo-macchina, traduzione sensoriale, comunicazione non verbale, digital twin, Physical Avatar, physical metaverse

Contents

1 | Introduction

1.1. Aim of this project

This work addresses a critical challenge within the realm of embodied non-humanoid avatars.
A person that embodies a non-humanoid avatar is inevitably going to perceive and act in the environment through a body that is not his own: in order to have the user feel such body as his own there has to be a technological system that flawlessly translates the avatar's perceptions for the user to sense, and sends the user's inputs for the avatar to actuate.
Furthermore this system, that we will call "Sensory Translation System", has to maintain meaningful communication between the Avatar's Controller and other people that are in physical proximity of the Avatar. We want to focus on one-to-one interaction, so there will be just one person interacting with the avatar that we are going to call "The Visitor". For the moment, we have also considered only the visual channel for communication, so interaction happens through non verbal cues. In order to assess the quality of communication between the parties we engaged them in a structured activity with a shared objective.

This work is a vital component of the larger ongoing Physical Metaverse project and it builds upon the achievements of previous theses within the project. Specifically, it draws upon the integration of an initial Sensory Translation System developed by Giuseppe Epifani as his Master thesis [6], then further developed in this thesis, and two different robots: Odile, an emotional robot built as her thesis by Erica Panelli [18], and Blackwings, a theatrical performance robot built by professor Andrea Bonarini.
What sets this thesis apart from previous work is the introduction of a shared cooperative goal oriented activity between two human subjects, the Visitor and the Controller, and the fact that the Controller will participate to the activity using a Physical Avatar (a Robot) mediated by the Sensory Translation System.

1.2. The Importance of this Project

This project, centered around the advancement of human-avatar interaction through sensory translation systems in the Physical Metaverse, holds significant importance in several key areas.

1.2.1. Innovative Integration of Technology

At the forefront of technological innovation, this project integrates advanced concepts such as virtual reality, robotics, and sensory translation. It not only demonstrates the practical application of these technologies but also explores their synergistic potential. By bridging the gap between the virtual and physical realms, the project sets a precedent for future explorations in this rapidly evolving field.

1.2.2. Enhancing Human-Robot Interaction

The project contributes substantially to the field of human-robot interaction. By enabling humans to embody non-humanoid robot avatars, it opens up new possibilities for remote communication, control, and collaboration. This has profound implications for industries such as telemedicine, remote surveillance, and disaster response, where human presence is risky or unfeasible.

1.2.3. Expanding the Boundaries of Nonverbal Communication

Through the focus on sensory translation in a structured activity, the project provides invaluable insights into nonverbal communication. It challenges and expands our understanding of how nonverbal cues can be translated and understood across different embodiments, which is crucial in a world increasingly reliant on remote and digital interactions.

1.2.4. Contributions to the Physical Metaverse Concept

As a vital component of the larger ongoing Physical Metaverse project, this research contributes to the conceptual and practical development of the Physical Metaverse. It explores the integration of physical and digital spaces in innovative ways, paving the way for future applications that could transform how we interact with and perceive our environment.

1.2.5. Educational and Research Implications

The project serves as a rich educational resource for students and researchers in fields such as computer science, robotics, virtual reality, and human-computer interaction. It offers a practical case study in integrating diverse technologies and provides a platform for further research and exploration.

1.2.6. Social and Cultural Impact

Finally, the project has the potential to make a significant social and cultural impact. By exploring new forms of interaction and communication, it can help overcome barriers imposed by distance, physical limitations, and even language. This opens up new avenues for social interaction, entertainment, and cultural exchange in a globally connected world.

In summary, the importance of this project lies in its innovative approach to combining cutting-edge technologies, its contribution to human-robot interaction, its exploration of nonverbal communication, and its potential impact on education, society, and culture. As we continue to venture into increasingly digital and interconnected worlds, the insights and developments from this project will undoubtedly have far-reaching implications.

1.3. Thesis structure

This dissertation consists of 8 chapters.
In order:

- Chapter 1 - **Introduction**
  Provides an overview of the project, its objective and its utility.
- Chapter 2 - **Theoretical Framework**
  Explores the theoretical foundations relevant to the project.
- Chapter 3 - **Project Implementation**
  Describes the practical aspects and methodologies employed in implementing the project.
- Chapter 4 - **The System**
  Details the functionalities of the system developed.
- Chapter 5 - **The Hardware**
  Provides specifications and details about the physical components used in the project.
- Chapter 6 - **External Software**

Discusses the external software tools and frameworks integrated into the project.

- Chapter 7 - **The Experiments**
  Presents information on the experimental procedures, testing, and analysis.
- Chapter 8 - **Conclusions**
  Summarizes key findings, outcomes, and possible directions for future development.

2 | Background

This project emerges as an engineering approach to the concepts outlined in the pilot paper *Towards a Framework for Embodying AnyBody Through Sensory Translation and Proprioceptive Remapping: a Pilot Study* [8]. The paper represents the initial formalization of a larger research project called *Physical Metaverse*.

Our objective is to translate the insights and components studied in the paper into tangible software and hardware implementations. This effort aims to create a system for experimentation and validation of the theoretical framework. We want to deliver a system that can be expanded and enhanced with a more design-oriented approach in the future, moving beyond the technical challenges we are currently tackling. Subsequently, the intention is to build upon this foundation to advance further research in this domain. The interim results of this work have led to a more advanced formulation presented in the paper titled *First Contact: Crafting Authentic 'Otherness' with Embodied Narratives through Play and Theatre* [7].

In this chapter, we will recall the main concepts introduced in the paper, review relevant fields of technology and introduce a series of notions that will be helpful in understanding the subsequent content and the work at hand.

2.1. The Sense of Embodiment

The Sense of Embodiment, the perceptual integration with a virtual or physical body, encompasses three key dimensions [8]: Sense of Body Ownership, Sense of Agency, and Sense of Self-Location.

2.1.1. Sense of Body Ownership

Referring to the perception that one's body is the origin of sensations, Sense of Body Ownership has been extensively studied, with origins in experiments like the Rubber Hand Illusion [2]. Its malleability, influenced by factors like avatar appearance and perspective, underscores the complex interplay of multisensory perception.

2.1.2. Sense of Agency

Defined by the feeling of causing one's actions, Sense of Agency ties closely to real-time visuomotor correlations. Contrary to assumptions, it is not solely contingent on the fidelity of avatar representation but thrives on efficient control [21]. Virtual reality can maintain this sense through timely visual feedback, irrespective of the degrees of freedom in control.

2.1.3. Sense of Self-Location

Capturing the spatial aspect of embodiment, Sense of Self-Location delineates the experience of inhabiting a body. Influenced by factors like first-person perspective and visuo-proprioceptive correlations, it provides a reference frame for bodily sensations. The malleability of self-location emphasizes the role of synchronous visuo-proprioceptive cues.

2.2. Sensory Translation

As Sutherland [22] suggests, virtual objects need not adhere to the physical rules we know. Merely substituting a camera for an avatar's eyes proves insufficient for a complete experience in a radically different body. Sensory Translation becomes a cornerstone in our framework, acting as a system that collects data from the avatar's perception (robot sensors in our case) and translates it into user-perceivable information [8].

Sensory Translation's design prioritizes interaction channels to amplify the three dimensions of the Sense of Embodiment. Leveraging first-person view, vision through a VR headset, sound, and haptic sensations, Sensory Translation aims to provide a comprehensive and congruent sensory experience. Emphasis on synchronous representation of control signals ensures a user's sense of agency and facilitates motor learning.

The crucial integration of visual, auditory, and haptic feedback in Sensory Translation aims to establish a unified source of body ownership. Studies supporting minimal representations highlight their utility in providing relevant information [26] and fostering experimentation [14]. The intentional removal of verbal communication encourages users to explore diverse channels, promoting emergent behaviors and a richer embodied experience.

2.3. Human and Robotic Perception

The intricacies of human perception, encompassing the sense of embodiment discussed earlier, extend to the profound capacity to interpret stimuli received through sensory organs. This process involves collaboration between different parts of the brain and sensory systems to create a coherent understanding of the world around us. In the field of robotics, perception involves machines using sensors and algorithms to sense and understand their environment.

Recent advancements in robotic perception, driven by improvements in sensors, machine learning, and computational power, have had a significant impact on various applications like industrial automation and autonomous vehicles. These technological strides enable robots to navigate and interact with their surroundings, mimicking some aspects of human sensory abilities [12].

This progress not only enhances teleoperation in challenging environments but also improves the overall control and manipulation of robots, especially in remote or inaccessible settings. The evolution of both human and robotic perception is creating new possibilities and blurring the lines between biology and technology.

2.4. Teleoperation

Teleoperation, the act of remotely guiding devices, finds its strength in empowering human operators to maneuver in hazardous and unpredictable environments. This approach not only facilitates real-time adaptability but also opens avenues for remote experts to guide less experienced personnel, enhancing overall performance and minimizing errors. However, teleoperation grapples with challenges such as latency and limited situational awareness [19]. While various technologies, from basic remote control to advanced methods like haptic feedback and augmented reality, exist, each has its unique limitations. Embracing embodiment in teleoperation, be it physical or virtual, emerges as a promising avenue to overcome these challenges, providing operators with an immersive and intuitive experience.

2.4.1. Embodiment in Teleoperation

Embodiment, the sense of "being there" in a remote environment, emerges as a game-changer in enhancing teleoperation. Whether through a physical robotic avatar mirroring real-time movements or a virtual representation in a digital realm, embodiment offers valu-

able feedback, elevating the operator's sense of presence and situational awareness [24].

2.4.2. Virtual Reality in Teleoperation

The synergy between Virtual Reality (VR) and teleoperation holds promise in creating a seamless embodiment experience [10, 20]. VR's immersive environment, coupled with haptic feedback and multimodal interfaces [5, 13, 24], addresses challenges posed by latency and enhances the operator's control and presence in the remote setting. Careful consideration of sensor selection, data processing, and feedback mechanisms is crucial for designing effective VR-based teleoperation systems.

2.5. Computer Vision

Computer vision is a multidisciplinary field that empowers machines to gain high-level understanding from visual data. It leverages the capabilities of computers to interpret and take decisions based on visual information. The primary goal is to enable machines to replicate aspects of human vision, allowing them to perceive, interpret, and respond to the visual world.

Central to computer vision is the development of algorithms and models that can analyze images and videos, extracting meaningful insights and information. This involves tasks such as image recognition, object detection, facial recognition, and scene understanding. The field intersects with various domains, including artificial intelligence, machine learning, and image processing.

In recent years, computer vision has undergone rapid advancements, driven by the surge in computational capabilities, the availability of vast datasets, and breakthroughs in machine learning techniques. These developments have propelled applications across industries [16], from healthcare and automotive to security and entertainment [15]. In the context of robotic systems, computer vision plays a pivotal role in endowing machines with the ability to perceive and interact with their environment, facilitating tasks like navigation, object manipulation, and human-robot interaction.

2.5.1. Computer Vision in Teleoperation

Computer vision makes computers "see" and understand images or videos. Hence it plays a pivotal role in advancing the field of teleoperation, contributing significantly to the sense of embodiment. Through the integration of computer vision technologies, teleoperators can leverage visual information for a more immersive and intuitive remote control

experience [24].

2.5.2. Visual Perception and Feedback

Computer vision algorithms enable the extraction and interpretation of visual data from the remote environment. This information can be utilized to provide teleoperators with real-time visual feedback, enhancing their understanding of the surroundings and facilitating better decision-making. For instance, object recognition and depth perception algorithms contribute to the identification of obstacles, allowing operators to navigate through complex environments with increased precision [25].

2.5.3. Challenges and Considerations

While computer vision enhances teleoperation, it comes with challenges such as managing diverse environmental conditions and ensuring robustness in real-world applications. Additionally, the integration of computer vision should be aligned with the overall teleoperation system design, considering factors such as latency, accuracy, and the specific requirements of the teleoperation task.

2.6. Human-Robot Interaction

Human-Robot Interaction (HRI), a dynamic interdisciplinary field, aspires to develop robots capable of intuitive and natural interaction with humans. This pursuit requires a deep understanding of human psychology, social behavior, and the technical intricacies of robot design and programming.

2.6.1. The Role of the Physical Avatar

The introduction of Physical Avatars, remotely controlled robots facilitating communication and interaction, marks a transformative development. With applications ranging from teleoperation and telepresence [23] to social interaction and therapy [1], Physical Avatars promise to redefine human-robot communication. Their potential to enhance natural and intuitive interaction positions them as catalysts for future research and development.

3 | Theoretical Framework

In this chapter, we will delve into the new theoretical foundations formulated to steer the implementation of the project. As the Physical Metaverse theoretical framework had not yet considered any targeted activity between the Physical Avatar and the Visitor, we initiated a study to select the activity for our two participants. Subsequently, we will present a formalization of the system to implement, complete of our activity, accompanied by alternative reformulations of it that proved instrumental during the development process.

3.1. Study on Activities

The main new component we wanted to introduce in the interaction between Physical Avatar and Visitor is the goal-oriented activity. We required an activity that could test our sensory translation, enabling us to study the fundamental components in human communication and explore how much they could be altered while still maintaining effective communication.

The goal-oriented activities we explored are essentially games, and henceforth, we will also use this term to refer to them.

However, we weren't looking for just any kind of game. To start with, we were looking for a game for two players, and one in which no form of verbal communication would have to be contemplated by the rules. Through brainstorming sessions, we compiled an initial list of games. Then we tried to break down each game into components that are relevant to our system, as we will later explain.

It is important to note that this was primarily a qualitative study, serving as an initial analysis aimed at integrating a targeted activity into the Physical Metaverse framework. We emphasize that the main goal of this thesis was to overcome the technical challenges of sensory and Human Translation in a real-world application. Therefore, the sole essential requirements for our activity were to prompt the Controller (through the avatar) and the Visitor to interact non-verbally with each other, referencing a physical environment

perceived by both, albeit with different affordances. The interaction should lead to a meaningful exchange of information that is comprehensible and interpretable by both parties. Further exploration and refinement of this study shall be addressed in a more targeted discussion on the subject.

3.2. Games and Activities

This study served to provide numerous insights into the choice of the activity for the final experiments of the thesis that led to the development of Odile [18]. The results of that thesis, in turn, helped us validate the choice of the activity and incorporate it into our system. The goal was also to begin standardizing the experimentation framework in the physical metaverse, aiming to focus on a specific activity to build a baseline for reference in subsequent experiments.

In our quest to identify activities that emphasize physicality and movement we gathered a variety of options to explore. We aimed to select games that present unique challenges, contributing to our goal of testing the limits of sensory and Human Translation in a real-world application. The list of selected games, initially chosen from a broader range with fewer restrictions, is provided below with brief descriptions of each game:

1. **Twister:** A physical game that involves players placing their hands and feet on colored circles based on a spinner's instructions, often leading to precarious positions and elimination if players fall.
2. **Escape Room:** A team-based game where players discover clues, solve puzzles, and accomplish tasks in a limited time to achieve a specific goal, often involving escaping from a confined space.
3. **Build a Tower:** Teams compete to build the tallest tower using only newspapers and masking tape.
4. **Laser Tag:** A recreational shooting sport where participants use infrared-emitting light guns to tag designated targets, highlighting physical movement and precision.
5. **Blind Man's Bluff:** A variant of tag where the player who is "It" is blindfolded, requiring reliance on other senses to locate and tag other players.
6. **Airplane Game:** Blindfolded "airplanes" navigate a runway with guidance from a "navigator" who provides step-by-step directions, in our case we will restrict communication to touch.
7. **Three Leg Run:** A race where pairs of participants are physically connected, with

their left and right legs strapped together, aiming to reach the finish line before other contestant pairs or in our case when the time is up.

8. **Hide and Seek:** Players hide, and one designated player ("It") seeks to find them.
9. **Tag:** A classic playground game where players chase and tag others.
10. **Pictionary:** A drawing and guessing game where one player illustrates a word or phrase for their partner to guess, in our case we would replace drawing with imitating.
11. **Capture the Flag:** Teams aim to capture each other's flag and have to avoid getting touched by opponents while carrying the flag.
12. **Smaugh's Jewels:** A game where players attempt to steal a "jewel" guarded by a "dragon".
13. **Trust Fall:** A trust-building exercise where one person falls backward, relying on another person to catch them.

These games present a diverse array of physical and cognitive challenges, with potential to provide valuable insights into the dynamics of human-avatar interactions during their play. Before starting our discussion on the activity's selection, we explored relevant literature for guidance. Fortunately, we didn't need to search far, as the book "Robot Play For All - Developing Toys and Games for Disability" by Professors Andrea Bonarini and Serenella Besio [1]offered us an initial mindset on how to approach and break down games into their components.

To begin our analysis, we divided the activities into cooperative and competitive (blue for cooperative and red for competitive in Figure 3.1); then we delved into the specific components of each activity, identifying five dimensions particularly relevant to our system, as depicted in Figure 3.1:

- **Information from robot to human:** How much information needs to be conveyed from the Physical Avatar to the Visitor for the activity's purpose.
- **Information from human to robot:** How much information needs to be conveyed from the human to the Physical Avatar for the activity's purpose.
- **Robot proprioception:** What level of proprioception by the robot is required.
- **Robot action complexity / control complexity:** How complex are the actions, and therefore the control by the user, that the robot needs to perform.

- **Robot environment knowledge:** How much the avatar needs to perceive about the environment and report to the Controller.

Figure 3.1: First formalization attempt on our study on activities.

Having built an initial framework to guide us in choosing the activity, now we just had to use it.

Each of our activities could be positioned higher or lower on each dimension depending on how developed it was.

However, we also had practical limitations with our system, so we couldn't choose an activity that simply maximized every category to achieve the richest possible interaction, we needed to find the best tradeoff.

The first most and important component we identified was the transfer of information from human to robot, which aligns closely with our goal of enriching our sensory translation with more advanced Human Translations.

There was also the complementary component of information from the robot to the human. Here, we couldn't push too far since the expressiveness of our robot would still be limited, given that the primarily studied communication direction is the other way around.

In the information from the environment to the robot, we wanted to go higher as the goal of this work was to consolidate and improve sensory translation, thus expanding the Physical Avatar's ability to perceive affordances in space.

Proprioception and control are two connected dimensions because a high level of control over one's actions requires the ability to perceive one's body and its effects on the immediate environment. We would have liked to develop these components further, but due to the scope of our work, we had to leave them more in the background, focusing on other aspects. Nevertheless, we decided to be ambitious about these components to leave a modular and extensible system and also to arrive at a final choice that would be shared by other theses.

As previously mentioned, this was a preliminary, qualitative study primarily aimed at guiding us towards an initial choice that could be put into practice. The idea was to see the results and then potentially re-evaluate and enrich this framework with experimental findings.

After comparisons and evaluations, our choice fell on the Escape Room activity. It seemed like a reasonable compromise between requirements and limitations. While it might not appear as the absolute best choice based on the graphs, we decided to make modifications to adapt this activity, which already generally aligned with our goals.

We particularly liked the mystery inherent in this activity. It would have justified, from a narrative perspective, the interaction with an alien Entity in the virtual world and the interaction with a non-humanoid robot in the physical world.

We then made the activity asymmetric, giving a guiding role to one of the two players, the Visitor, and the role of executor to the other, the Controller. This way, we could effectively overcome the challenges presented by this activity in the dimension of information transfer from robot to human.

Figure 3.2 shows a very naive prototype we developed in Unity just to play out our idea. The black and gray cylinders represent the person and the Physical Avatar. It was possible to embody the two protagonists on two separate computers and play in multiplayer. To

experiment with elements that may come in future projects, we designed a room with doors or grids and buttons to open them, giving participants an idea of progression. This goes a bit beyond the simple guiding and executing roles that we later formalized for our Escape Room, and, in fact, the prototype is included here for completeness of presentation of the development process. This approach of creating a virtual room will later inspire the development of a proper simulation of the system.

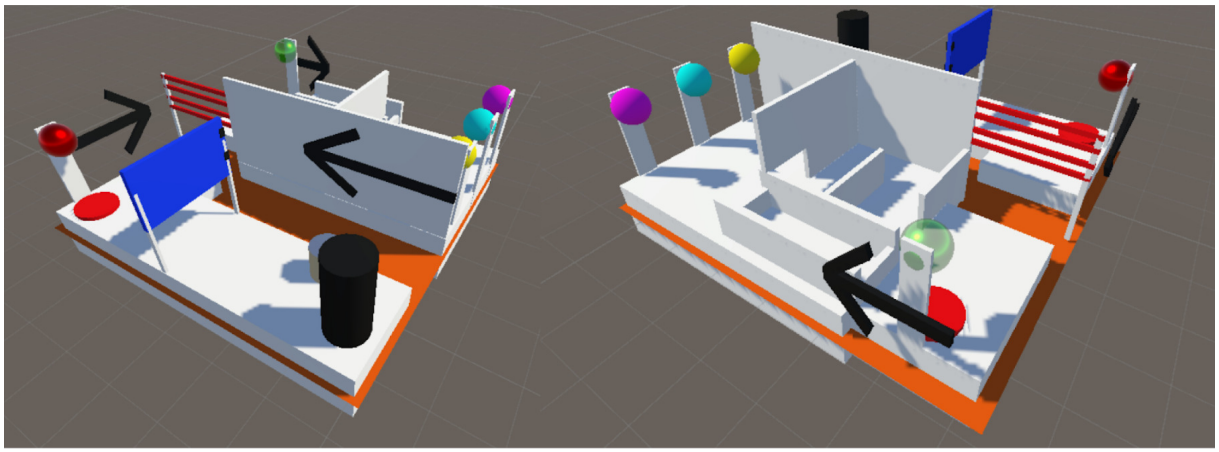


Figure 3.2: Pictures of a very first interactive digital draft for an Escape Room.

3.3. The System

The ultimate goal of this project is to have two individuals, a Visitor and a Controller, collaboratively engaging in a goal-oriented activity. During the activity, the Controller will not be physically present, unlike the Visitor. Instead, the Controller will participate through a Physical Avatar (a robot in our case) controlled remotely, mediated by a Sensory Translation System. Image 3.3 offers a graphical representation of our description.

The Visitor will be unaware that the robot with which he is interacting is, in fact, controlled by another person. Simultaneously, the Controller will not know that the Entity she saw in the virtual world, namely a Human Translation of the Visitor, is also a person.

We can say that our system has succeeded if, at the end of the tests, we can confidently assert the following:

- The two participants were able to understand each other.
- The two participants were able to collaborate.

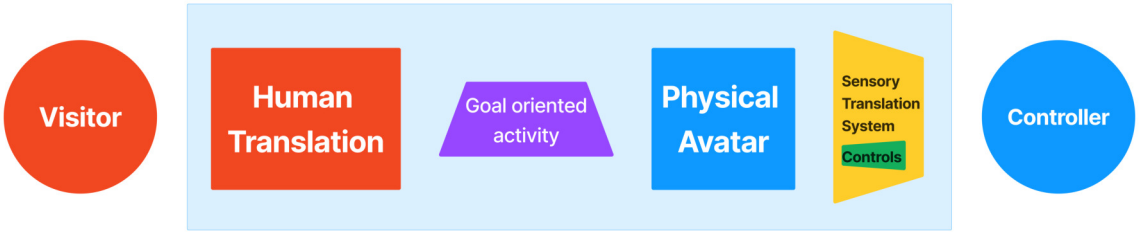


Figure 3.3: Overview of the Full System.

3.4. Components Virtualization

In the exploration of the landscape of human-avatar interaction, this document offers an analysis of alternative formulations for the components for our system. These formulations are not just theoretical concepts but also practical setups that have been tested and refined, pushing the boundaries of what is achievable in virtual interaction spaces. The core of these systems lies in their capacity to seamlessly integrate the physical and virtual realms.

The main drive to develop these alternatives arose in the very early stages of design: we had envisioned the possibility of creating a purely virtual version of our system that could run on a laptop without external hardware. This version would have implemented an online multiplayer system, allowing two players to connect and play as the Visitor and the Controller in a purely virtual world.

This system aimed to isolate and explore in more detail the newly introduced element, namely the goal-oriented activity.

In practice, however, this work evolved differently: more opportunities than initially expected arose to gather results from in-presence real user experiences, and these, in turn, led us to redirect our efforts toward the direct implementation of the final system, leaning on taking our choice of activity for granted, at least in this phase.

Anyway we came to the conclusion that virtualizing various portions of our system to a greater or lesser extent would in any case be useful in order to isolate possibly arising software and hardware problems, and not to be tied to specific hardware present just in the laboratory whenever we wanted to carry on our work while not physically there.

3.4.1. Virtual Visitor and Virtual Avatar

In our target real system, we have a Visitor who is physically near the Physical Avatar to carry out a goal-oriented activity. The Physical Avatar, in turn, is teleoperated by the Controller through the Sensory Translation System.

Therefore, our idea was to develop virtual versions of the Visitor and the Avatar. In practice, this virtualization translates, always considering a real person behind the actions of one or the other, into transposing their spatial movement component to the virtual world. This means that, for example, a Visitor, instead of walking around the play area, will control their virtual replica using an interface such as a keyboard or joystick, and similarly, a Controller will not move a physical robot but an exclusively virtual version. In fact, for the Controller, little will change in terms of controls because even in the real version of the system, they have the illusion of moving in a purely virtual world.

3.4.2. Virtual Visitor

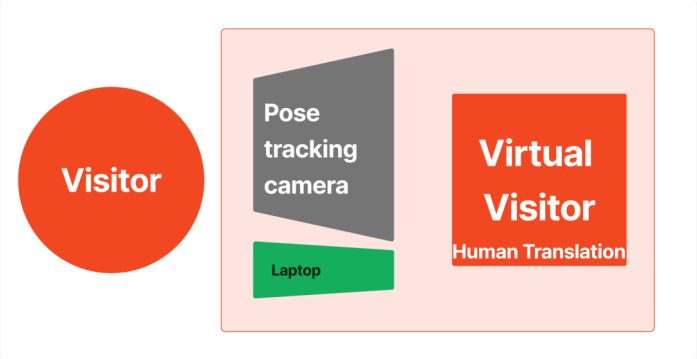


Figure 3.4: Virtualized version of the Visitor.

On the Visitor side, we have a person in front of a camera, along with the pose tracking software. With our system, it's possible to use any camera for pose tracking, not just the one we used on the robot, as long as the output pose landmarks adhere to the same format. More importantly, the camera will not have to be anymore located on the Physical Avatar; in fact, we could simulate [see later in Section 4.4] the entire system to have a Virtual Avatar. The key is that the relative position assumed between the camera, Avatar, and Visitor in the virtual world remains consistent.

Pushing the Virtual Visitor concept to the maximum extent, it would even be possible to skip pose detection altogether and generate consistent pose landmarks in other ways, for

example with a GUI to select specific emotes. However, this approach has not been used in what we present here.

3.4.3. Virtual Avatar

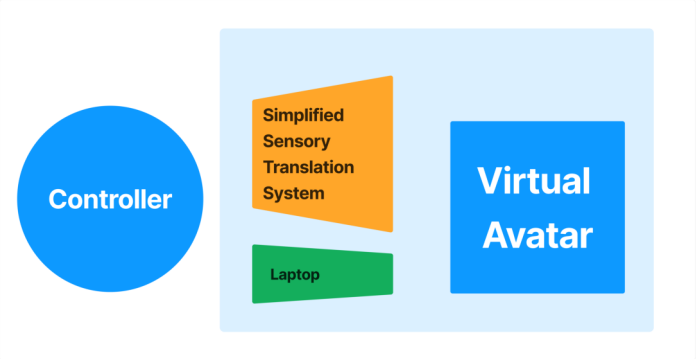


Figure 3.5: Virtual version of the Avatar.

On the Controller side, we have a user commanding the Avatar through a Sensory Translation System. Again it is possible to detach from the actual hardware and obtain two new components: a Simplified Sensory Translation System, represented by a computer screen where controls are keyboard and mouse-based, and a Virtual Avatar, in place of the actual robot.

Throughout the development, we used combinations of our initial Controller/Physical Avatar approach and this virtual one. Sometimes, we utilized the VR headset (main component of our Sensory Translation System) to navigate the exclusively virtual world (that we will later introduce as the Simulation), and at other times, we used keyboard and monitor to control the Physical Avatar. We often also just tested in the virtual realm by controlling our Virtual Avatar with monitor and keyboard.

3.4.4. System Variants

Reflecting on the newly formulated virtual components, two variants have emerged outside the framework of the final system that could be of interest in future developments.

The first one is an environment in which there are only Virtual Avatars with Sensory Translations that, instead of performing Human Translation, simply display the other Avatars [Figure 3.6]. The utility of this environment lies in the fact that it can lead to the emergence of new behaviors from the Controllers that were not initially considered.

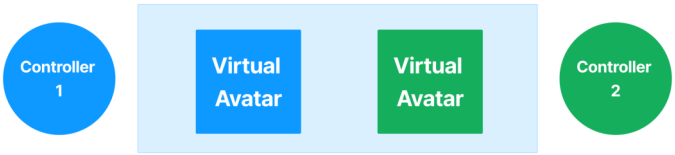


Figure 3.6: Virtual Avatars playground.

The second one is an environment complementary to the previous one, where users would see and be seen through Human Translation [Figure 3.7]. This version, too, can lead to the emergence of behaviors that might not be observed otherwise.

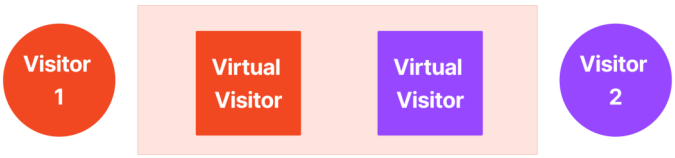


Figure 3.7: Human Translations playground.

3.4.5. Conclusion

In conclusion, this analysis has helped us gather multiple perspectives on our system and realize that this project has significant potential to blur the boundaries between physical and digital reality. In the following sections, we will revisit many of the theorized concepts, especially during the development phases where we needed a certain level of agility and didn't want to be hindered by the practical limitations of the hardware. Furthermore, the virtualization of components has allowed us, as good engineers, to apply the principle of the superposition of effects to the best extent.

4 | Project Implementation

This chapter outlines the whole development of the project, explaining step by step how we reached the final state of the system. We will walk through the stages that, from becoming familiar with the already existing systems, led to the results of the latest experiments conducted with the new system. First we will make a brief introduction about the final result we want to obtain.

4.1. The Objective

Sensory translation is the main focus of this project and it primarily focuses on vision, with some auditory and tactile components, given that this is still an initial approach in the development of the Physical Metaverse. Thoroughly covering all senses would require a much more extensive treatment and will probably come in later stages of this research.

The sensory translation system works to translate the perceptions of the Physical Avatar for its Controller. To achieve this objective, it leverages a synergy of electronic devices and software solutions, which will be detailed in the following sections.

In general, our Controller wears a VR headset to see what the avatar is perceiving. Additionally, sounds related to the activity are emitted, and the Controller can control the avatar's movement using joysticks. The Physical Avatar, in the final version represented by the robot Blackwings, is equipped with numerous sensors and actuators:

- A LIDAR to detect obstacles in the environment
- A webcam to detect "Stations", the objectives of the goal-oriented activity
- A Luxonis DepthAI camera for processing the Visitor's pose without computational burden on the onboard computer
- A mobile base to enable the robot to move in the environment
- Two servomotors for the pan and tilt of the DepthAI camera
- More servomotors for expressiveness

4.2. Starting Point

The implementation of the system began by building upon the foundation laid in the previous thesis, *Non-Verbal Communication Through a Robotic Physical Avatar: A Study Using Minimal Sensor Information* [6]. The initial setup involved a Physical Avatar and its sensory translation system, featuring a VR headset and two joystick controllers.

Expanding upon this existing framework, we introduced new sensors and features informed by the insights gleaned from past tests. A pivotal addition during this phase was the incorporation of the DepthAI camera, providing a fresh perspective on the perceived human pose.

(a) The past experimental setup. (b) The past robot.

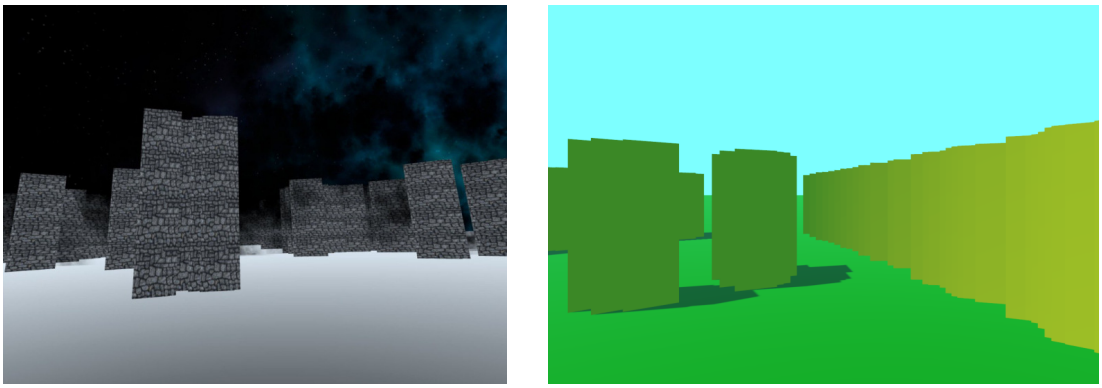
Figure 4.1: The starting point of this thesis.

The work seamlessly resumed from where it had been left after the aforementioned thesis' final experiments [Figure 4.1]. By retrieving both hardware and code, the system was restored to its state during the final experimental phase. After conducting preliminary exploratory tests to assess various features, the results were analyzed to identify focal points for further development.

An early observation from these experiments highlighted a recurring concern: the lack of a defined "purpose." This absence adversely affected the embodiment of the Controller in the Physical Avatar, as well as the interaction of the Visitor with the robot. Recognizing this, the subsequent work aimed to address and fix this issue to enhance the overall effectiveness of the system.

4.2.1. LIDAR Visualization

Before delving into the addition of new features, we opted to focus on refining existing elements. This approach aimed to familiarize ourselves with the system and discern the already available tools, minimizing the need for redundant development. Guided by test results, attention was then directed towards stabilizing the LIDAR visualization.



(a) Previous system's LIDAR Visualization

(b) New LIDAR Visualization. The shadows are projected by the newly added sun.

Figure 4.2: First experiment on the visualization's graphics.

Drawing from insights gained in the Computer Graphics course, we recognized the impact of manipulating light and color in a virtual environment on perceptual immersion. Consequently, we initiated changes to the textures of the pillars representing LIDAR-detected points. A uniform tint and geometric shapes that blended seamlessly proved instrumental in enhancing the stability of the overall visualization. Experimentation with fade-in and fade-out effects was also conducted. While these effects appeared to stabilize the environment, they introduced an unavoidable delay in visualization, negatively impacting the quality of embodiment.

While working on the visualization of these pillars, considerations were made on extracting information about distances to elements, s detected by LIDAR, and potentially grouping them into walls. However, a notable limitation emerged – the system lacked a "memory." The Controller remained consistently centered in the virtual world without any rotation. The surrounding pillars approached or receded, mirroring the 360-value array (one distance for each angle) received from the LIDAR. Crucially, the pillars only moved closer or farther away; none shifted laterally. This characteristic led to the characterization of the system as "memoryless."

Aware of SLAM algorithms, we hesitated to adopt them at this stage. We viewed them as potentially diverging from the envisioned Sensory Translation System, as relying heavily

on computer processing could distort the perception we aimed to maintain.

4.2.2. The Sun

After these initial trials, we were ready to develop new features. The first addition was the incorporation of a "sun" to provide an additional reference point for the Controller, who could easily lose orientation in such a homogeneous environment. To display the sun in the visualization, we needed the current orientation of the robot. From this orientation, we could rotate the sun in the virtual world (remembering that the Controller neither rotates nor translates; it's the world that changes around them).

To obtain this data, we utilized an onboard IMU on the robot, acknowledging that even a slight drift would not significantly compromise the experience. The results were satisfying; it became possible to constantly understand the orientation relative to the starting point by observing shadows in the surrounding environment.

4.2.3. Human Pose Estimation

When the DepthAI camera, long considered an essential component for advancing the Sensory Translation System, became available, we immediately began experimentation. After some research, we discovered the versatile Python repository for Human Pose Estimation, known as "DepthAI Blazepose" [9]. We seamlessly integrated the code into the robot's Jetson pipeline. The software loads a neural network onto the camera, which starts running it and sends the recorded video images and an array of pose landmark coordinates (containing features of the first detected human) via USB connection. We transmitted this array to Unity via UDP, adhering to the Key:Value format already used in the original Unity project. Upon receiving these values in the Unity project, we positioned spheres representing the pose's joints in the virtual world in front of the Controller.

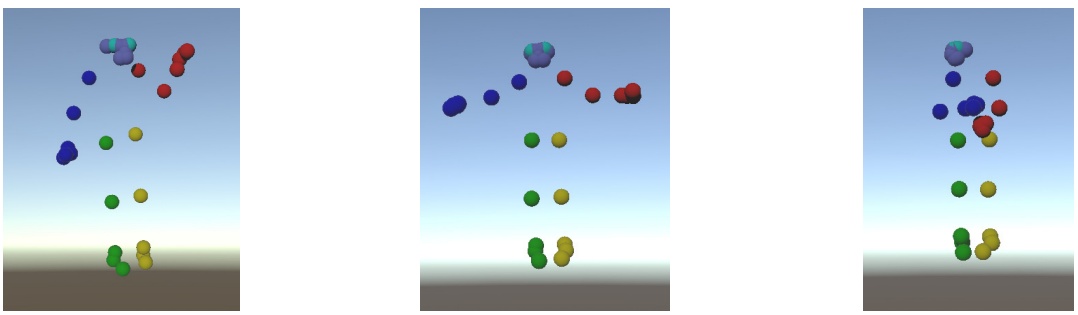


Figure 4.3: Examples of detected pose.

The primary challenge encountered was a Computer Vision issue. Although it was possible

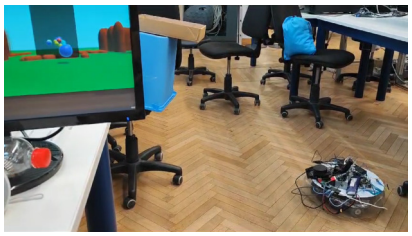
to recognize the person captured by the camera in the 3D reconstruction, their position and orientation relative to the observer needed adjustment through extrinsic calibration.

Using Unity's graphical interface, this calibration was relatively straightforward. Instead of explicitly calculating the transformation matrix representing the calibration, we opted to expose adjustable slider variables. With each update, these variables modified the position and orientation of the spheres, essentially the offsets and scales on the x, y, and z coordinates of the received points.

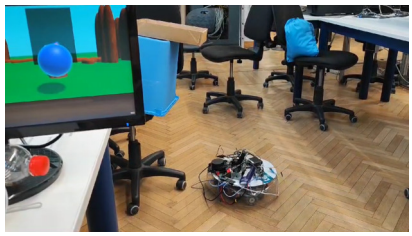
In this initial phase the results were satisfactory, placing the virtual person in a coherent position relative to the observer.

4.2.4. Color tracking

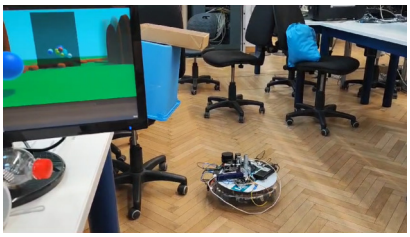
In anticipation of introducing a real-world activity through this sensory translation, we decided to work on the Controller's perception of relevant objects in the environment. The aim was to enable the Controller to extract useful information, not just for navigation but also to understand the affordances of the shared space. An initial, perhaps "naive" experiment involved detecting colored shapes in the camera image.



(a) The Robot is far from the blue box, the biggest blue ball in the screen represents the box.



(b) The Robot approaches the box, the blue ball is visualized as closer.



(c) The Robot rotates right, leaving the box to its side, the blue ball is now on the side of the view.

Figure 4.4: Reconstruction of the blue box's location.

In practice, the results were immediately interesting and proved highly valuable later in the project. The practical goal was to display the position of blue boxes in the virtual environment that the physical robot could see. These boxes were placed on chairs of similar heights for camera angle consistency.

To achieve this, we applied a color mask, in this case, blue, to the image and detected the largest "blob" (group of pixels) corresponding to that color. The assumption was that the largest blob represented the box the robot was observing. This, too, turned out to be a Computer Vision problem, perhaps even more complex than the previous one. The precise position of the object was crucial this time, and relying on color made it susceptible to changes in light intensity, potentially invalidating any calibration.

Concerning the translation of position from image space to world space, we quickly resolved it by making a simple assumption: the height of the boxes above the ground and their size were fixed. This way, a distant box would appear lower in the image, while a closer one would appear higher, hence given size and height in two dimensions we could recover their distance and in general position in three dimensions.

Another issue arose as the boxes easily exited the camera's field of view when the robot rotated. To address this, we experimented with integrating the position of a box lost by the camera using a combination of dead reckoning from sent movement input and IMU, the same used for the sun. As expected, the estimated position's reliability quickly dropped when undetected, but still it was better than nothing.

The process was evidently empirical and exploratory, but the visualization effectively worked, allowing for the identification of the boxes.

4.3. Preparation of Odile

After becoming familiar with (and testing) new features with the old system, we conducted the study described in the previous chapter to choose the activity for our system. Then we decided to collaborate with the student that was finishing the development of the Odile robot. There were communication issues between Arduino and the servomotors that we were able to resolve. Our primary contribution to completing the robot was the development of a system that connected the pan and tilt of the robot's camera to the orientation of the headset worn by the Controller. This allowed for a much more natural and expressive control in Odile's final experiments.

From the first-person control of the robot, a latency issue in video transmission became apparent. To address this, we developed a GUI that enables real-time adjustments to the

transmitted image size and JPG quality. By reducing these parameters, latency decreases. As expected it was observed that latency grows linearly with size and sublinearly with quality. Playing with the sliders allowed for quickly finding a functional compromise for real-time robot control.

The chosen activity in Odile's experiments was very similar to what we will implement and was in fact influenced by our choice. The difference was that the Physical Avatar and the Visitor's roles were switched, with the former being the guide in this case.

The results were promising for the continuation of this project.

4.4. The Simulation

The Simulation is a digital twin of the real system and has been a crucial element in the project's development. Thanks to it, the number of times it was necessary to actuate the real robot to test the sensory translation system was drastically reduced. More importantly, it allowed the isolation of any hardware-related complications from the real system, focusing on software challenges in terms of algorithms.

Since the Simulation does not have strict requirements in terms of physical fidelity with the real system, it was possible to develop it directly in Unity, in close contact with visualization.

An important factor for the Simulation is that it is completely transparent to the sensory translation system, meaning it can be used interchangeably with the real robot, avoiding unnecessary project development overhead.

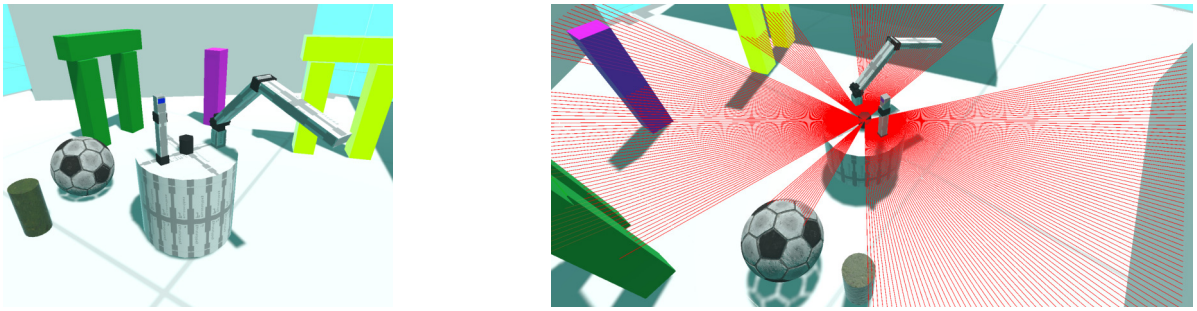
During development, there were instances where we worked with a special version of the avatar—half of its components in the physical robot and the other half in the virtual simulation.

4.4.1. Virtual Odile

"Virtual Odile" is the main avatar in the Simulation, a virtual copy of the Physical Avatar "Odile." It resides in Unity and is equipped with a "virtual Jetson" and virtual copies of sensors. This virtual Jetson receives data from the sensors and sends it to localhost via the UDP protocol, leveraging the existing infrastructure for data reception in the visualization. The visualization simply receives and displays the data without differentiating from which IP it comes.

Virtual Odile [Figure 4.5a] is placed in a virtual room with obstacles, similar to the envi-

ronments where we position the real robot. To simplify development and communication between modules, it is located in the same virtual space as the visualization. It is positioned away from the Controller and is invisible. This arrangement is somewhat consistent with the fact that the person wearing the VR headset is in the same physical space as the real robot, even if not in immediate proximity. IN the following Subsequently, the virtual components are presented.



(a) Virtual Odile. (b) Virtual LIDAR's ray casting.

Figure 4.5: Virtual Odile and Virtual LIDAR's ray casting.

Virtual LIDAR

Virtual LIDAR is an interesting case of sensor modeling. It uses a ray-casting [Figure 4.5b] to detect obstacles in Unity's 3D virtual environment, in order to simulate the real LIDAR. We also modelled noise to mimic the technical challenges of the actual LIDAR and prevent the development of algorithms that would only work in the "ideal case." It's worth noting that the real LIDAR presents a "scattering effect" specifically when the robot rotates and, although not modeled due to its complex behaviour, is not entirely negligible. In any case, the algorithms developed in simulation have proven to be robust enough to withstand this complication, provided that the robot rotates slowly enough, which is necessary to avoid disorienting the Controller.

Virtual Camera

Virtual Camera relies on Python code from the real robot. It is used to test computer vision features in the virtual world. The Unity component is a camera that transmits images of the virtual room via the UDP protocol to localhost, where the Python script is listening. The only significant difference this script has from the one running on the real robot is the "channel" on which it receives the images: in the real robot, it's the serial port where the webcam is connected, while in the Simulation, it's the UDP socket. This

component was crucial in the extrinsic calibration of the camera, meaning the translation of the position of an object in the image with respect to its three-dimensional position relative to the Controller in the virtual world.

Virtual Pose Recognition

Virtual Pose Recognition also relies on Python code. However, in this case, it is an entirely different script from the one running on the robot. The output is nonetheless the same to keep the interfaces between components identical. Its purpose is to test the functioning of displaying humans in the virtual world when the precision of the DepthAI camera is not required. The Python script receives the feed from a webcam connected to the computer and extracts human pose features using a neural network. Afterward, it sends these features to the visualization in the exact format used by the real robot.

Virtual Bump and Virtual Sun

These two functions, which are performed by an IMU in the real robot, are carried out in Unity through the analysis of transformations or collision control—both known data in the virtual environment. The respective messages are then sent via the virtual Jetson.

Virtual Body

The virtual body of the robot is a faithful reproduction of the Physical Avatar "Odile" in terms of degrees of freedom and dimensions. It is possible to move Virtual Odile in space using the same commands sent to the real robot, which is crucial for testing algorithms on the virtual LIDAR and calibrating cameras more rapidly.

The reproduction of Odile's arm also allowed testing the translation of human movements to Virtual Odile, the first Human Translation used during the Digital Week.

4.5. Preparation for the Digital Week

At the beginning of October in Milan, the Milan Digital Week takes place, an annual event dedicated to digital innovation and technology. During this week, conferences, meetings, workshops, and exhibitions are organized to showcase the latest digital and technological trends, involving professionals, businesses, and enthusiasts from various sectors. The goal is to promote knowledge and the dissemination of technology and digital innovation in industries such as manufacturing, art, culture, and more.

One year prior to this work, we participated through the Politecnico di Milano with

our event to present the Physical Metaverse to the public: Connect to the Machine. In this event, external visitors had the opportunity to participate in two experiences. In the first, they wore a VR headset and immersed themselves in a virtual world, the very first Sensory Translation System of the Physical Metaverse. In the second, they could physically interact with Siid, a robot serving as a Physical Avatar controlled through sensory translation. The two experiences were presented separately to visitors to minimize bias when completing questionnaires.

This year, the Physical Metaverse team participated again in the Digital Week, showcasing the new system developed in this thesis. The event was named "Playful Machines," and similarly to the previous year, visitors could participate in two experiences. This time, the Sensory Translation System was the one developed in this thesis, and the robot was Odile, with the modifications made in this work. Additionally, this time, the interaction with the Physical Avatar was structured in a goal oriented activity, also allowing just one Visitor to physically interact and be in proximity of the robot.

In preparation for this event, a tremendous amount of work was required to ensure that a version ready for external users was available well in advance with the timelines set by this thesis.

4.5.1. The Activity

With a meeting for discussion, we decided to clearly outline what we would bring to the event [Figure 4.6]. To effectively and efficiently delineate everything, we focused mainly on the interaction between the Visitor and the Virtual Avatar, considering that Sensory Translation would follow.

The activity would thus take place in two locations simultaneously: on one side, there would be the Controller, who, through the Sensory Translation System, would embody the Physical Avatar, while on the other side, there would be the Physical Avatar with the Visitor.

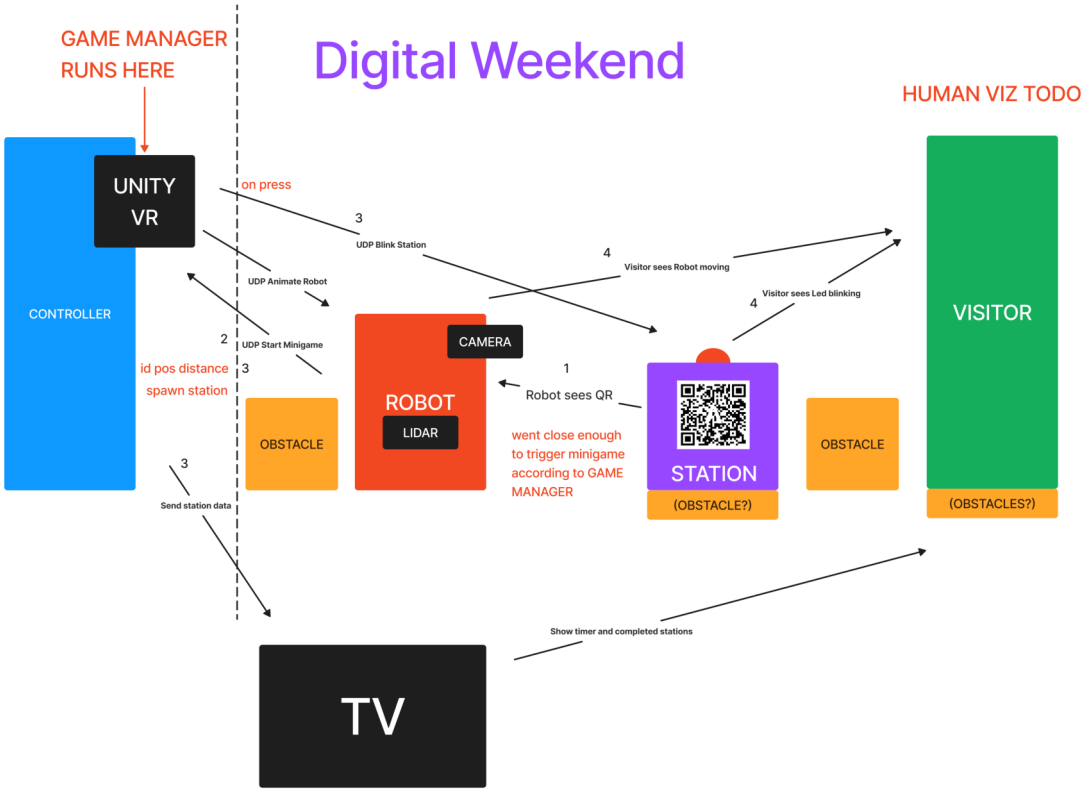


Figure 4.6: Scheme we agreed for the Digital Week event setup.

The selected activity is an "Escape Room," which involves a timed challenge where the two participants must find and activate specific objectives before time runs out. The caveat is that only one of the two participants, namely the Controller with his Physical Avatar, can activate the objectives. This setup shifts the focus towards communication and teamwork between the two, as the Visitor has a clearer perception of the environment compared to the Controller teleoperating the Physical Avatar.

The choice of activity mirrors the experiments conducted in the final stages of the thesis that resulted in the construction of Odile [18]. However, in this scenario, roles are reversed: the human serves as the guide, and the robot follows instructions. This decision aligns with past experiments, ensuring continuity, and, to some extent, establishes a foundation for standardizing the setup for experimenting with systems developed in the Physical Metaverse.

We decided to call the objectives of our experience "Stations." The presence of new "Obstacles" with a well-defined affordance was a new feature in the system that needed

clear definition. Fortunately, experiments had been conducted in the past in anticipation of this feature, so Color Tracking was immediately ruled out. The simplest and quickest solution at the time seemed to be placing QR codes on the Stations so that the robot, seeing them with the camera, could not only locate them in space but also differentiate their meaning. Just like in Odile's experiment [18], not all Stations were valid objectives.

In terms of the interface between Stations and the Visitor, we decided to add an ESP with a LED on top of the Stations, which would be green or red depending on whether the Station was correct or incorrect. This way, the Visitor could intuitively decide which Stations to direct the Physical Avatar towards.

Upon completing a Station, we wanted to provide feedback as explicitly as possible by flashing the LED until it turned off and also execute a programmed movement of the robot's arm. In a sense, we drew inspiration from science fiction droids.

From Odile's experiments, we also adopted the addition of a screen in the room, called the Game Manager, which gives the Visitor an overview of the progress in the game, showing the remaining time and the number of completed Stations.

The next section describes the development process of the features necessary for the event.

4.5.2. The Role of the Simulation

Fundamental requirements for development were the possibility to test the avatar in the environment more agilely and the re-configurability of the avatar and environment. Everything was still to be developed, and potential implementation difficulties were not clearly known. Therefore, all initial tests were carried out in the Simulation described earlier.

Thanks to the Simulation, we could test the developed software features well in advance on physical implementations, minimizing trial and error in the more time-consuming and resource-intensive phases of the project.

The following features were conceived and then tested for the most part in simulation, transitioning to the physical system only in the final tests.

4.5.3. Calibration of the Cameras

The system aims to translate spatial information from two-dimensional camera images to positions in the three-dimensional virtual space. This reconstruction is a well-known problem in the field of Computer Vision, and algorithms exist that, by capturing multiple images of a known object from different angles, provide the transformation matrix from

image space to world space [Figure 4.7]. However, there are several assumptions we can make in our system that simplify the problem. We do not need to faithfully reconstruct all three dimensions, as it is always possible to infer the Y coordinate of each object in our virtual world.

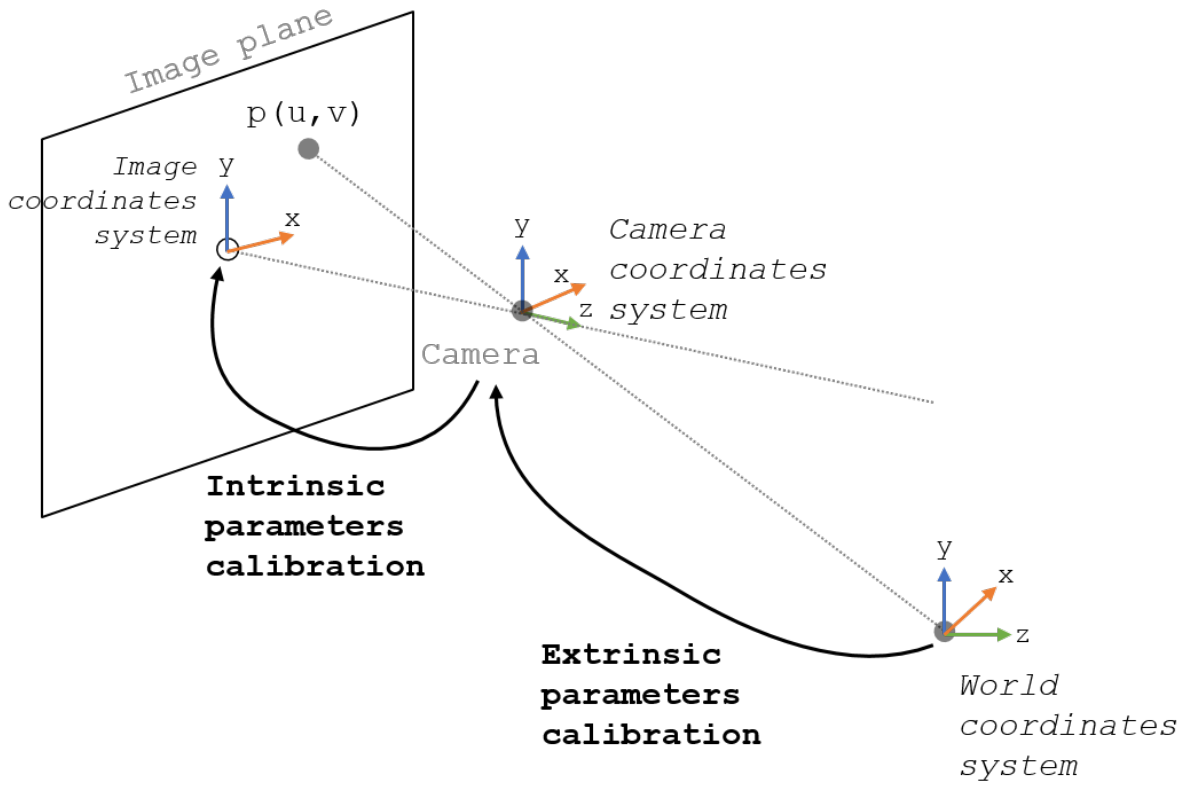


Figure 4.7: Camera calibration coordinate systems [4].

Considering our simplified version of the problem, we decided to expedite the process by manually calibrating using Unity's Inspector tool. We exposed a series of variables as sliders that act on the coordinates of an object in the image to translate them into three-dimensional space. Thanks to the LIDAR reference, we could quickly calibrate both cameras by hand.

For the development and identification of variables and mathematical functions needed for our task, we used Simulation. This allowed us to test numerous angles and configurations with ease.

In practice, the variables needed for calibration are 7. We have offsets on X, Y, and Z, scales on X and Z, and finally, a variable representing the distortion factor due to perspective in the lateral portions of the vision.

With these variables, we have effectively implicitly formulated the linear transformation

matrix required for extrinsic calibration.

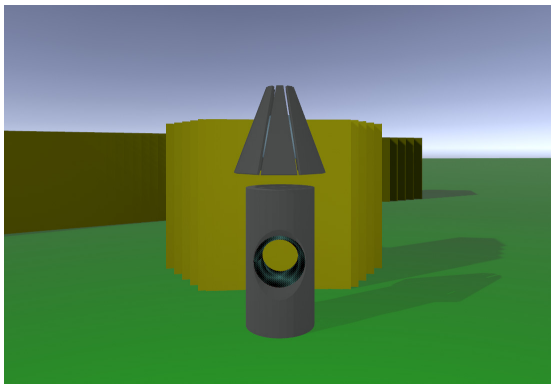
The translation from image space to world space is as follows:

$$\text{PositionXYZ} = \left\langle \begin{array}{c} (\text{ImageX} \times \text{ScaleX} + \text{OffsetX}) \times (\text{DistanceZ} \times \text{PerspectiveCorrection}), \\ \text{OffsetY}, \\ \text{DistanceZ} \end{array} \right\rangle \tag{4.1}$$

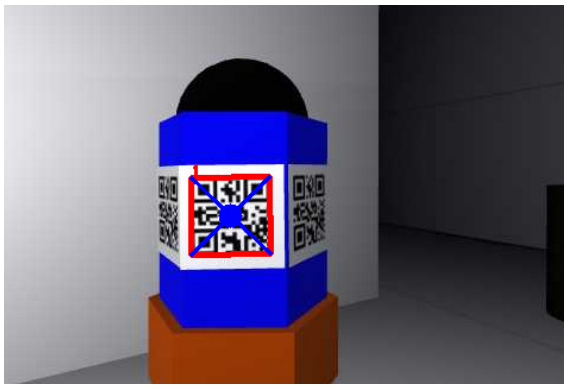
DistanceZ is not calculated in Unity but directly on the robot and differs for each camera. The QR camera finds DistanceZ in meters by performing a calculation on the length of the QR diagonal and the camera's focal length. In contrast, DepthAI calculates DistanceZ using a combination of infrared stereo vision and human pose detection: it takes three evenly spaced points (experimentally these proved to be enough without adding weight to the computations for the analysis of larger portions of the heatmap given by the infrared camera) by making a correspondence between the images from the two cameras, chooses the one with smaller distance value and, through a linearization process on the greyscale value, returns the approximate distance in meters. It is possible to see this result more in detail in Figure 4.17.

4.5.4. QR Code Stations

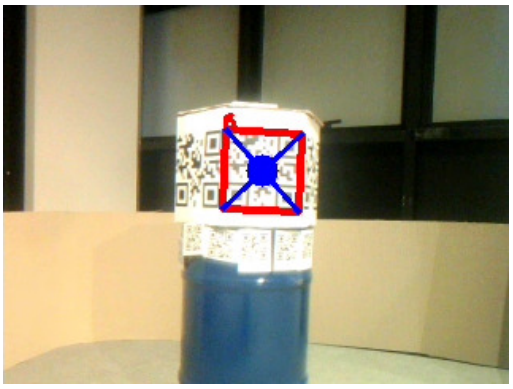
For the analysis of QR codes, we decided, after researching online, to rely on the Python library "PYZbar" [11]. This library takes a video feed as input and outputs the four vertices of the squares that enclose each QR code, along with their encoded values. By processing this information, we created an array containing the encoded value, X and Y coordinates of the QR code centers in the image, distance from the camera, and sent it to VR in the Key:Value format [6]. In simulation, we then addressed the Computer Vision translation issue from Image Space to World Space by exposing the necessary variables for the calibration of a new camera. After a quick calibration of the extrinsic calibration sliders, the results were satisfactory. However, it was possible to track a Station only while it was actually visible to the camera, and given the relatively narrow field of view, it was easily lost. In simulation, exploratory tests were conducted using dead reckoning [See later 4.5.5], and as expected, they worked very well. However, as dead reckoning is rarely effective in the real world, the idea was to use it locally—confined



(a) Station Visualization



(b) Station Detection in the Simulation



(c) Station Detection in the real world

Figure 4.8: Comparison of Station Visualization and Detection.

to short periods—and to gradually fade the Station tracked with this technique. This approach aimed to communicate how the information's reliability decreased over time without confirmation from the camera.

The Simulation allowed us to determine the most efficient physical configuration to ensure that a Station was actually detected when within the camera's field of view. Experimentally, the conclusion was to give the Stations a hexagonal shape with a QR code on each face. This way, if one face was not visible, its adjacent faces would be reliably detected.

Station Interaction

To be activated, the Stations have a cavity that must be first opened and then touched. Functionally, the Stations present a larger invisible spherical area called the "interactable area,": once the Controller enters in it, the cavity opens, indicating that the Station can be activated. There is a second much smaller area corresponding to the cavity called the "activation area," which, once touched, activates the Station. During the Digital Week, we simplified the interaction to the mere proximity of the Controller to the Stations, as

there were visualization scale issues that were later resolved. Subsequently, we introduced a virtual staff held in the user's hand, that replaces proximity activation with actually having to touch the Station in the virtual world, in order to make the interaction more natural and voluntary.

Audio Feedback

An important component of our Stations is the audio feedback they provide when reaching their proximity and when activating them. This is a fundamental component in our study as it introduces the sense of hearing, enriching our sensory translation. In practice, the Stations emit a sound when opening or closing their cavity, i.e., at the entrance to the "interactable area" to draw the Controller's attention to them. Subsequently, they emit another sound upon completion, which will vary depending on whether the Station was correct or incorrect. These auditory elements have proven to be crucial in situations where users did not notice or lost sight of the Stations, still returning them a form of feedback.

4.5.5. Basics of Proprioception

As anticipated previously, this system does not consider proprioception as its primary study objective. However, it inevitably has to develop it to some extent to ensure the success of the Controller's embodiment. In the following, we will discuss two elements that we introduced to initiate a future, more in-depth, development of this system component.

The Proprioception Arrow

Humans are accustomed to see their own bodies, and the complete absence of it was deemed unacceptable in the virtual world. To give users a reference even of just in which direction their body was oriented, we decided to place an arrow at the feet of the Controller for direction and a cylinder to show the body's encumbrance when trying to navigate through closely spaced obstacles. The result is show in Figure 4.9.

Dead reckoning

This module estimates the robot's displacement simply from the translation and rotation user commands. The significant advantage it offers lies in providing immediate visual feedback on the avatar's movement in the most natural way possible, enhancing the Sense of Embodiment.

The dead reckoning module collects input on movement. It found practical application

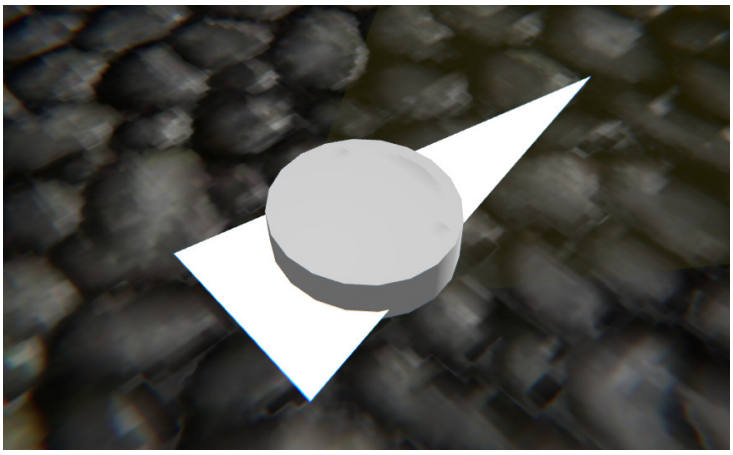


Figure 4.9: The Controller's virtual body.

in the improvements we made during the Digital Week and its most effective use lies in using this information to rotate or translate the floor beneath the user and the sun in the sky. This effect is much more noticeable than the one given by the LIDAR updates and serves to provide immediate feedback on movement input even when the robot moves slowly or by little. It is essential to remember that in the virtual world, the Controller is always at the center and the world has to move around it.

In the past, we rotated the sun using information gathered from an IMU on the robot. However, to streamline our system and make room for other features in this rapid development phase, we opted for dead reckoning as it is effectively functional in practice. In the future, it would be wise to consider re-implementing the IMU to increase sensory information, especially with a view to incorporate additional features.

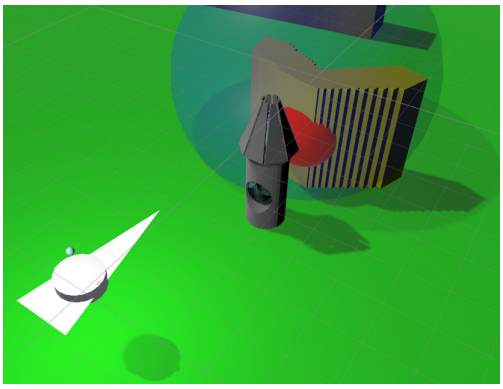
There is one last important way in which we leverage dead reckoning. Recognizing the significance of the visual feedback provided by the smooth movement of the floor, and how it seemed inconsistent with the "jerky" movement of the LIDAR pillars, we decided to apply the same principle to the LIDAR. The implementation involves continuously translating or rotating the pillars according to the user's input, just like with the floor. However, in this case, we have the subsequent reception by the LIDAR to "recapture" our dead reckoning, effectively closing this feedback loop. The achieved effect is a perceived movement that feels much smoother. Additionally, it is much easier to realize when you are stuck in a certain position because you will feel like you are moving toward the walls, but they will continuously move back. An interesting result that can be observed when the robot disconnects from the system is that you will be able to navigate freely (exclusively in the virtual world) in the last "snapshot" provided by the LIDAR.

An implementation challenge encountered with this feature, only partially addressed by

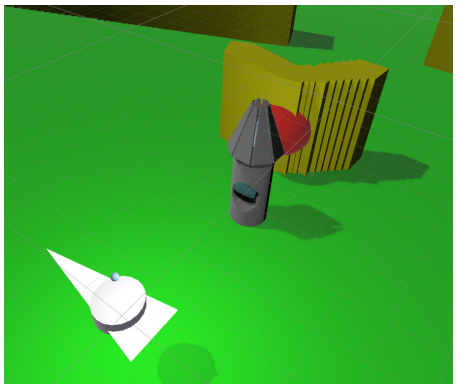
introducing acceleration in dead reckoning, is that the robot has its dynamics in movement. Therefore, it will not always move at the speed we might infer from the input with a simple first approximation. Moreover, as the battery charge decreases, the robot will tend to move more slowly. In the future, one could consider incorporating all these features into our system calculations, studying the proprioception of our robot at a much deeper level, delving into its electromechanical dynamics. Or, more simply, mount a rotary encoder on the robot to measure actual odometry.

4.5.6. LIDAR Tracking

Given the limited applicability of dead reckoning in the real world due to its open-loop nature, we decided to switch to a different method in order to track our Stations: using the 360-degree information from the LIDAR. The insight came from observing that when a Station is detected by the camera, it inevitably overlaps with a group of pillars detected by the LIDAR, and these pillars represent the physical shape of the Station. When the Station is "lost" by the camera, the pillars are easily traceable by eye. As long as one remembers which group corresponds to which Station, it is possible to keep track of the Stations based solely on LIDAR information. What we are discussing probably



(a) LIDAR Tracking with Camera detection



(b) LIDAR Tracking without Camera detection

Figure 4.10: Visualization of LIDAR Tracking algorithm.

makes those working with mobile robots and LIDARs wonder why we didn't implement SLAM (Simultaneous Localization and Mapping). The reason is that the "break-even" point between a handcrafted technique and the use of a SLAM algorithm was not yet in sight. Given the set deadlines and the original idea of minimizing the alteration of sensory information by the avatar, we decided to stay consistent with our initial choice, as changing technology would have required a lot of refactoring, among the other things.

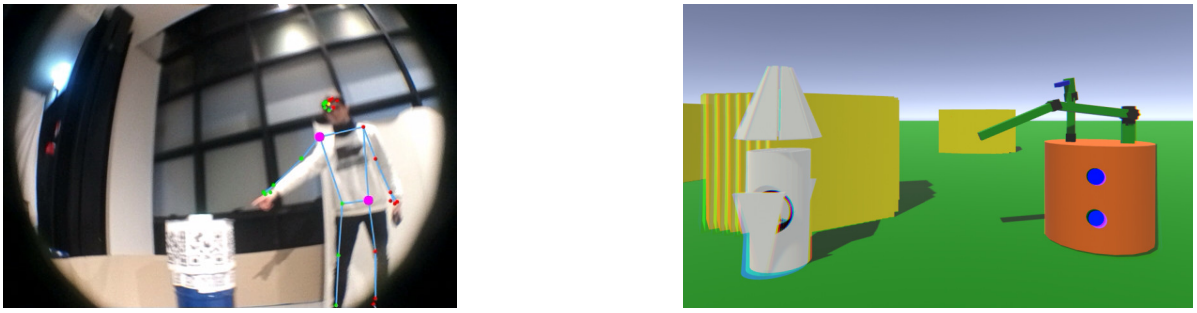
Returning to LIDAR Tracking, the insight about integrating LIDAR information proved entirely valid and led to the development of an algorithm capable of tracking Stations by merging information from camera and LIDAR. The results provided more stable and frequent Station position updates than those obtained solely from the camera. As explained earlier, individual Stations have multiple QR codes, and as one or more of them can be detected at any given moment, the visualization used to appear quite shaky near the Stations. With LIDAR tracking, it became possible to display the Station simply at the center of the physical obstacle it represented. In Figure 4.10 you can see the algorithm in action. The white triangle represents the direction in which the Controller is oriented: the larger transparent sphere represents the area where the camera indicates the Station should be. The algorithm selects the largest group of pillars intersecting with the sphere and places the placeholder for the Station, represented by the red sphere, at its center. The actual Station is then displayed with an offset from the obstacle in the direction of the Controller. When the robot moves, causing all the pillars to shift, the algorithm continuously recalculates the intersecting largest group of pillars and its midpoint to keep the Station centered. In essence, this algorithm introduces the much-needed memory component to the LIDAR visualization.

At times, it happened that in the absence of confirmation from the camera, the algorithm would start to "drift" the Stations along the walls. To overcome this problem, we added an invisible area called the "invalidation area" inside which a Station can be visualized only if it is actually detected by the camera and is hidden in the absence of reconfirmation from it.

4.5.7. Human Translation

Outlining the translation from a detected human to a non-humanoid Entity required another collaborative meeting. We decided to reintroduce Odile himself as a companion in the activity, this time in a virtual form, serving as a guide as it did in the previous tests. Apart from the translation in space, which we wanted to keep consistent with what the DepthAI camera observes, we had access to the degrees of freedom of the head, where the camera was located in the real Odile, and the arm. Mapping the head turned out to be immediate, simply by tracking the Visitor's gaze and translating it to the virtual Odile's gaze.

However, determining how to command the arm was much less immediately solvable. We needed to decide how to translate the pose of the real person to Odile's arm while introducing the least possible bias. After some less-than-satisfactory attempts with direct



(a) Camera detected pose (b) Visualization

Figure 4.11: Virtual Odile Human Translation.

kinematics of the human's right arm, we decided to try inverse kinematics. The position of the hand in space relative to the person would indicate that of Odile's arm's end effector relative to the robotic body. The issue of bias remained, whether to choose which arm or to use both, and if so, how. At least in this phase with limited time, we simply opted for the right arm.

Inverse kinematics yielded promising results from the outset. However, it was clear that much more time than was available before the Digital Week would be needed for a truly effective calibration and even more time to study translations that were meaningful in our research. So, for the time being, we settled for something that was at least functional.

4.6. Playful Machines

The event took place over two days, and we decided to offer participants a questionnaire, which remained essentially unchanged for the final experiments and will thus be analyzed later in the Experiments chapter of this thesis. In this phase of full development we focused on improving the system, allowing ourselves to make changes to the system during the event. Therefore, we didn't provide a fixed base for participants' responses. The purpose of the questionnaire in this phase was just aimed to guide us through the development with valuable feedback from real users.

Tuning the system during these days proved to be of enormous importance because it provided us with insights that we might not have gained otherwise.

In the upcoming sections, we will first outline the setup for the experiment, briefly detailing the physical arrangement of the room and the setup of the virtual reality Stations to leave a more in-depth description for Chapter 7 about the final experiments. Additionally, we will describe the physical implementation of the Stations mentioned earlier. Then we

(a) Camera detected pose

(b) Physical experience

Figure 4.12: Setups of the virtual and digital experiences at Digital Week.

will proceed with descriptions and considerations about the two days of Playful Machines.

4.6.1. Room Setup

For the space in which to conduct the activity, we delimited an area of approximately five meters by eight with sheets to provide a well-defined border for the LIDAR positioned on the robot within the playing area. Subsequently, we organized our space by placing five Stations, three of which were correct and two incorrect, along with the Game Manager screen. Figure 4.13 reports the room layout for the event that we reproduced in the Simulation, in order to test its flow and Stations' detectability.

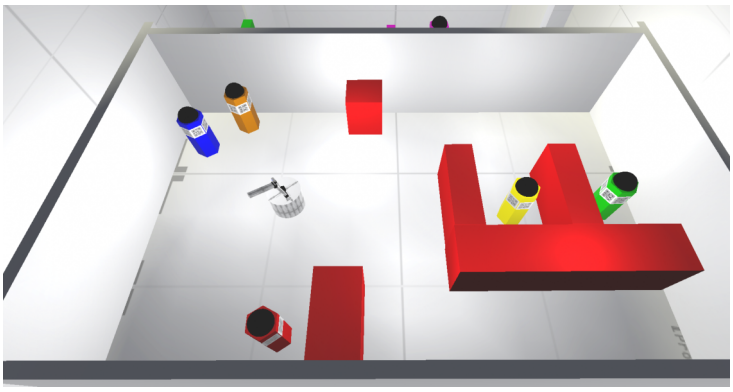


Figure 4.13: Digital week physical room setup reproduced in the Simulation.

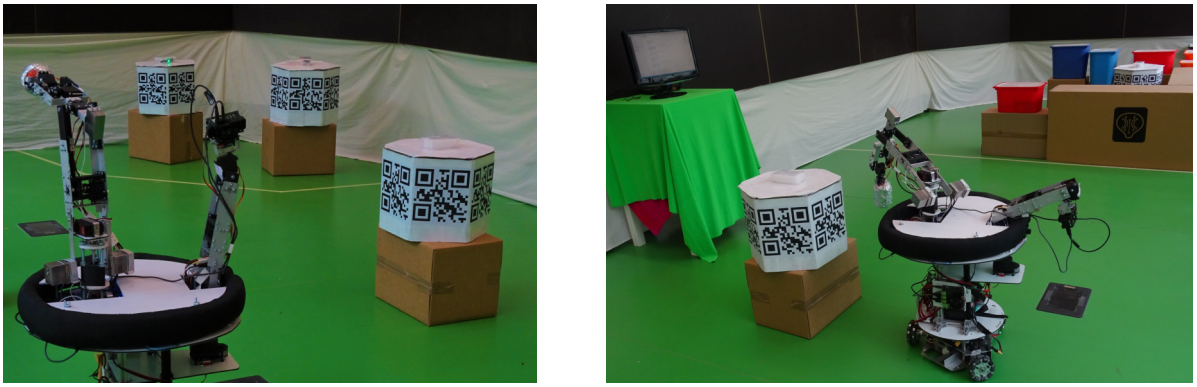
The first correct Station was positioned very close to the entrance (bottom left corner in Figure 4.13), where the Physical Avatar would first encounter the Visitor. Following this, two other Stations, one correct and one incorrect, were placed not far from each other.

Lastly, we created a kind of "F" starting from the wall with obstacles, where the Visitor would have to guide the avatar through to the final corridor containing a correct Station, avoiding the first corridor where the incorrect Station was located.

4.6.2. QR Code Stations

The Stations serve as the focal points representing the goals in the goal-oriented activity. Structurally, they consist of hexagonal cardboard prisms with QR codes printed on A4 sheets, affixed to each face of the prism.

The hexagonal shape is chosen strategically, considering the camera and algorithm's QR detection angle. Beyond a certain angle, code detection becomes challenging. To address this, we introduced redundancy by placing QR codes on multiple faces of the Station.



(a) QR Stations and Odile (b) Other side of the maze and Game Manager screen

Figure 4.14: QR Stations and Digital Week Physical Maze.

In a subsequent phase, smaller mapped QR codes were added alongside the hexagonal Stations. These smaller QR codes corresponded to the respective Stations, facilitating detection by a camera positioned closely, especially when larger codes were temporarily cut in the view.

The Station design was entirely driven by the Simulation. Prior to constructing all Stations, a prototype validated the design, proving its efficacy. This initial prototype remains in use alongside the other Stations in the final system.

To convey information to the Visitor regarding correct and activated Stations, an ESP with LED indicators is positioned above each Station. This implementation serves as a visual cue, allowing the Visitor to discern the status of each Station based on the illumination of the corresponding LED on the ESP module.

4.6.3. Day One

Considerations The first day started with numerous hardware problems that forced us to delay the beginning of the event by a few hours. The most significant issue was the breakage of the pan servo motor of the camera due to the collapse of the portion of the neck that we did not like to actuate. In an attempt to identify the problem (we couldn't rule out software errors), we unfortunately also broke the tilt servo motor, to which we sent the pan signal for diagnostic purposes. It found itself working in a range for which it was not structurally implemented, and after numerous unnoticed stalls, it fused. The lesson learned was that too much haste (some people were already arriving and asking when we could start) can lead even a team of a graduate student, a post-graduate student, and a professor to make gross mistakes.

In any case, in the early afternoon, the system was back in operation, and we could start the activities.

The delay due to breakdowns prevented us from making the necessary calibrations on the robot in the real environment scheduled for the morning (for organizational reasons, it was the first useful moment when we could perform them). So the first Visitors unfortunately couldn't try the best version of the system. Nevertheless, they were generally satisfied.

The choice not to build the Sensory Translation application to run on the headset but to execute it in Unity's play mode through a serial connection with the computer proved to be fundamental. This choice allowed us to perform calibrations during user experiences using the game engine's inspector without disturbing them, significantly improving the user experience within three or four trials.

The most problematic aspect of the system was the latency in transmitting the pose from the human pose recognition camera, making it difficult to follow Visitors who moved particularly quickly.

Fixes and Improvements Considering the feedback received in person from users during the day and the observations made about their behavior and performance in the virtual world, we deemed it necessary to make essential changes to improve the user experience for those wearing the headset.

The first and most substantial change was adding a texture to the floor. Choosing the right texture brought us to also change the pillars' color for aesthetic motivations. Considering that the Controller doesn't actually move in the virtual world, but it's the world that changes, giving a sense of movement, we needed a way to translate and rotate the floor

to provide a feeling of movement. Therefore, we decided to use input from joysticks as anticipated earlier when discussing about dead reckoning. Precision in this effect was not as crucial as responsiveness to user commands. The goal was to provide immediate feedback to motion controls, often not very visible from the representation of the LIDAR, which could appear slow or change only slightly at times.

Other modifications improved the overall control of the robot, adding a dead zone to the analog stick and removing the "watchdog" that, in Arduino, periodically stopped the base during movement to avoid continuous movement in case of disconnection.

4.6.4. Day Two

Considerations For the second day, there was a general sense of optimism, and it proved to be well-founded. Thanks to the improvements made the night before, users were much better able to navigate the virtual world, and overall, participants' appeal toward the experience increased. Calibrations and improvements continued during the use of the system on this day as well, always without affecting or distracting users from the experience.

The latency issue with the pose recognition camera persisted, although in more severe cases, we tried to minimize it by manually moving the virtual Entity from Unity's inspector, relying on points displayed by the LIDAR, similar to LIDAR tracking for Stations but without a dedicated algorithm for this task.

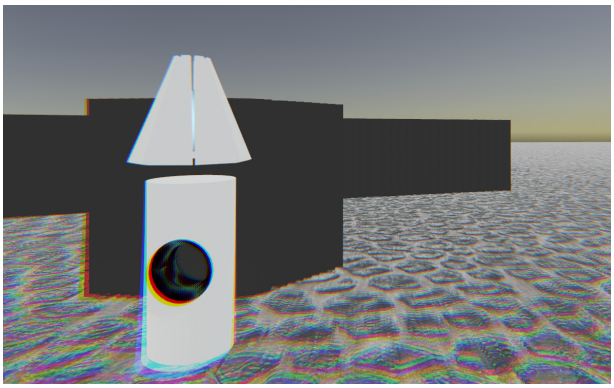


Figure 4.15: New graphics and floor texture after the Digital Week.

4.7. Post Digital Week improvements

After the Digital Week, we had clear points to work on to improve our system: better interaction with the Stations and a more stable visualization of the Entity in space.

Additionally, we found that the Odile robot was too complex for our system, and we were not fully utilizing its features. Therefore, we switched to the Blackwing robot, adapting it to accommodate all the new hardware and the including the pan and tilt mechanism of the camera.

4.7.1. Improved Station interaction

Realizing that interaction solely based on the proximity of the Controller to the Stations was not only unintuitive but also led to unintentional activations, we revisited this aspect. We decided to introduce a virtual staff into the experience that the Controller would hold in their right hand, and at its end, there is a luminous sphere.

When the staff is in the interaction area with the Stations, they open. Subsequently, if the Controller inserts the staff into the Station's cavity, the sphere will flash, and if kept long enough, the Station activates. We introduced two new important components during this holding phase before activation: the first is a movement in the robot to indicate to those in the physical room that the avatar is actually interacting with the Station, while the other is a haptic feedback that vibrates the joystick in the Controller's hand. The new robot movement is detailed in Figure 7.8, towards the end of this thesis.

Haptic feedback

The introduction of haptic feedback has fundamental relevance in our system: it represents the translation of another new sense. Haptic feedback implemented in this system consists of continuous vibration as long as the user keeps the staff in the Station's cavity, and it ends if the user removes the staff or if enough time passes and the Station activates. It will be worth studying this aspect more thoroughly in the future as we have only touched on it in the final stages of our project.

4.7.2. Person Tracking

Similarly to what we did with LIDAR Tracking (see Subsection 4.5.6), we decided to integrate information from the LIDAR to stabilize the position of the Entity in the visualization. Once again, we considered that the Visitor, like the Stations, is perceived by the LIDAR as an obstacle, and therefore, in the virtual world, it will be shown with corresponding pillars.

We made a strong assumption to develop an effective algorithm: the Visitor is the only element that moves in the physical maze. Based on this, we developed a system that

detects which pillars have undergone movements beyond a certain threshold, and at least in simulation, this intuition has proven to be very effective. The complexity of developing this algorithm led us to think in more visual terms: in the context of this work, we developed a technique that we named "Visual Debugging." Before continuing with the description of Person Tracking, we focus on this technique, as the concepts it introduces will greatly simplify the presentation of the algorithm.

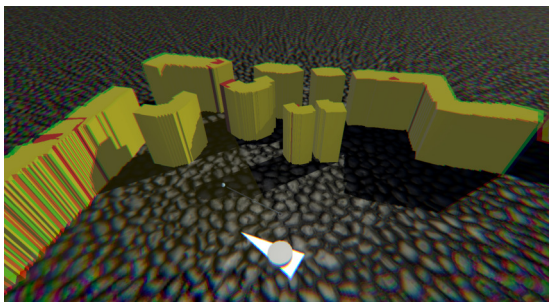
Visual Debugging

This technique was born when we realized that it was not effective to simply print the results of the calculations performed behind the scenes by our algorithm to the console, as the data were too complex for a purely numerical and textual representation. Unity, among its tools, has the Inspector: this allows you to view and modify the values of program variables in real-time during execution. However, even this approach had its limitations because we needed a much more intuitive representation if we really wanted to understand how the developed algorithm was working and how to improve it.

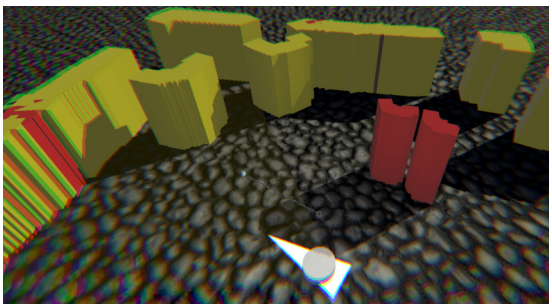
Therefore, leveraging Unity's collision system, which triggers certain events when meshes intersect, we decided to take a highly visual approach. Initially, in the debugging phase, the pillars of our visualization were all yellow. We then introduced new colors to indicate their current state, semantically representing what they were representing in our algorithms. The new colors introduced are three: red, blue, and orange.

- **Yellow** represents default pillars. [Figure 4.16a]
- **Red** represents pillars that have undergone a **movement** beyond a certain threshold. [Figure 4.16b]
- **Orange** shows which pillars correspond to the position of the **Visitor**. [Figure 4.16d]
- **Blue** indicates pillars that are still providing information for a **Station**. [Figure 4.16c]

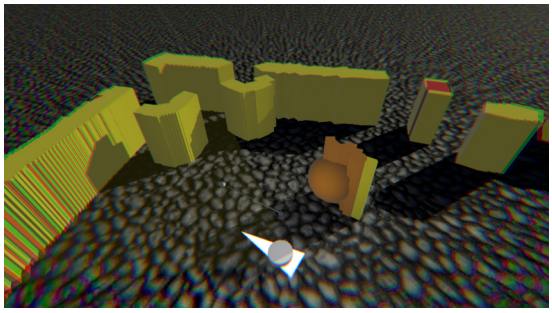
We make a consideration about this technique: when developing a system, its first users are the developers themselves. And developers, like the users who come later, are human beings. For this reason, it is important that those who develop the system are the first to have an interface as understandable as possible and, if necessary, even an interface dedicated solely to them.



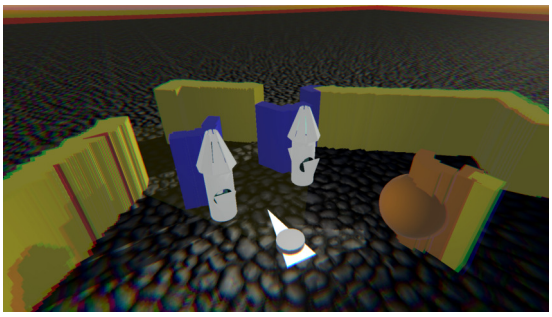
(a) No movement, just LIDAR noise



(b) Person moves towards Controller



(c) Person got tracked



(d) Stations detected

Figure 4.16: Visual debugging.

Person Tracking continuation

Let's now return to Person Tracking. We will describe the algorithm in terms of Visual Debugging, primarily referring to the colors of the pillars. By default, the environment consists of yellow pillars. When they undergo a movement beyond a certain threshold, influenced even by noise, this movement is shown with red pillars. When a sufficiently large group of adjacent red pillars is formed, according to parameters adjusted with sliders during a calibration phase, an orange sphere representing the position of the Visitor—referred to as the Person Tracker—will be placed on this group. Since the sphere, in turn, colors the pillars it collides with in orange, we have practically introduced another memory component into the system. We have also assigned "scores" to each color. Thus, when there are differently colored pillars under evaluation, it is not really the size in terms of the number of pillars in a group that decides our sphere's positioning, but the total sum of their scores. Opening our project in the Unity editor it is possible to analyze in detail the calibrated values and make sense of them inside their context. However, here we provide an explanation purely in terms of discourse and intuition.

Experimentally, we noticed that the memory component (orange) is more important than the movement component (red) in tracking the Visitor. Therefore, in terms of score, the orange pillars will weigh more than the red ones.

However, this system required additional components to work once we moved from simulation to LIDAR and real measurements. For this reason, we introduced a new concept of maximum jump distance for the Person Tracker so that it would remain more stably in one position without jumping elsewhere simply due to noise. Another component we added was the dynamism of this maximum jump distance: we decided to introduce a concept of the "certainty" of the tracked position that would increase or decrease the maximum jump distance in a way directly proportional to the total score of the pillars in the chosen group of the algorithm.

Closing the loop

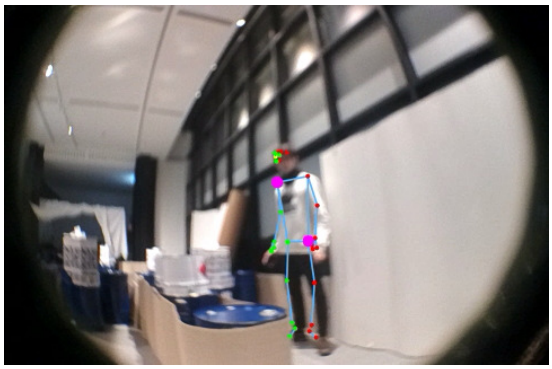
In the development of Person Tracking we had completely set aside the positioning of the Entity in the virtual world by the camera, because its high mobility due to the pan and tilt mechanism returns spatial position information that is too unstable to be used effectively. With pure Person Tracking, the system was tracking very well a person as they moved in the environment; however, when they stopped, the tracking was quickly lost.

We decided to merge the two approaches very similarly to how we did for the Stations: the camera would roughly indicate the point where the Visitor is, while Person Tracking would take care of a more stable and precise placement given the camera information.

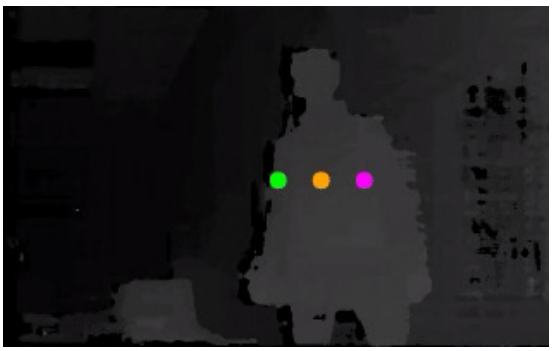
In terms of the algorithm, what happens is that when the camera detects a person, we significantly increase the scores of the pillars in the field of view so that the probability that the algorithm calls the Person Tracker on a group located in this area is greatly increased. Also, every time the camera detects the person, it positions the Person Tracker according to the coordinates found with the extrinsic calibration, and then immediately calls it to the nearest valid group of pillars. To visually smooth the result, we made sure that the Entity model moves gradually and not instantly toward the person tracker. This achieves a more soft movement. Additionally, we also hid the pillars corresponding to the position of the Entity so as not to make it appear as an obstacle in the eyes of the Controller.

As with the Station tracking algorithm, we introduced an 'invalidation area' to avoid situations where, in the absence of confirmation from the camera, Person Tracking would position the Entity along the walls.

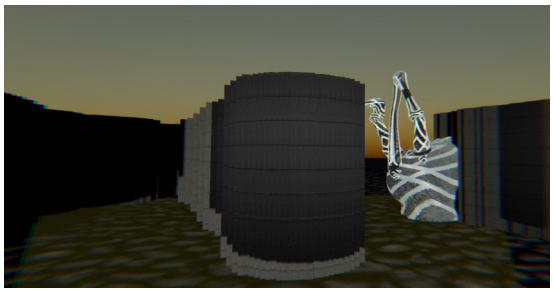
To wrap up, in picture 4.17 it is possible to see the state of pose detection that we would bring to the final experiments.



(a) Detected pose



(b) Heatmap of depth perceived by camera



(c) Result inside the visualization

Figure 4.17: Results of post Digital Week improvements.

4.8. X-Cities Exhibition

Following the impactful showcase at the Milano Digital Week, the subsequent and final opportunity for public engagement with our system presented itself at the X-Cities exhibition.

The X-Cities exhibition offered a forward-looking perspective on the integration of physical and digital spaces, shaping the identity of future urban environments. This immersive showcase investigated the profound impacts of the digital revolution on our cities, lifestyles, and perceptions of space. It featured two main trajectories: the digitalization of the physical world, presenting digital replicas and models of buildings as intelligent, sensor-equipped entities, and the materialization of the digital world, where virtual realities and digital identities are grounded in physical contexts. The exhibition also included interactive installations by AIRLab, our laboratory, allowing visitors to engage with alien creatures and explore the interplay between human forms and digital metamorphosis. Designed sustainably and inclusively, X-Cities catered to a diverse audience, combining narratives, videos, avatars, and participatory experiences to envision the symbiotic future of urban and digital realms.

Our contribution centered around two interactive experiences, serving as the culmination of this thesis. Each experience was crafted to engage and immerse visitors, providing a tangible glimpse into the potential of human-robot interaction and the translation of human expressions in a shared space. Our dedicated space in the exhibition was named "First Contact". The experiences were the following:

1. **The Mirrors** The second experience drew inspiration from our theoretical framework, specifically the concept of a shared space where Human Translations interact. In this setup, two large screens oriented vertically served as mirrors, reflecting how the visitors' images were translated in real-time.

2. **The Mazes** At the core of our presentation was the execution of a goal-oriented activity between a Visitor and a Physical Avatar in a combination of digital and virtual Mazes. This avatar, a teleoperated robot, would be controlled by a Controller using the Sensory Translation System in virtual reality.

The significance of the second experience extended beyond the immediate exhibition. It offered a valuable opportunity to develop and assess novel Human Translations with real users. By observing how individuals expressed themselves in these virtual mirrors, we gained insights into the adaptability and intuitiveness of our system. This user-centric approach contributed to the refinement of our technology and laid the groundwork for potential future enhancements. The exhibition also pushed us to revisit the design of our creatures as shown in Figure 4.18.

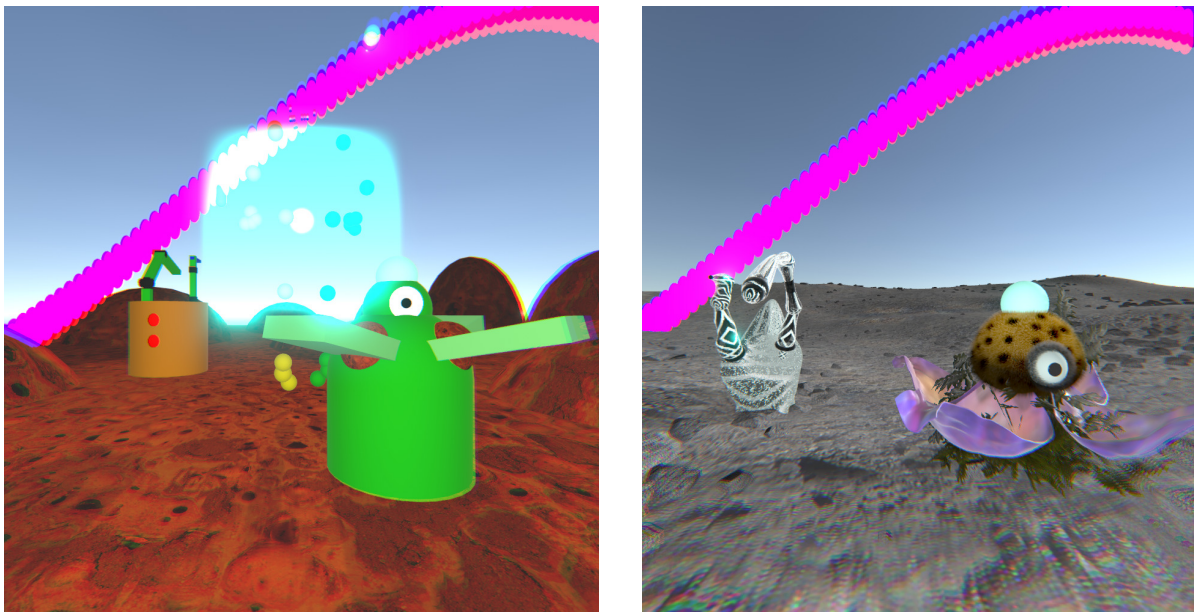
4.8.1. The Mirrors

The design and implementation of the interactive mirrors, or Mirrors, represented a fundamental element within our "First Contact" exhibition at the X-Cities exhibition at Politecnico di Milano. The main goal of this experience was to create a fluid and intuitive environment, where visitors could interact freely with technology, experimenting with the real-time translation of their physical expressions in the virtual world.

Overview

Each Mirror consisted of three key elements, along with a computer dedicated to running the software. Figure 4.19a shows two of these Mirrors side by side. Their components are:

1. **Vertical 16:9 Monitor** Positioned vertically, this display creates an interface that is immediately accessible to visitors, inviting them to interact spontaneously.



(a) Physical Metaverse of Madness (b) Mirror X Cities Models

Figure 4.18: Comparison before and after Entity design update.

2. **Full HD Webcam** Mounted above the monitor, the webcam captures facial expressions and gestures of participants in high definition, feeding the human pose detection process.

3. **ESP32 with Ultrasonic Sensor** Placed in front of the monitor, the ESP32 with an ultrasonic sensor detects the distance of visitors, helping to adjust the experience based on their position and presence. The structure to hold the ultrasonic sensor in position was constructed with a combination of Lego and 3D printing [Figure 4.19b.

4. **Computer and Software** The system is powered by a dedicated computer, connected to the monitor via HDMI. This computer runs the human pose detection software, contributing to the processing and instant transformation of physical actions into virtual inputs.

The Mirror Experience

The mirrors were designed to offer a seamless experience. Visitors could approach, stand in front of the display, and start their interaction spontaneously. The experience concluded after a preset time or when participants decide to step away.

(a) Mirrors, not active in this picture.

(b) Lego mount with ultrasound sensor.

Figure 4.19: X-Cities Mirrors setup and Lego ultrasonic sensor mount.

The Implementation

In contrast to the implementation of Human Translation on the robot, relying on neither a camera nor LIDAR for the distance of the detected person in the mirror presented a challenge. To address this, an ultrasonic sensor was introduced to provide distance measurements.

After several attempts, noise from the sensor was effectively filtered, providing a stable distance value for the mirror. Moreover, the sensor's detection cone proved to be sufficiently wide to cover the entire area where we wanted a person to "see themselves." While the camera used the same version employed in simulations and much of the development process, it is essential to note that, as the DepthAI camera was not used, the pose detection quality was slightly lower. Nevertheless, this solution proved entirely effective for the intended experience.

In detail, the mirror's operation was as follows: two Unity applications for the mirrors were executed on the computer, and through video ports, they were displayed on vertically positioned monitors. Simultaneously, on the same computer, two instances of pose detection in Python were executed. These instances received distance data from ultrasonic sensors connected to ESP32 through a UDP protocol. The distances were then appended to the pose landmarks. Subsequently, via UDP, the pose data was sent through localhost to the respective Unity applications.

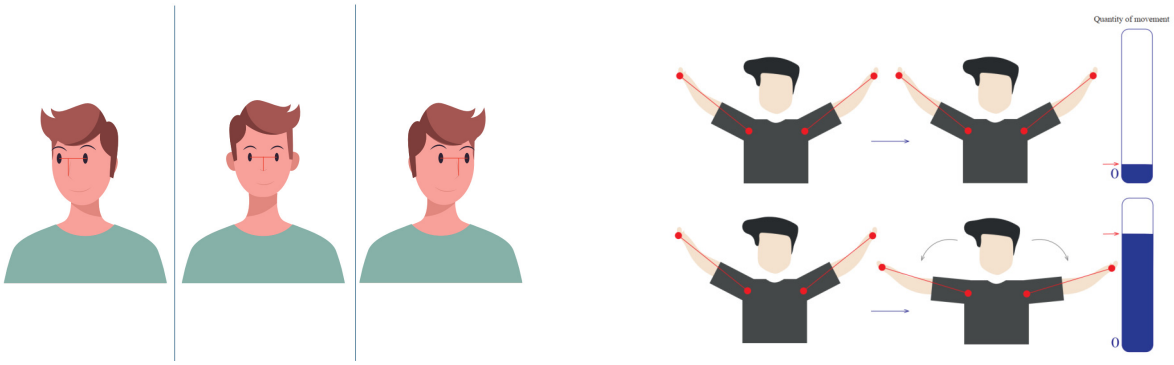
This architecture facilitated the real-time interaction between the mirrors and the pose detection system, creating a seamless and immersive experience for the participants.

The Human Translation

The application, regardless of the displayed Entity, extracts a series of data from pose detection that are then translated and used by different Human Translations.

Recovering work on pose detection that was done in the past project [6], we focused on what we deemed essential for this phase, namely: the person's position and rotation, gaze direction, and quantity of movement.

The "quantity of movement" is a concept introduced in the previous work and is a parameter indicating how much a person moved their head and arms in a given time frame. A new tool introduced in this work are the hand trackers, whose value is summarized in the three dimensional position vector of the right or left hands.



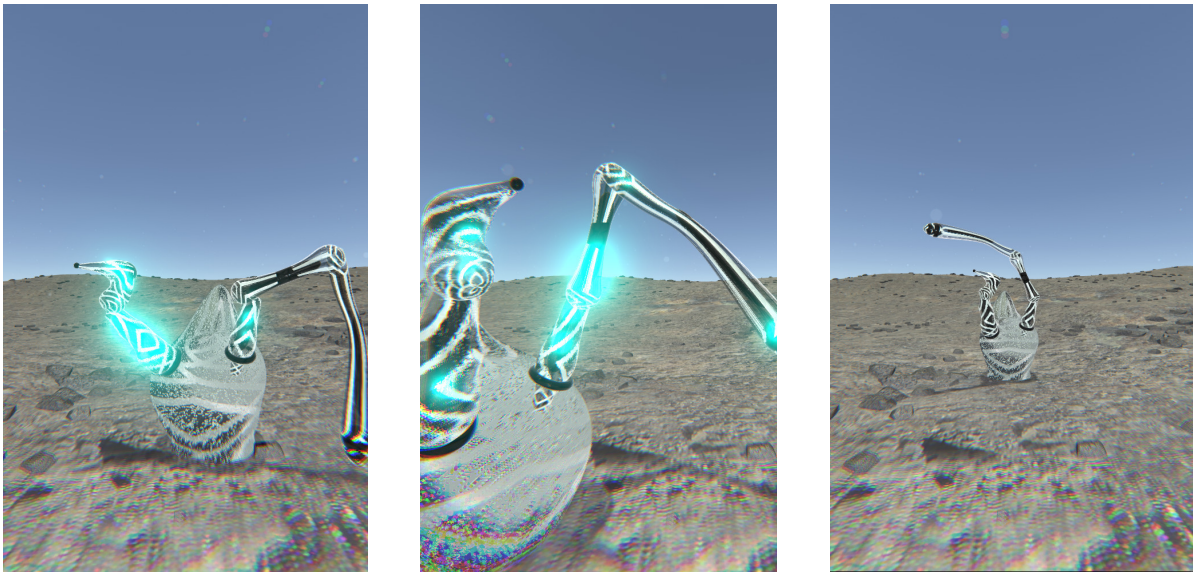
(a) Gaze direction (b) Quantity of movement

Figure 4.20: Gaze direction and quantity of movement [6].

We decided to present three different Human Translations for the exhibition: Odile, Siid, and Evangelion.

Odile

Odile has been entirely redesigned to make it more captivating and resemble an alien creature rather than a robot. The new design is show in Figure 4.21. In its new form, it appears to be made of rough-surfaced glass with gray/white and black stripes. Like before, it still has two arms, one ending with the creature's head and the other, longer, used for pointing. To map from a person to the creature, we connected the right and left arms to the respective counterparts of the robot. The arms move with inverse kinematics, so that the end effectors have a relative position end effector-joint base that is simply a scale of the relative position person hand-person hip. Odile's head once again reflects the orientation of the person's head. Rotation and translation in space correspond to those of the person relative to the mirror.



(a) Odile bending its neck. (b) Odile close up, looking up. (c) Odile far away.

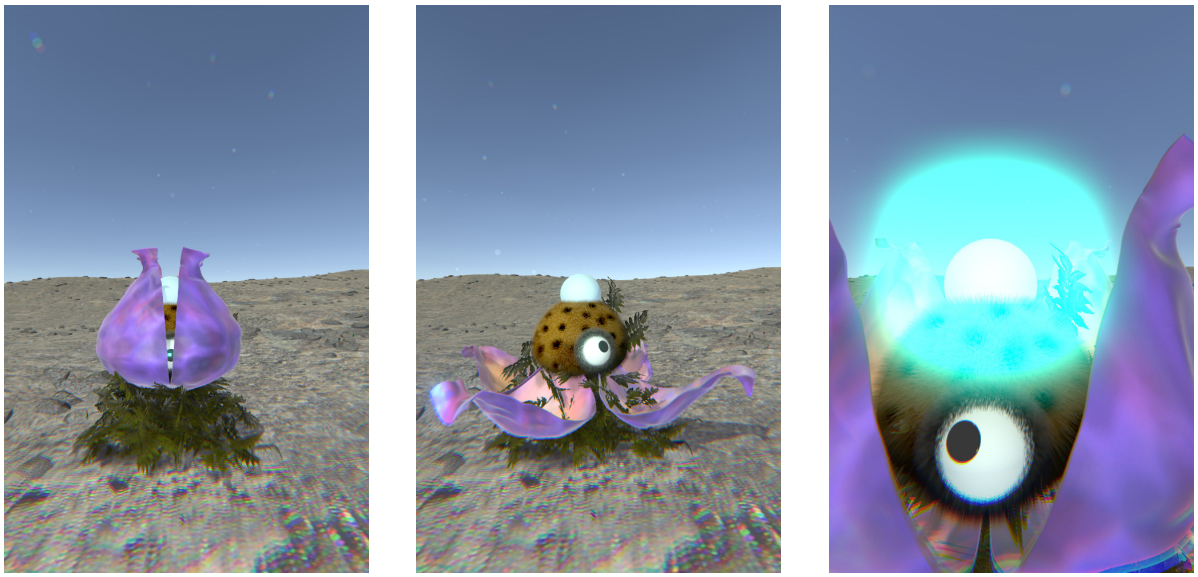
Examples of Odile's features. The short and long arm bend with inverse kinematics on the person's left hand and right hand respectively. Odile's head at the tip of the short arm is mapped to the person's gaze direction.

Figure 4.21: Odile Human Translation.

Siid

As a new Human Translation, we decided to introduce Siid, the robot that played the role of a Physical Avatar in the Digital Week of the previous year. Siid's new design is show in Figure 4.22. Like the real robot, its structure resembles that of a plant consisting of a bush topped with four large petals. Opening these petals reveals a hairy body with

a large eye, above which sits a luminous sphere. Therefore, it is possible to control the opening of the petals, the orientation of the eye, the brightness of the sphere, and, like Odile, movement and rotation in space. For the mapping, we connected the quantity of movement to the brightness of the sphere and the gaze direction to the orientation of the eye. We tested two alternatives for controlling the petals, determined by the position of the hand trackers: the average distance of the hands from the nose, covering the nose results in closed petals, or the average distance of the hands from the central axis of the person's body. We decided to experiment with both in the mirrors to see how people would behave. Once again, the rotation and translation of the Human Translation are determined by the actual movements of the person.



(a) Siid hiding behind its petals. (b) Siid looking up to the right. (c) Siid doing physical excercise.

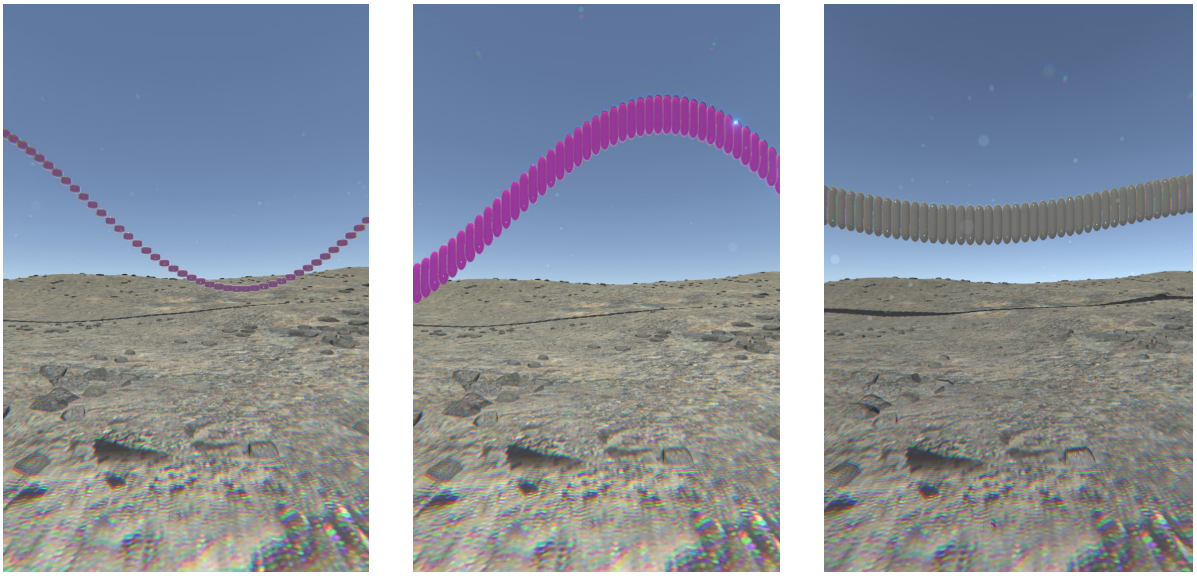
Examples of Siid features. Petals opening is mapped to position of person hands from nose, eye reflects the person's gaze direction, the ball blinks if quantity of movement is high.

Figure 4.22: Siid Human Translation.

Evangelion

Evangelion is a Human Translation that differs from the others due to the almost complete absence of spatial information. This Human Translation is show in Figure 4.23. It completely removes information about the person's rotation and, for translation, retains only the distance from the camera. It consists of a wave composed of numerous consecutive fuchsia capsules, aiming to push the limits of Human Translation to see how much can be reformulated while still conveying meaningful information to an external observer.

As a wave, more specifically a sine wave, among the comparable components are the amplitude, and there is also the speed at which the capsules that compose it move, their thickness, and the saturation of their color. The amplitude is determined by how centrally the observer is positioned relative to the vertical axis of the camera. The speed of the capsules is given by the quantity of movement, the thickness by how close the observer is, and the color is more saturated if the observer looks straight at the camera.



(a) Evangelion far away, looking slightly to the side.

(b) Evangelion moving a lot while looking at the camera

(c) Evangelion while the person is leaving

Examples of Evangelion's features. The color is more saturated if the person looks straight into the camera. The wave is more ample when the person is centered in front of the mirror. The wave also moves faster when quantity of movement is high.

Figure 4.23: Evangelion Human Translation.

4.8.2. The Mazes

The Mazes are the second part that we presented at the X-Cities exhibition, as well as the final work of this thesis. In Chapter 7 we provide a comprehensive description of them, followed by the final results.

5 | The Hardware

Figure 5.1: Hardware used for sensory translation.

5.1. The Physical Avatars

This project evolved across three distinct Physical Avatars as they became available through the efforts of other collaborators in the Physical Metaverse project. The modular and portable nature of this project, aligned with the Physical Metaverse philosophy, facilitated seamless hardware transitions. The Physical Avatars employed in this project are: "The Robot", "Odile", and "Blackwings". This adaptability allowed for the exploration of diverse robotic platforms, enabling the integration of updated hardware components without hindrance.

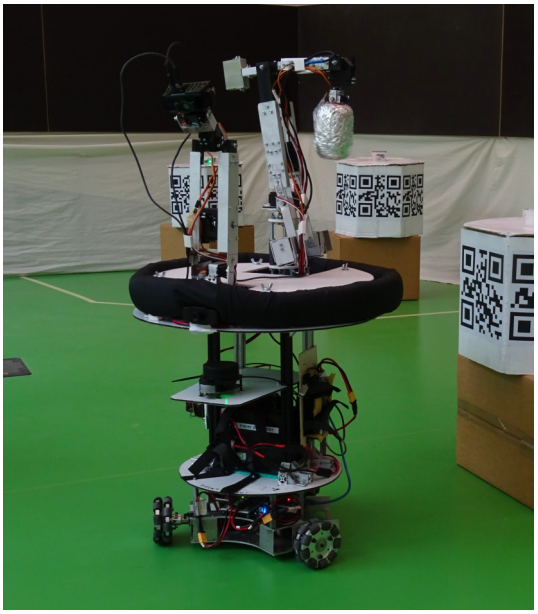
5.1.1. The First Robot

"The Robot" [Figure 4.1b] served as the foundational element of this thesis and was developed as the thesis of Giuseppe Epifani [6], embodying the inaugural sensory translation system within the Physical Metaverse. This robot empowers users to navigate the environment with a clear perception of obstacles.

This mobile robot exhibits the capability to detect obstacles, bumps, and human movements. At its core is a Jetson Nano, serving as the main processor. The robot receives user movement commands through a USB wireless Logitech Controller, and in turn, wirelessly transmits its sensor feed to the user's VR headset for translation.

Equipped with a webcam connected to the Jetson Nano, the robot utilizes a neural network running on the Jetson to recognize humans. Additionally a LIDAR, serially connected to the Jetson, plays a crucial role in conveying obstacle perception to the user during navigation.

To facilitate the robot's movement, an Arduino, plugged into the Jetson, transmits controls to a triskarino omni wheel base. This base is essential for the robot's mobility. The Arduino also detects bumps against obstacles through an Inertial Measurement Unit (IMU), enhancing the robot's spatial awareness during navigation.



(a) Odile at Playful Machines

(b) Blackwings at First Contact

Figure 5.2: Odile and Blackwings.

5.1.2. Odile

"Odile," the largest robot boasting the most degrees of freedom among the avatars in the Physical Metaverse, played a pivotal role as the Physical Avatar in the intermediate phases of this project. It was built as the thesis work by Erica Panelli [18].

Designed to be mobile with some manipulator robot capabilities, Odile's unique feature lies in its ability to express emotions. The robot is controlled by a Raspberry Pi3 and

wirelessly receives movement commands from the joysticks of a VR headset. Notably, Odile is equipped with a pan and tilt camera directly controlled by the VR headset's orientation. This camera system was collaboratively developed and integrated into this project.

Mounted on a Triskarone omni wheel base, controlled by an Arduino UNO, Odile also incorporates an Arduino MEGA to manage two arms. One arm features a camera at the end effector, while the other is equipped with a capacitive sensor. The expressive capabilities of Odile were put to the test in a similar Escape Room activity to that of this thesis. In this scenario, roles were reversed between the Physical Avatar, serving as the guide, and the Visitor, responsible for activating objectives.

5.1.3. Blackwings

Blackwings, crafted by Professor Andrea Bonarini, was initially conceived as a theatrical performance robot. Its distinctive feature lies in a large fabric, referred to as the "wing", which can extend or retract through the activation of two rods driven by a servomotor. Similar to the other robots, Blackwings is a mobile robot with a Triskar base. Upon integration into this project, modifications were made to accommodate new hardware, including the addition of the camera system with pan and tilt functionality. Furthermore, Blackwings was adjusted to increase its height for the incorporation of the updated components.

5.2. Oculus Quest 2

The Oculus Quest 2 emerges as a popular virtual reality headset developed by Oculus, a subsidiary of Meta Platforms, Inc. What distinguishes it is its standalone nature, eliminating the need for a PC or console for operation. This VR device boasts a high-resolution display, a potent processor, and supports 6 degrees of freedom, enabling lifelike movements within virtual environments. With its wireless and portable design, the Oculus Quest 2 provides convenient access to a diverse array of VR games and applications via the Oculus Store. For expanded content options, users can connect it to a compatible PC using Oculus Link. The inclusion of hand tracking and social features emphasizes its design for immersive and communal VR experiences.

In this research, virtual reality proves to be a potent tool for visualizing concepts that would be challenging, if not impossible, to convey through more conventional means. The integration of virtual reality serves as a crucial component, facilitating the complete

immersion of the Controller's sense of sight into the virtual world, effectively eliminating distracting elements.

5.3. Jetson Nano

The primary computational unit driving the robotic system is the Nvidia Jetson Nano board. This single-board computer is designed for AI and machine learning applications, emphasizing efficiency and compactness. The Jetson Nano features a Nvidia Maxwell GPU with 128 CUDA cores, accompanied by a quad-core ARM Cortex-A57 CPU, 4GB of LPDDR4 RAM, and a 16GB eMMC storage module. With support for various AI frameworks and programming languages, including Python and C++, it provides the necessary computational power for handling complex tasks. The operating system utilized is Ubuntu Linux.

Beyond its processing capabilities, the Jetson Nano offers customization and extensibility through expansion ports and connectors. This flexibility facilitates the integration of peripherals and sensors such as cameras, displays, and motors. All sensors within the robot are connected to the Jetson Nano, consolidating data processing within a centralized hub. This configuration ensures a systematic approach to data management and decision-making within the Physical Metaverse framework.

5.4. RPLIDAR A1

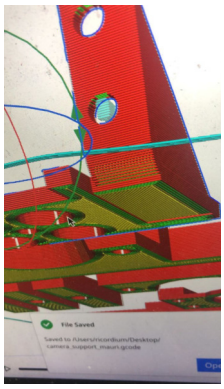
The RpLIDAR A1 Laser Range Scanner stands as a 360-degree scanning LIDAR crafted for applications such as indoor mapping, robot navigation, and obstacle avoidance. With a laser emitter and receiver working harmoniously, this sensor is capable of measuring distances with a resolution of 0.2cm, offering valuable spatial insights for diverse applications. Capable of scanning up to a maximum range of 12 meters, the RpLIDAR A1 delivers detailed spatial information crucial for mapping and environmental awareness. Emitting a modulated infrared laser signal, it captures reflections from objects, sampled by a Vision Acquisition System (VSC) featuring an embedded DSP. This system processes the data, generating rapid and dense point clouds that form a 2D map of the environment.

In essence, the RpLIDAR A1 emerges as a reliable and accurate sensor, presenting a robust solution for robotic perception and navigation. Its seamless integration into the Physical Metaverse project enhances the environmental awareness of robotic avatars, contributing to the creation of immersive and responsive shared physical spaces.

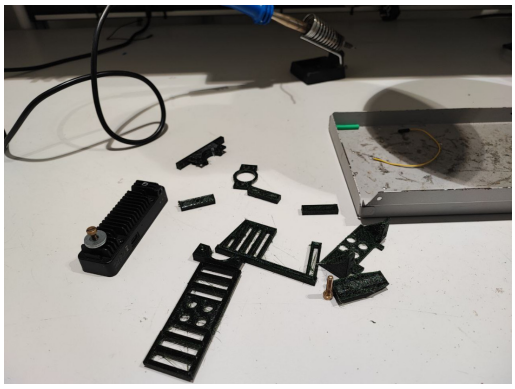
5.5. Cameras

5.5.1. Webcam

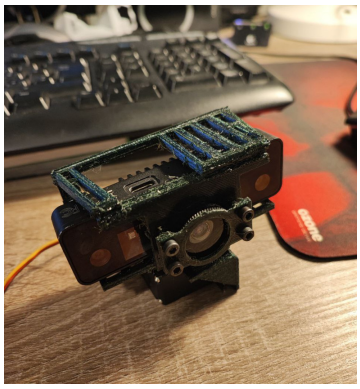
The USB camera used in this project is a straightforward device with a 1080p resolution. Its connection to the Nvidia Jetson Nano occurs through a USB port, serving the purpose of capturing video footage of the robot's surroundings and detecting affordances, more specifically the QR Stations. Although this camera lacks advanced features or capabilities, its reliability and cost-effectiveness make it a practical solution for obtaining visual information essential for the robot's functionality.



(a) Mount slice before 3D print



(b) Assembling the camera holder



(c) Camera mounted on its holder

Figure 5.3: DepthAI Camera mount.

5.5.2. DepthAI OAK-D Pro

The DepthAI OAK-D Pro is a sophisticated camera module designed for applications demanding advanced computer vision and depth perception capabilities. Featuring stereo depth cameras for 3D depth map creation, an onboard AI accelerator from Intel for real-time AI processing, and a color camera, this module caters to a range of applications in robotics, autonomous vehicles, and augmented reality.

Key Features:

- *Stereo Depth Cameras:* Enable the creation of detailed 3D depth maps.
- *Onboard AI Accelerator:* Facilitates real-time AI processing for enhanced computational capabilities.
- *Versatility:* Applicable in robotics, autonomous vehicles, augmented reality, and

beyond.

- *Open-Source Approach:* Encourages customization and innovation, aligning with the open-source DepthAI ecosystem.

The Oak-D Pro DepthAI Camera serves as a significant hardware addition to the existing system within this thesis. With the ability to run trained models on its hardware and transmit final results to the connected computer, it addresses the system's refresh challenges without necessitating a more powerful onboard processor.

In Figure 5.3, we showcase the creation of a mount for our camera, allowing it to be connected to the pan and tilt mechanism introduced in our work.

This device not only provides practical solutions but also aligns with a vision foreseen in a Computer Vision lecture. The professor's anticipation of chips handling both image capture and processing finds realization in the DepthAI OAK-D Pro, highlighting the evolution of technology in the realm of computer vision [17].

5.6. Arduino

Arduino is an open-source electronics platform designed for creating interactive and programmable projects. It uses a microcontroller that can be programmed to control various electronic components. Arduino is known for its user-friendly Integrated Development Environment (IDE), making it accessible to beginners and professionals. With a variety of board models, input/output pins, and expandable shields, it's widely used in applications ranging from robotics and home automation to wearable technology and art installations. Its low cost, educational value, and a large online community make it a popular choice for electronics enthusiasts and educators.

5.7. MPU6050

The MPU-6050 is a compact sensor module that's a powerhouse for tracking motion and orientation. Inside, it houses a 3-axis gyroscope and a 3-axis accelerometer, and it talks to microcontrollers through digital interfaces like I2C or SPI. This dual-sensor setup makes it versatile for applications such as robotics, drones, wearables, and gaming controllers, where knowing how something is moving and positioned is crucial.

5.8. Servomotors

The servomotors used in this project are compact electric motors that excel in providing precise control over angular position. The main models are MG996R and SG90 together with their alternative versions; they are widely used by makers, especially in robotics projects, remote-controlled vehicles, and model aircraft. These servos come with built-in control electronics, including feedback mechanisms to ensure accurate positioning. They are easy to interface with microcontrollers, typically using PWM signals, and are available in various sizes and torque capabilities, making them a versatile choice.

5.9. Esp32 and Esp8266

The ESP8266 and ESP32 are popular microcontroller platforms for IoT and embedded projects. The ESP8266, with its single-core processor, is a more budget-friendly option, perfect for straightforward Wi-Fi-connected projects. In contrast, the ESP32, powered by a dual-core processor and featuring Wi-Fi and Bluetooth, offers more versatility and capabilities. It's an excellent choice when you need both wireless connectivity options, additional GPIO pins, and more memory for larger applications. They are the default Physical Metaverse choice for devices that require wireless capability and that are not directly connected to a more advanced device like Raspberry or Jetson.

5.10. Triskar Omni Wheel Base

AIRLab-made omni-wheel base that was chosen as default mobile base for the Physical Metaverse robots named Triskar.

6 | External Software

In this section, we will outline the key external software tools employed in the development of this project.

6.1. Git

Git is a distributed version control system (DVCS) conceived by Linus Torvalds in 2005 [27]. It plays a pivotal role in tracking changes within source code during software development. Widely acknowledged as one of the most popular version control systems globally, Git is designed to streamline collaboration among multiple developers, manage codebase changes, and meticulously track the historical evolution of these changes. While the majority of this project was developed individually, Git emerged as an indispensable component that significantly contributed to its realization. The complexity of the project, encompassing numerous components and features, surpassed the capacity of a single individual's cognitive load. Hence, having a version control system that allows the creation of "save points" became crucial, enabling a return to specific checkpoints in case issues arose after the addition or modification of certain features.

Moreover, Git fosters the potential for project reuse within the Physical Metaverse framework, aligning with efforts to standardize processes within this overarching initiative. The utilization of Git, therefore, not only streamlines individual development efforts but also aligns with broader collaborative goals and future project scalability.

6.2. Unity and C#

Unity stands as a versatile game development engine and application framework, renowned for its role in creating video games, simulations, AR/VR applications, and interactive 3D/2D experiences. Equipped with a user-friendly visual editor and broad platform support, Unity simplifies the development process. Its extensive ecosystem of assets and plugins solidifies its position as a preferred tool for game development and interactive applications.

C# is a modern, strongly-typed programming language developed by Microsoft. In the context of Unity, C# serves as the primary scripting language, celebrated for its versatility not only in game development, but across various software applications. Key features such as garbage collection, exception handling, and memory management contribute to C#'s reputation as a secure and efficient choice for defining game behaviors and interactions within the Unity environment.

6.3. Python

Python, a renowned high-level programming language, is celebrated for its simplicity and readability. Its versatility spans a spectrum of applications, from web development to data analysis and scientific research. Python's strength lies in its extensive standard library, offering pre-built modules and functions that enhance developer efficiency. As an interpreted language, Python enables rapid development, and its cross-platform nature ensures compatibility across various operating systems. Being open-source, Python thrives on a vibrant community and provides third-party libraries for diverse tasks, including data analysis, web development, and machine learning. Its widespread adoption, coupled with its user-friendly syntax, positions Python as a premier choice for developers across different domains.

Python proved to be of vital importance, imparting agility to the project's development. Predominantly executed on the real robot, Python facilitated swift adjustments to parameters and modifications to behaviors through remote SSH terminal access. This adaptability significantly streamlined the development process, allowing for rapid testing and iteration of features on the physical robot.

6.4. VSCode and GitHub Copilot

Visual Studio Code stands out as a widely-used code editor, esteemed for its speed and extensibility. It serves as a versatile platform for coding in multiple languages, and its functionality is further enriched through an extensive library of extensions.

GitHub Copilot represents an AI-powered coding assistant, working seamlessly alongside your code. It provides real-time suggestions, autocompletions, and even generates code based on your comments and contextual information. GitHub Copilot is meticulously designed to save time and enhance coding efficiency.

The mention of these two tools together emphasizes that power is nothing without control. In the development of this project, GitHub Copilot, omnipresent in the IDE through the

VSCode extension, played a pivotal role. It allowed the delegation of numerous low-level development tasks, tasks that would have consumed valuable time and resources, from aspects of the project that necessitated human intellect and creativity. The integration of VSCode and GitHub Copilot exemplifies a powerful synergy that significantly streamlined the development process.

6.5. UDP Protocol

UDP, or User Datagram Protocol, stands out as a connectionless network protocol known for its emphasis on speed and efficiency. Even though it's not precisely "software" we included it here to avoid creating a dedicated chapter for it. It finds applications in scenarios where a slight loss of data is acceptable, such as real-time video streaming or online gaming. In contrast to TCP, UDP operates without ensuring data reliability, making it akin to sending a letter without requiring a receipt. This lightweight approach contributes to its speed, but it may not be suitable for situations where every bit of data must be delivered in a specific order.

For this project, UDP emerged as the ideal candidate. Given that almost every component continuously transmits and receives data, necessitating minimal latency, UDP's characteristics align well with the project's requirements. In the event of errors, the approach is simply to wait for the next message in the stream.

Certain components, like the Stations, send a limited number of messages and require these messages to be effectively received. Through careful verification, it was determined that UDP's occasional "fallibility" was not statistically significant enough to warrant the project's burden with additional code and sockets dedicated to a more reliable protocol like TCP. Throughout the development of the project, no packet loss was observed, and even if it occurred, it did not manifest any noticeable repercussions. The pragmatic choice of UDP contributed to the project's efficiency without compromising its reliability.

6.6. Tinkercad

Tinkercad: Streamlined 3D Design, Electronics, and Coding

Tinkercad stands out as a web-based platform designed to streamline 3D design, electronics, and coding processes. Particularly popular in educational settings, it provides a user-friendly 3D modeling tool that simplifies the creation of objects, offering an accessible introduction to the fundamentals of 3D printing. Beyond its 3D modeling capabilities, Tinkercad offers a virtual electronics lab where users can construct and test electronic

circuits.

In practical terms, I found this software exceptionally user-friendly, especially for the specific application I intended. It facilitates conceptualizing 3D designs through operations like addition and subtraction of predefined volumes, avoiding the intricacies of manipulating individual vertices. This approach ensures a highly parametric workflow, making Tinkercad an invaluable tool for various users, with both free and paid plans catering to a wide audience.

6.7. DepthAI BlazePose

DepthAI Blazepose, hosted on GitHub by developer geaxgx, emerges as a pivotal repository integrating Google Mediapipe's single-body pose tracking models with DepthAI hardware—a versatile platform enabling AI and depth vision on embedded devices. Geared towards computer vision and machine learning enthusiasts, this repository offers two distinctive modes: host mode and edge mode.

- Host Mode: In this mode, neural networks execute on the device, but the majority of processing occurs on the host computer. This configuration facilitates inference on external inputs like videos or images, broadening the application's flexibility.
- Edge Mode: Conversely, edge mode prioritizes on-device processing, leveraging DepthAI's scripting node feature for heightened speed and efficiency. While limited to the device's camera, this mode minimizes data exchange, transmitting only the essential landmarks of the detected body.

Key Features:

1. *Render Module:* The repository includes a render module capable of displaying the body skeleton in 3D. Users can opt for either inferred 3D coordinates from the landmark model or measured 3D coordinates from the depth map.
2. *Landmark Models:* Users can choose from various landmark models, each with distinct accuracy and speed trade-offs, including full, lite, or heavy.
3. *Customization:* The repository empowers users with customizable parameters such as internal camera fps, smoothing filters, output video file settings, and more.

Underlying Technologies: DepthAI Blazepose harnesses the power of open-source frameworks like Mediapipe models and utilities, designed for creating cross-platform solutions across mobile devices, embedded systems, and the web. Complemented by OpenCV for real-time computer vision and Open3D for 3D data processing, the repository showcases

a robust integration of cutting-edge technologies.

This repository stands as an exemplary use case, demonstrating how DepthAI hardware coupled with Mediapipe models can form a powerful and versatile pose tracking system. Its applications span diverse fields, including fitness, gaming, augmented reality, and beyond, underscoring the transformative potential of edge computing and computer vision in various domains.

7 | The Experiments

This project underwent two separate user experimentation phases: an intermediate one during the Milan Digital Week with "Playful Machines" and the final one at the X-Cities exhibition with the "First Contact" experience. In the following, we will report and analyze only the results of the second experimentation phase. The first phase, primarily due to a significantly low number of total responses, does not allow for a truly meaningful analysis. Mainly because of the public nature of the event we couldn't ensure participants would complete the tests. Additionally, as mentioned earlier, the system underwent significant changes between experiences of that phase. This was because the primary goal was to refine the system as much as possible with continuous and immediate feedback from real users. Please refer to Section 4.6 for a thorough description of the intermediate experimentation phase.

7.1. Methodology

In our final experiments, we aimed to gather results from both the virtual experience of controlling the Physical Avatar and the physical experience of interacting with the robotic avatar. This approach allowed us to gain a comprehensive view of the effectiveness of our system and better understand which components performed better or worse.

The questions we sought to answer through our experiments were as follows:

1. Were the two participants able to understand each other?
2. Were the two participants able to work together?

Due to time constraints and the number of participants, we structured the experience as follows: participants would first perform the activity in virtual reality, isolated from the physical room. Subsequently, they would be guided to the room with the Physical Avatar for the second experience, without revealing the connection between the two. Acknowledging the possibility that participants might notice the link between the two experiences, we intentionally decided to conduct them in this order. Note that due to organizational reasons it was not always possible to follow said order, but the cases were

just some exceptions. The virtual experience is the one we prioritize for evaluation, so we arranged it to be influenced by the least possible bias given the circumstances.

7.1.1. The Activity

The chosen activity is the "Escape Room," as seen in the Digital Week [4.5.1], involving a timed challenge where the two participants must find and activate specific objectives before time runs out. The caveat is that only one participant, that is the Controller with the Physical Avatar, can activate the objectives. This design emphasizes communication and teamwork between the two, leveraging the Visitor's unique capability to distinguish correct from wrong Stations, as opposed to the Controller teleoperating the Physical Avatar which can't notice any difference between Stations but is the only one capable of activating them.

Following the established format of the Digital Week, we divided our environment into two spaces to simultaneously conduct our experiments: The Virtual Maze and the Physical Maze

7.2. The Virtual Maze

The Virtual Maze was set in the first space. Participants were seated next to a desk with our PC streaming the application to the VR headset. After providing a brief description of the experience, participants were instructed to put on the VR headset, and the session commenced. To ensure there were no communication issues between the application and the robot, and also to reassure those unfamiliar with virtual reality, one of us remained in the room, emphasizing that we would not interact with the participant to avoid influencing their experience.

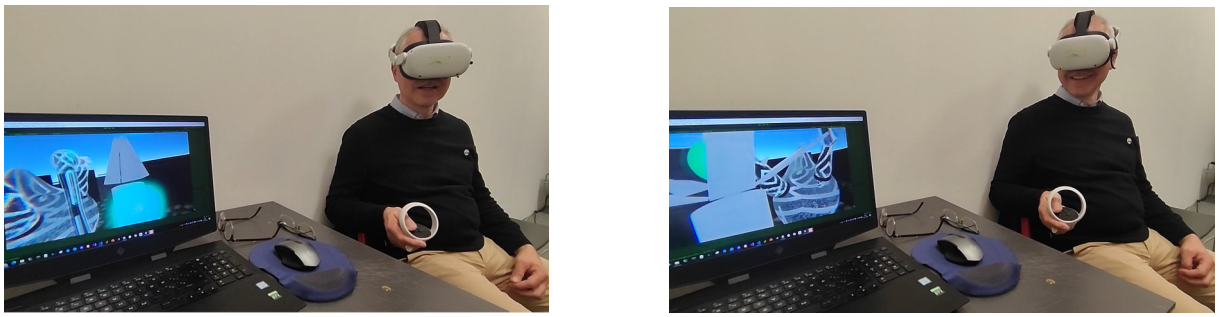


Figure 7.1: Virtual Experience setup.

We introduced the virtual experience with the following description:

"In a short while, you will find yourself in a virtual environment with a specific purpose: much like in an escape room, you must do the right actions before a 10-minute timer runs out, symbolizing nightfall in the virtual world. Be cautious: some actions may be incorrect and will deduct time.

This virtual world differs from the one you're accustomed to, with its own unique logic. To comprehend it, you'll need to explore. Within it resides an Entity that will attempt to guide you toward your goal.

To navigate through space, you have this controller: use the directional joystick to move forward and backward and rotate right or left. Wearing the VR you will notice you are holding a virtual staff in your hand: this will be your tool to interact with certain elements of the world.

Good luck!"

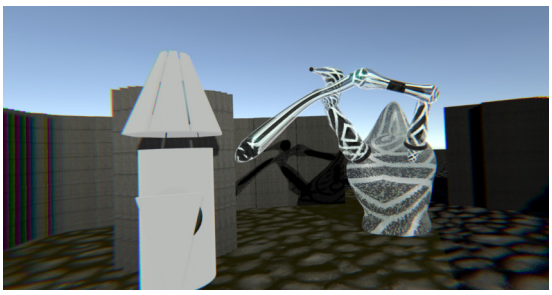
As mentioned earlier, the virtual experience involved navigation a virtual environment in order to figure out the correct actions—activating the three correct Stations, while avoiding incorrect actions—the activation of the other two Stations. The total duration was of ten minutes, with each incorrect Station subtracting 100 seconds. Additionally, the progression of time was reflected in the transition from daytime to nighttime. During the experience, the Entity would appear, as the robot recognized the Visitor in the other room. We incorporated three different Human Translations for the Entity, dividing the experiments accordingly. The three previously developed Human Translations were:

- Odile
- Siid
- Evangelion

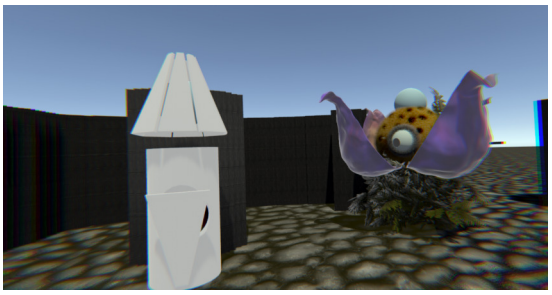
You can refer to their description in Section 4.5.7.

We decided to allocate a larger portion of the tests to Odile and Siid, as Evangelion would require further adaptation for functional compatibility with the Escape Room activity. In its current form, there was a risk that participants might simply ignore it. The limited experiments conducted on this last Human Translation nonetheless provided valuable insights for future developments.

In the following figures we present several screenshots collected from the virtual experience to showcase what participants saw during the experiments.

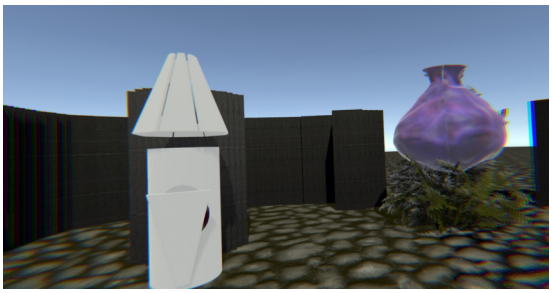


(a) Odile pointing at the Station

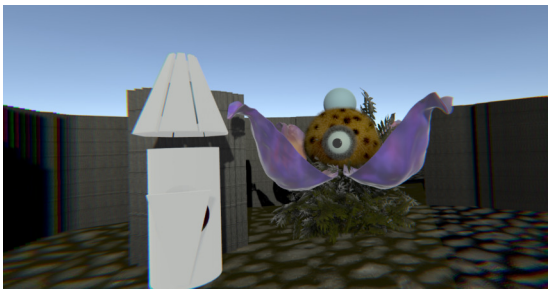


(b) Siid looking at the Station

Figure 7.2: Odile and Siid.

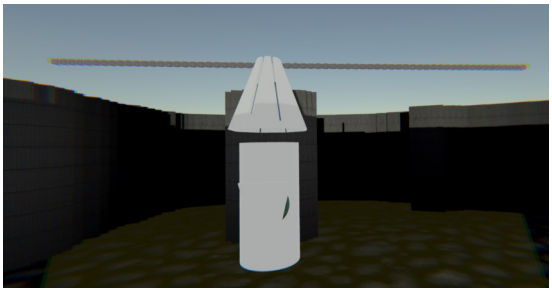


(a) Siid with petals closed

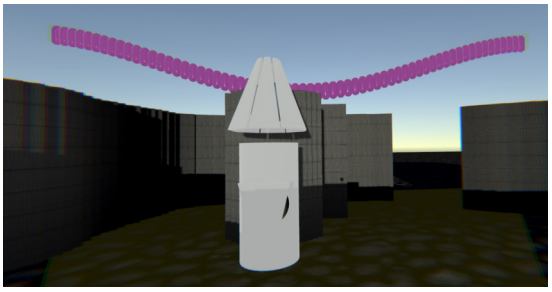


(b) Siid looking forward

Figure 7.3: Siid's different poses.

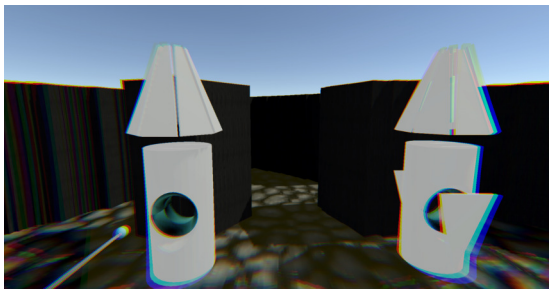


(a) Evangelion when Visitor is out of view

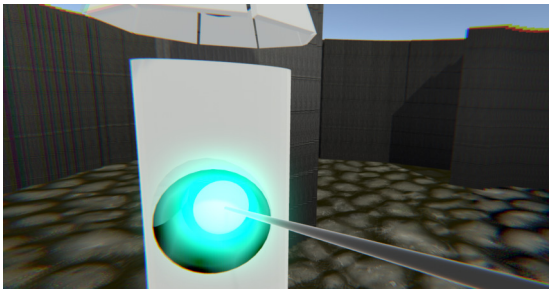


(b) Evangelion when Visitor is in view

Figure 7.4: Evangelion.

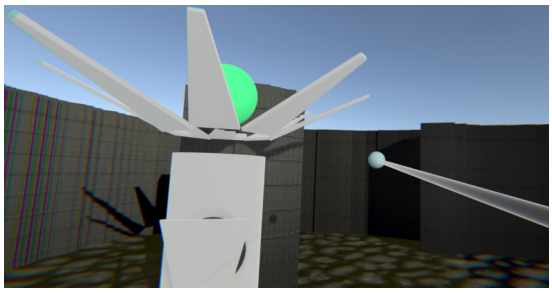


(a) Two Stations, one opened for interaction



(b) Staff's ball blinking during interaction

Figure 7.5: Stations functionality.



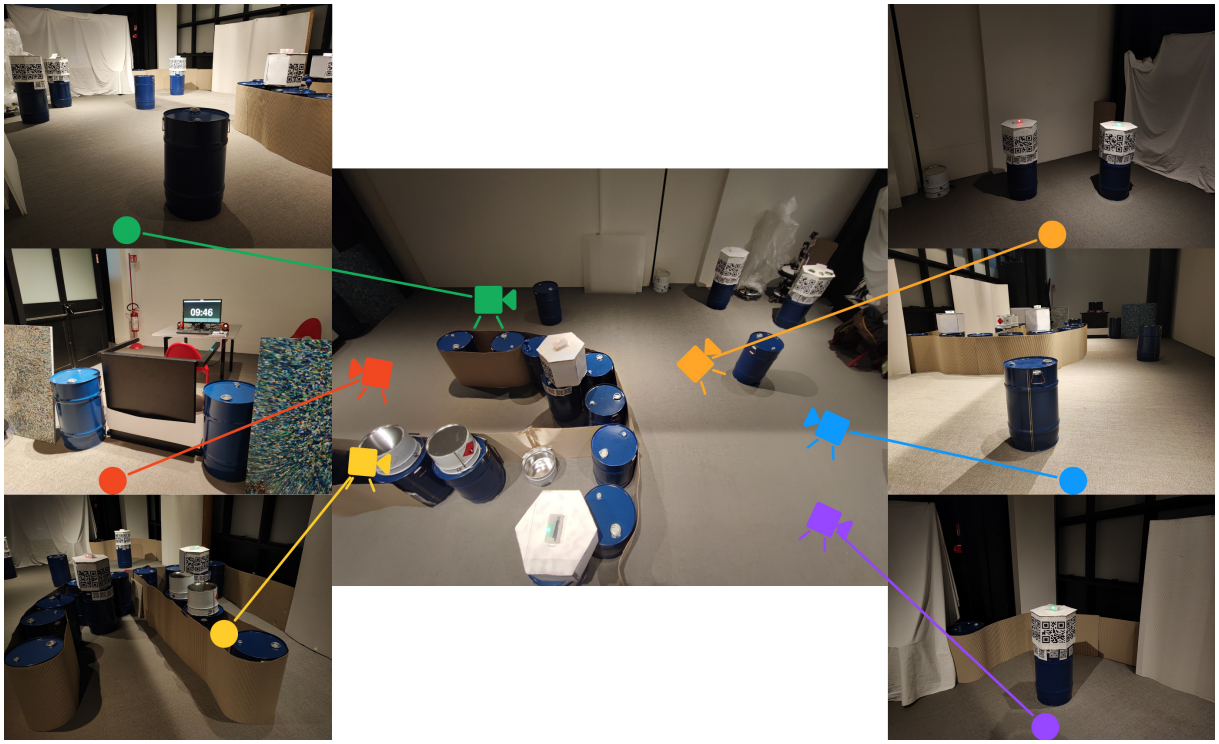
(a) Correct Station activated

(b) Wrong Station Activation activated

Figure 7.6: Activated Stations different graphics.

7.3. The Physical Maze

The Physical Maze [Figure 7.7], complementary to the virtual one, took place in a larger second space. Here, we assembled a maze similar to that built during the Digital Week. Once again, the Stations were represented by our hexagonal boxes with attached QR codes. We added smaller versions of our QR codes to compensate for the loss of perception on the entire large QR code when the robot's camera was too close to the Station.



Maze overview in the center. The stylised cameras represent the location and orientation we used to capture the detail pictures on the sides.

Figure 7.7: Full overview of the Physical Maze.

Above our Stations, LEDs indicated the completion status and whether it was correct or incorrect. In the room, there was a screen running the Game Manager application, displaying the remaining time and graphically representing the number of completed Stations. The application emitted sounds at the start, end, and completion of Stations, with different sounds for correct and incorrect completions.

The physical experience was introduced with the description:

"Soon, you will enter a room from which you can emerge victorious only with the help of a robot present there.

In the room, you will see white boxes with green or red lights on top. Your goal is to ensure that the robot turns off three green lights before time runs out.

The robot is not entirely autonomous, relying on you for assistance. The only way to communicate with it is by being visible to its camera, so make sure it's looking at you if you want to convey something. The robot doesn't have a microphone.

You are not allowed to move or alter obstacles.

On a screen, you can see the remaining time and how many correct lights the robot has turned off.

Good luck!"

(a) Robot approaching a Station guided by the Visitor

(b) Visitor and robot moving to the final Stations together

Figure 7.8: Visitor guiding the robot.

In the room, the Physical Avatar, represented by a different robot from the one used during the Digital Week (this time, Blackwings), was present. The robot would extend and retract its wing multiple times during the virtual interaction with the Stations, not just at completion, signaling that the robot was actively engaging with the Station.

7.4. The Questionnaire

Below, we present the points of the questionnaire presented to the participants. The statements without question mark originally offered alternatives to choose from: "Strongly disagree," "Disagree," "Neither agree nor disagree," "Agree," "Strongly agree." When reporting the data we translated the answers to numerical values ranging from 1 to 5, where "Strongly disagree" corresponds to 1 and "Strongly agree" corresponds to 5. The questionnaire is divided into three parts: one to collect general information and subsequently those for the Virtual Maze and the Physical Maze.

(a) Wing closed

(b) Wing opened

Figure 7.9: Robot interaction routine loop.

General information

- How old are you?
- What is your gender?
- What is your country of origin?
- If you took part in both experiences, which one did you do first? (if you only did one, select that one)

The Virtual Maze

- Had you ever used a VR headset before today?
- I felt uncomfortable using the VR headset today (motion sickness, dizziness, etc).

The Environment

- I was able to navigate the environment.
- I was able to detect and avoid obstacles.
- I felt like I was trapped.
- I felt lost.
- I felt like I was able to go wherever I wanted.
- It was easy to move around.
- I felt like the obstacles were dangerous.
- The environment was interesting/stimulating.
- The environment was annoying.

The Game

- It was easy/intuitive to understand how the game worked.
- It was easy/intuitive to understand how to interact with the Stations.
- Playing the game was frustrating.
- I felt immersed in the game
- Did you win?

Questions related to the other Entity

- I felt alone in the environment.
- I felt a living being was in the environment with me.
- I felt an intelligent being was in the environment with me.
- I felt another human being was in the environment with me.
- I felt a connection with the other being.
- I felt the other being was trying to communicate with me.
- I was able to understand what the other being was doing.
- I was able to understand what the other being was trying to do.
- I wanted to communicate with the other being.
- I tried to communicate with the other being.
- I felt like I was able to communicate with the other being.
- You said that you felt alone. What was in your opinion the Entity you saw in the environment?
- How would you describe the other being's appearance?

Extra questions

- Overall, how much did you enjoy the experience?
- What did you enjoy the most?
- What did you enjoy the least?
- How would you improve the experience?
- Feel free to add any comment.

The Physical Maze

- Did you take part in "The Physical Maze" experience: the game in the real world with the robot?
- Had you ever seen a robot before today?
- Had you ever interacted with a robot before today?
- I liked the robot appearance.

Questions about interaction with the robot

- I felt like another intelligent being was in the environment with me.
- I felt that the robot was trying to learn how to behave.
- I was able to understand what the robot did.
- I was able to understand what the robot was trying to do.
- I was able to understand what the robot communicated.
- I was able to understand what the robot was trying to communicate.
- The robot had a consistent response to what I was doing.
- I felt like the robot was able to understand my intentions.
- I was able to make the robot understand my intentions.
- I felt a connection with the robot.
- I wanted to communicate with the robot.
- I tried to communicate with the robot.
- I felt like the robot was trying to communicate an emotion.
- Which emotions did you feel the robot was trying to communicate?

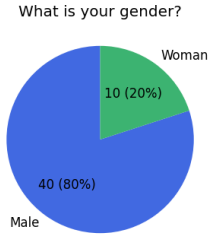
Extra questions

- How much did you enjoy the experience?
- What did you enjoy the most?
- What did you enjoy the least?
- Feel free to add any comment.

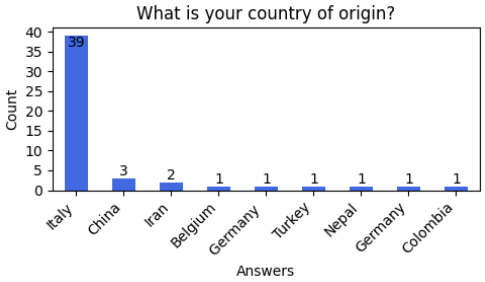
7.5. General Information

In this section, we review participants' responses concerning general information. As mentioned earlier, some individuals completed the Physical Maze activity first, but we ensured that all participants engaged in both activities. Additionally, we provide information on participants' nationality, age, and gender.

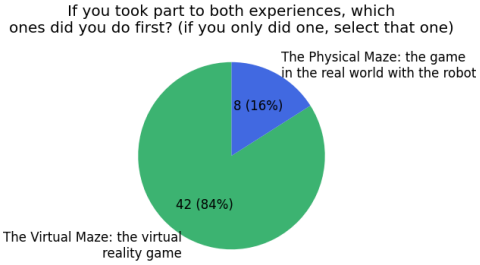
(a) Participant age



(b) Participant gender



(c) Participant country of origin



(d) Experience order

Figure 7.10: General information.

7.6. The Virtual Maze

Before delving into the analysis of the virtual experience, we present in Figure 7.11 the responses to general questions posed at the beginning. A significant portion of the participants had not used a VR headset before today, which may have contributed to the difficulty some participants experienced in controlling their movement in the virtual world, as we will discuss later. Fortunately, the majority of participants did not report issues such as motion sickness or dizziness during their experience.

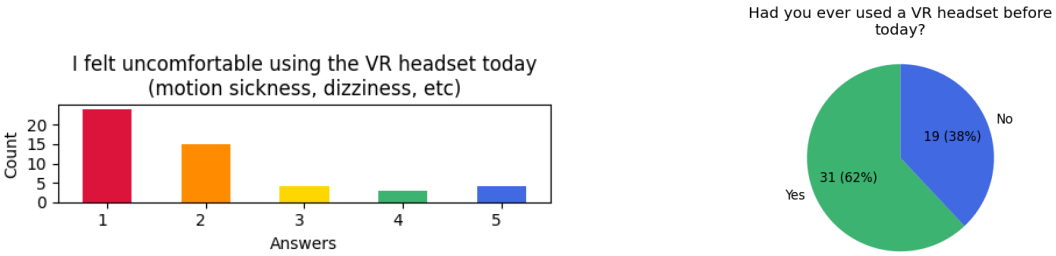


Figure 7.11: General questions about VR.

7.6.1. The Environment

Overall, navigation within the environment went well. The results are reported in Figure 7.12. Some difficulties emerged in control but it is noteworthy that, given the focus of this thesis, the emphasis was not primarily on this aspect but rather on perception itself. Anyway we managed to provide a solid foundation for future, more targeted work. The majority of responses opposing the statement "I was able to detect and avoid obstacles" can be attributed to the functioning of our application. The fact that the Visitor is perceived as an obstacle by the LIDAR led us to hide the corresponding pillars and position the Human Translation model in their place. Consequently, when the camera failed to detect the person, the controllers faced obstacles that were moving in the environment – the unmasked legs of the Visitor. This could create an impression of a more complex and intricate environment than reality.

"I felt like the obstacles were dangerous" received mixed responses. Our system caused the screen to flash red upon collision with an obstacle, but not all participants connected this effect with its cause. Future efforts should focus on better communicating collisions to users for effective avoidance.

A commonly reported issue by participants was the slowness of movement, a challenge we haven't yet resolved. While this is ultimately dictated by the actual speed of the

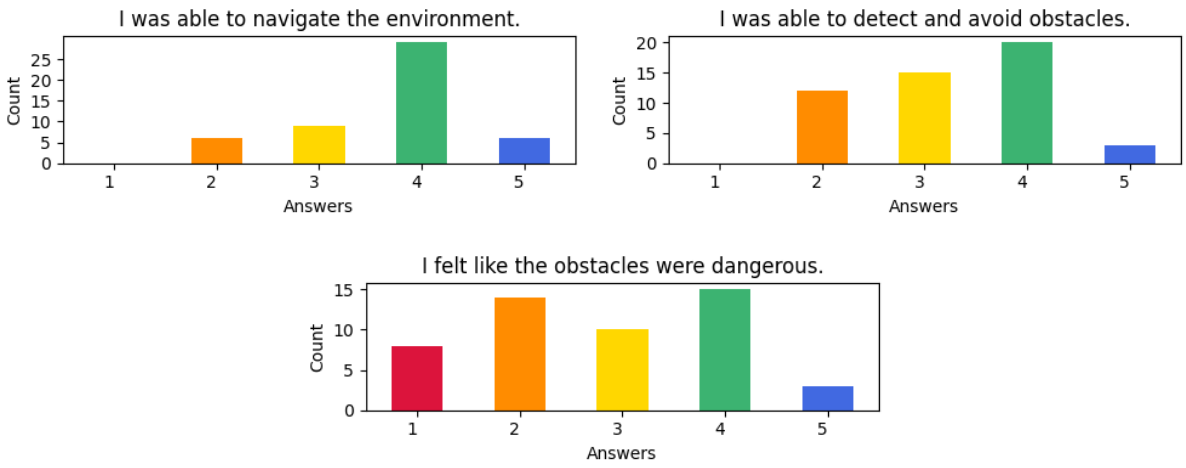


Figure 7.12: Environment survey responses part 1.

robot, which we aim to keep moderate. One potential solution involves significantly enlarging the virtual world, as this would create larger spaces, providing a greater sense of movement. However, this approach, previously used during the Digital Week, presents interaction problems with Stations due to scaling issues. This challenge is further complicated when manual interaction with Stations, as in these experiments, rather than proximity activation, as in the previous version, is desired.

An important note on controls: they often seemed non-intuitive due to the pan and tilt mechanism of the camera as it introduces a relative look orientation with the robot body, and how we represented it in the virtual world. Tilting the analog stick forward does not move the robot in the direction the camera facing but rather in the direction its body is facing. Many participants found this system non-intuitive, and in the future, we need to explore a solution if we want to maintain pan and tilt to separate gaze direction and the actual orientation of the robot's body.

With the responses in Figure 7.13, we continue the analysis of feedback related to the environment. The answers paint a picture of difficulty and significant physical constraints in movement, accompanied by feelings of entrapment and disorientation. These results may show a more challenging system than the one developed in the previous work [6], but it's essential to consider crucial factors: due to its nature, our sensory translation system conveys a highly homogeneous environment, and the maze we constructed was particularly challenging in some areas compared to the past arrangement.

Most importantly in the previous study the interaction was unstructured, allowing participants to navigate freely without pressure and for a much shorter duration (maximum three minutes instead of ten). The combination of the new factors likely introduced "per-

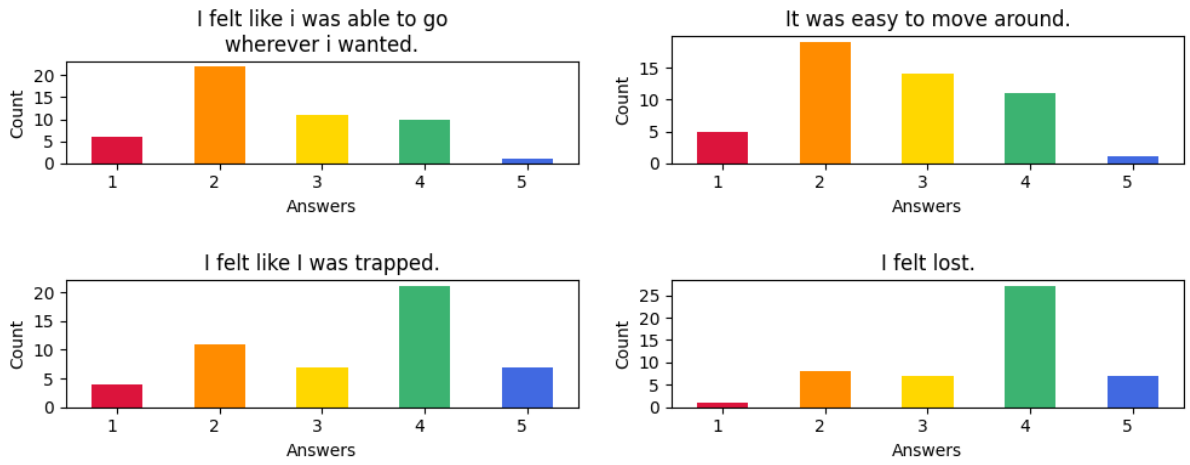


Figure 7.13: Environment survey responses part 2.

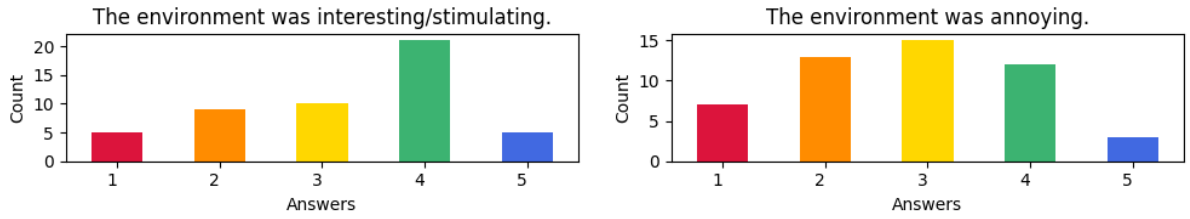


Figure 7.14: Feedback about the environment.

formance anxiety" in participants. Despite reporting this sentiment in their responses, participants generally succeeded in finding Stations and activating them.

We conclude the analysis of the environment with the feedback presented in Figure 7.14. Participants described the environment as interesting and stimulating, yet simultaneously annoying. While these two responses may appear contradictory, in future studies, we aim to refine our inquiries to better distinguish the specific elements contributing to these evaluations.

7.6.2. The Game

The responses in Figure 7.15 provide a general overview of how participants perceived the structured activity. Overall, the feedback is positive: achieving a strong sense of immersion was crucial for our study, and it was successfully attained. The interaction with the Stations was intuitive for the majority of participants. However, some reported frustration, likely due to slight instability and occasional non-reproducible Station detection. Addressing these issues will be a focus of future work.

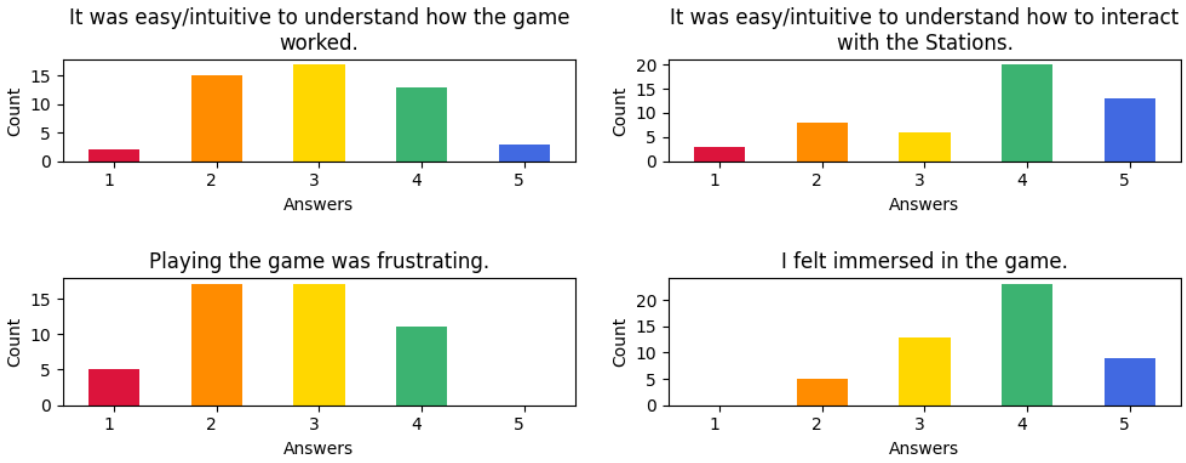


Figure 7.15: Game survey responses.

The game, in general, was perceived as intuitive by less than half of the participants. Hence, future efforts will also involve refining the game structure and perhaps providing clearer instructions at the beginning of the experience.

7.6.3. Questions Related to the Other Entity

We now delve into the central aspect of our study. Let's first analyze the result in Figure 7.16: there has been a significant improvement compared to our previous work [6]. Now, the majority of participants no longer feels alone in the environment, indicating that we are moving in the right direction. It's important to note that in the introduction to the activity, we mentioned the presence of "another Entity" in the space: this likely introduced some bias in our responses. However, we will confirm our assertion through the analysis of subsequent results.

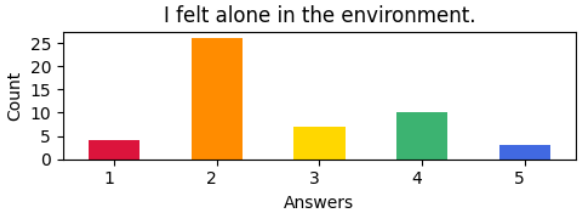


Figure 7.16: General feeling of loneliness.

In the graph shown in Figure 7.17, we get an overall picture of the total occurrences for each Entity. As mentioned earlier, we allocated a smaller portion of the tests to Evangelion because it was not deemed ready for goal-oriented activities. We leave the completion of this Human Translation to future studies, while still presenting the results obtained in

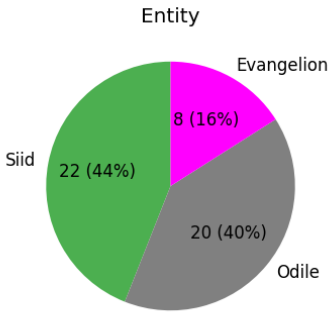
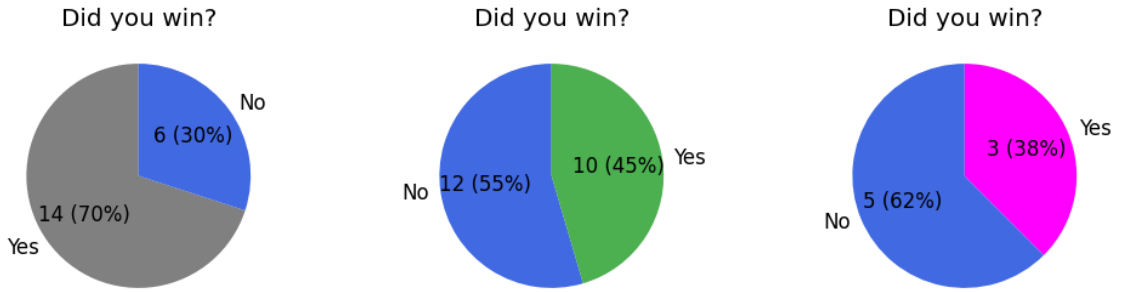


Figure 7.17: Total recurrences of each Entity in the experiments.



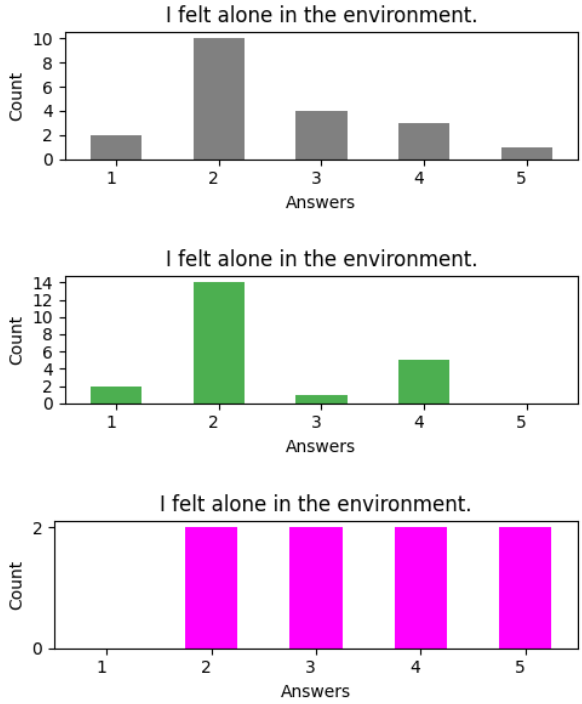
(a) Victories with Odile (b) Victories with Siid (c) Victories with Evangelion

Figure 7.18: Total victories per Entity.

this phase.

In Figure 7.18, we showcase the total victories achieved by participants with each Entity. It's immediately evident that the overwhelming majority of victories were reported with Odile, underscoring the importance of the Visitor's ability to manually indicate targets in our activity, which is captured just by said Human Translation out of our three.

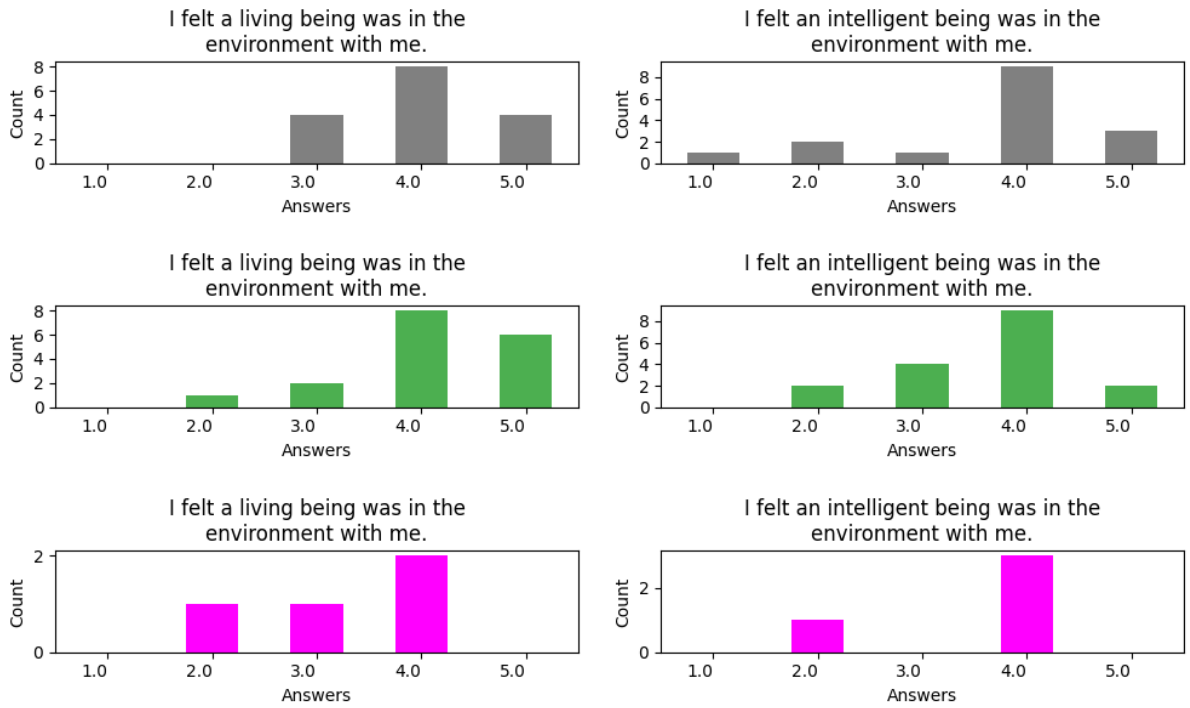
Subsequently, we will provide a detailed analysis of the interaction for each Entity. The colors will be reused to differentiate Entities in the graphs, offering a more intuitive understanding for the reader.



From top: Odile, Siid, Evangelion.

Figure 7.19: Feeling of loneliness with each Entity.

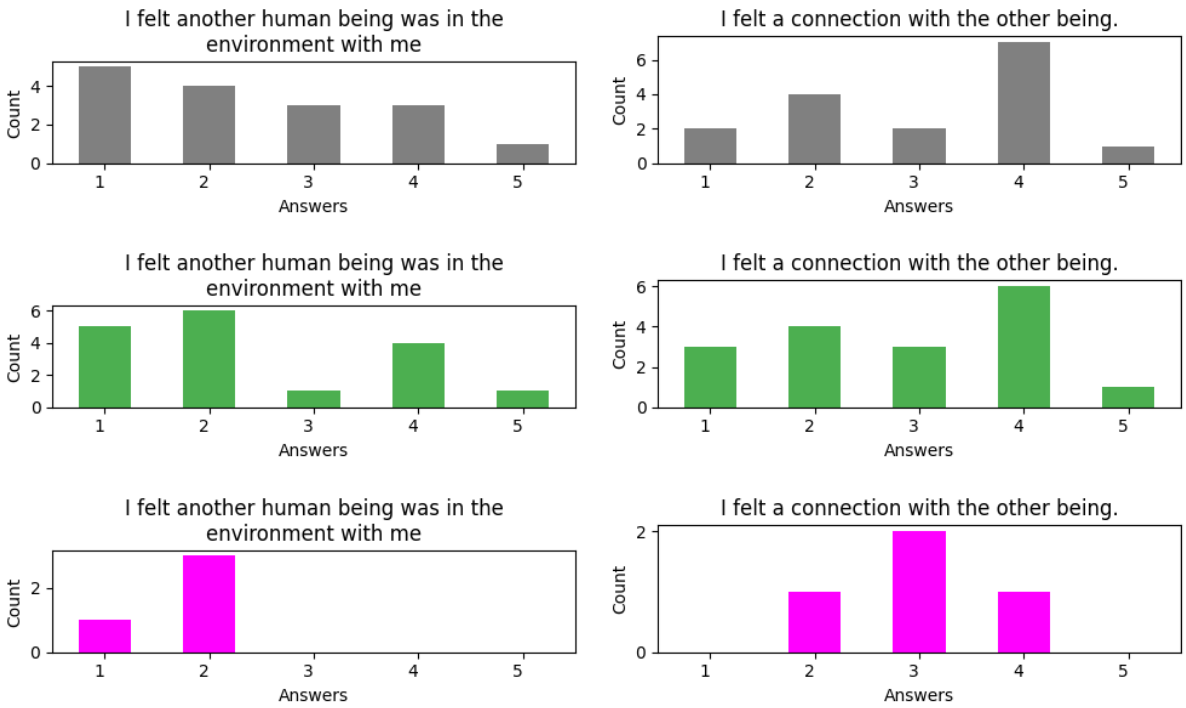
Returning to the feeling of loneliness, as shown in 7.19, we now conduct a more focused analysis, separating the responses given in experiences with the three different Entities. The results between Odile and Siid are quite similar and positive in terms of not making the Controller feel alone. In contrast, the responses related to Evangelion, considering the limited number of total responses, generally indicate a perceived sense of loneliness. The component that likely played a crucial role in this differentiation is the possibility of our Entities to navigate the virtual space and come closer or further from the Controller, catching his attention. However, we will now continue the analysis to extract more detailed information.



From top: Odile, Siid, Evangelion.

Figure 7.20: Feelings of living and intelligent being presence.

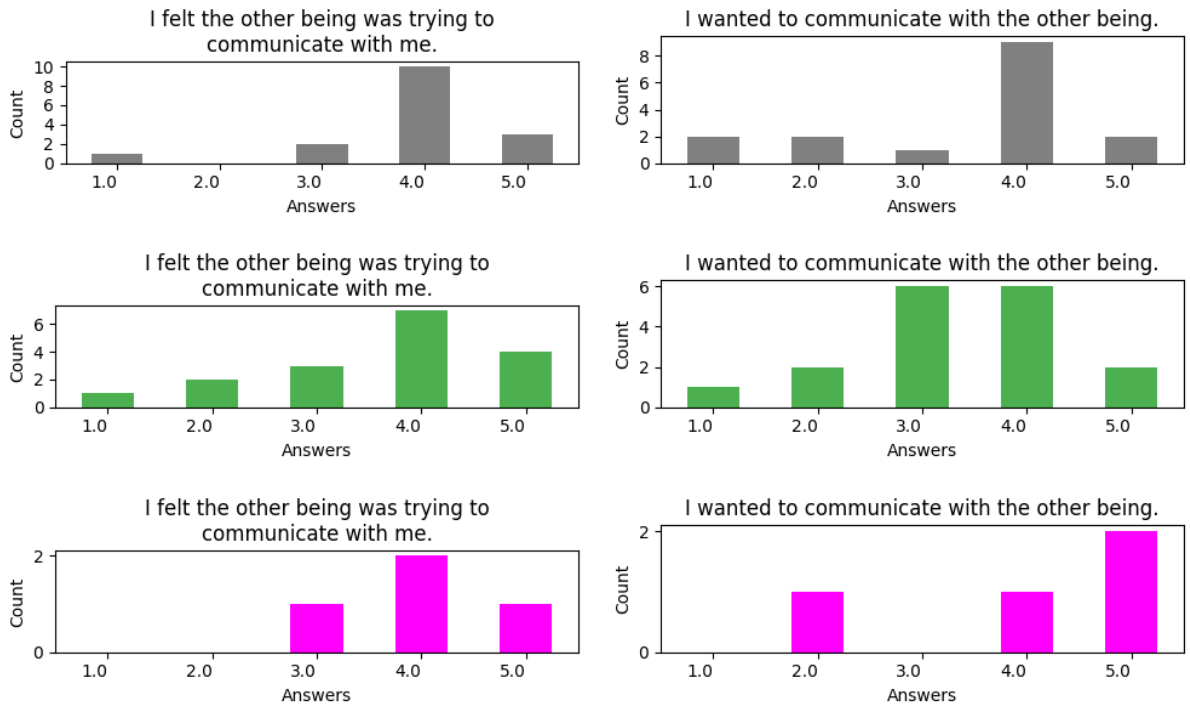
With 7.20, we present the feelings of the presence of a living and intelligent being. For all Human Translations, there is a correspondence between the responses given to both questions, leaning predominantly towards agreement. This suggests that our system has performed well, effectively capturing the movements of the Visitor in a fairly consistent manner.



From top: Odile, Siid, Evangelion.

Figure 7.21: Feelings of human presence and connection with the other being.

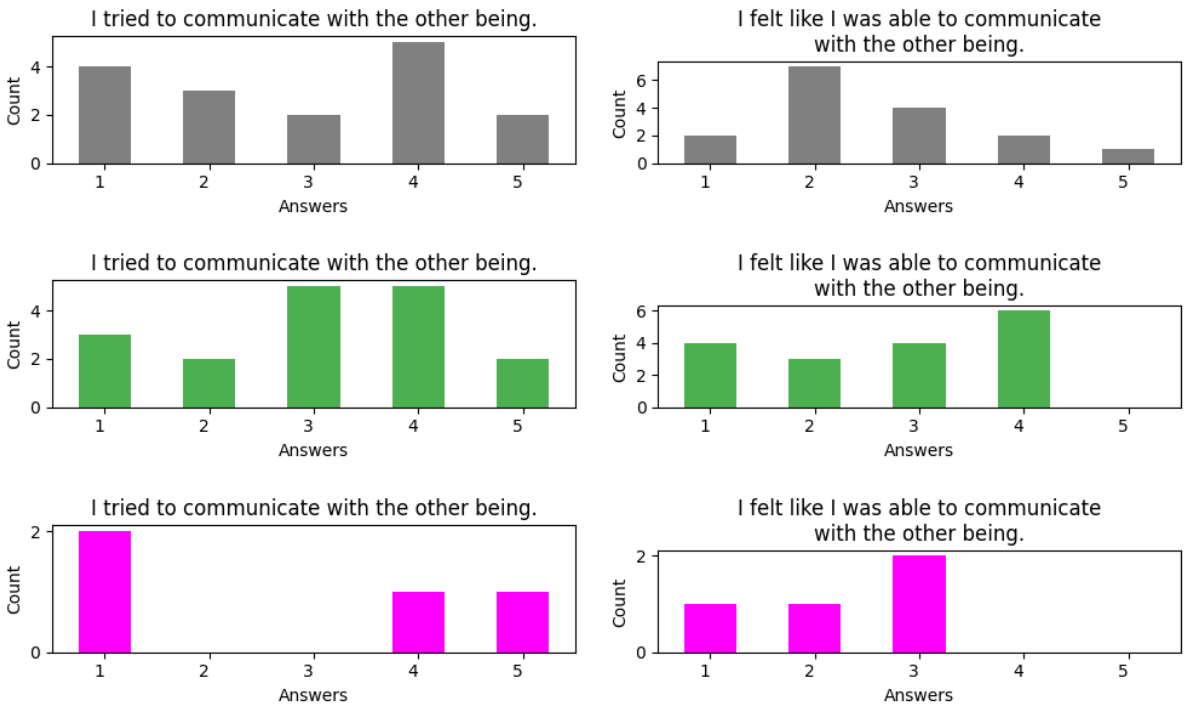
In Figure 7.21, we observe a particularly intriguing result, especially when combined with the responses we've just examined. While users perceive a living and intelligent being, the strong confirmation of it being perceived as human is not as pronounced. This result indicates that our Human Translations are indeed headed in the right direction. They go beyond a simple remapping of human joints, translating movement and pose onto different components. Another interesting aspect is the sense of connection felt by users towards the other being. While not entirely affirmed, as we will analyze in the next figure, it is certainly present.



From top: Odile, Siid, Evangelion.

Figure 7.22: Communication with the other being part 1.

The results in Figure 7.22 are promising regarding the setup of the interaction between Controller and Entities. The perception of a willingness to communicate from both sides is evident, laying a solid foundation for the development of bidirectional interaction. Unfortunately, in this phase of development, we were not able to fully cultivate this aspect as we will explain with the next answers' analysis

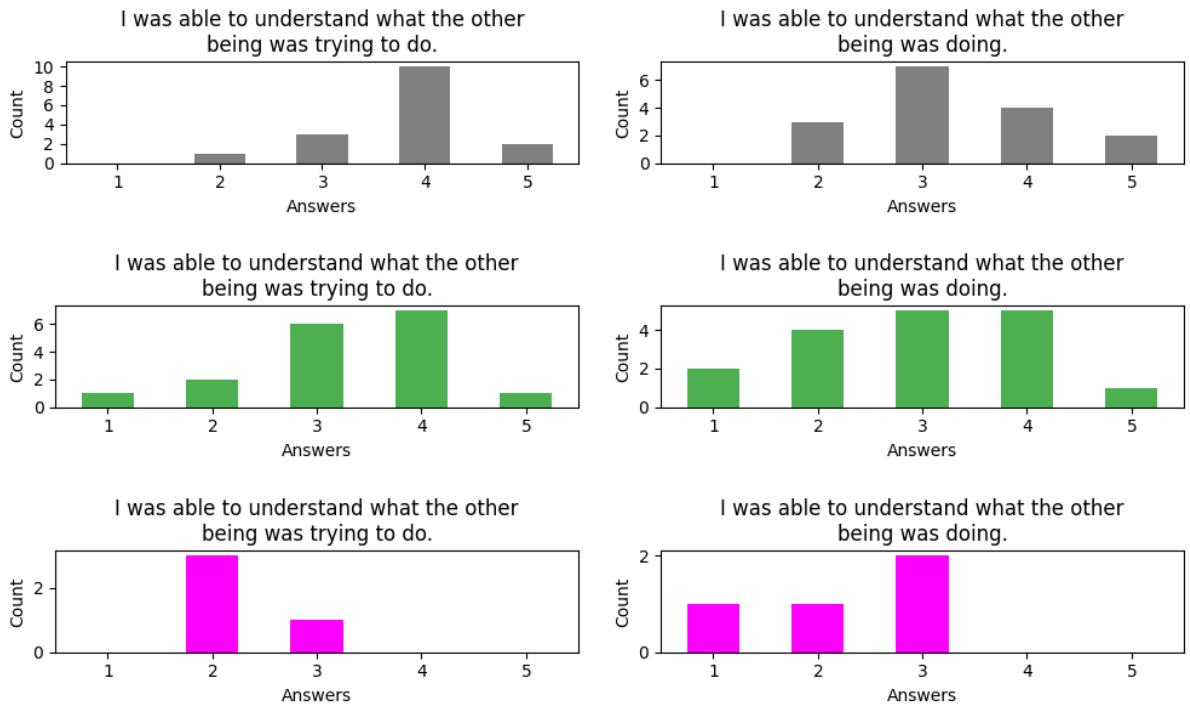


From top: Odile, Siid, Evangelion.

Figure 7.23: Communication with the other being part 2.

With the results presented in Figure 7.23 we delve into the details of communication. The findings underscore the need to develop interaction further because a significant portion of users indeed attempted to communicate with the other Entity. During tests, we observed that they primarily did so using the virtual staff in their hand. However, most users realized that it was not possible to interact effectively with the available controls. The only real methods to communicate something to the Visitor in the other room, or better his Human Translation in the virtual world, were either by looking in his direction or by moving towards or away from him.

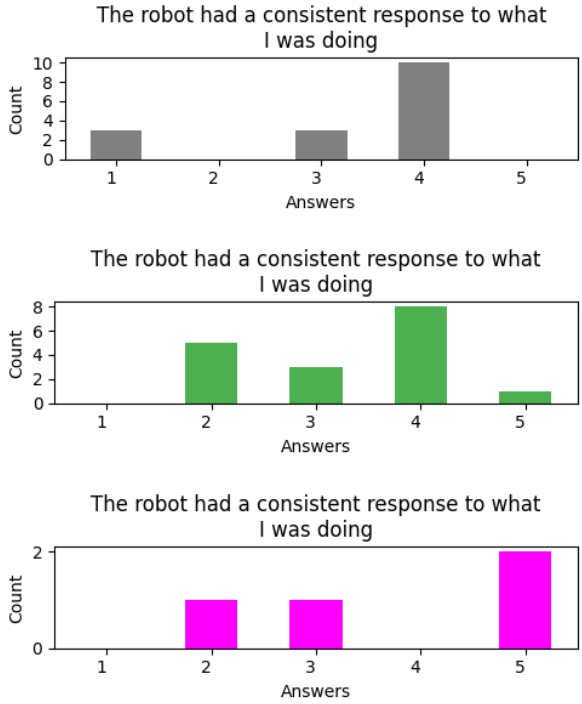
Differences in the perceived ability to interact with the various Entities are noticeable. Evangelion, likely due to its distance, was the least interactive, but there's also a difference between Odile and Siid. This discrepancy is probably attributed to the shapes of the two Entities: Odile maps the Visitor's arms to its pointer, while Siid maps them to the opening or closing of its petals. Consequently, when the robot approaches the Visitor, slight arm movements cause Siid to fully open, revealing its eye. In contrast, Odile gives the impression of merely waving its arm. The varied perception of communication capabilities likely stems from the fact that Siid's petal opening is perceived as more expressive compared to Odile's pointing gesture when communicating face to face.



From top: Odile, Siid, Evangelion.

Figure 7.24: Understanding of the other being's action.

Now, let's turn our attention to the understanding between the Visitor and the Controller. In Figure 7.24, we observe the results of their respective responses. As anticipated, Evangelion was generally less understood, indicating a need for additional calibration work. Regarding Siid and Odile, however, there's a slight discrepancy between the understanding of "trying to do" and the actual "doing." This discrepancy may arise from the fact that the Entity cannot directly interact with elements in the virtual world or even those in the physical room. It merely directs the Physical Avatar toward the correct objectives. Participants may have perceived an intention from the Entity that did not translate into practical actions, as suggested by the responses to "I was able to understand what the other being was doing." In the future, allowing the Visitor to interact with the physical room and enhancing the visualization of the Entity to allow even graphical interaction with elements in the environment could provide some feedback to users and improve these aspects.



From top: Odile, Siid, Evangelion.

Figure 7.25: Robot response consistency with action.

As we delve into the results of 7.25, we transition to the physical room side of the experiment. We opted to differentiate Visitors' feedbacks on the robot by Entity because when controllers face one specific Entity, especially recognizing it as their guide, may behave differently, reflecting a distinct Physical Avatar behavior. The results obtained with Odile and Siid are quite similar. An interesting finding emerges for Evangelion, albeit to be cautiously considered due to the extremely limited data: two out of four users perceived a strong correspondence between their actions and those of the robot picturing them as Evangelion. In the future, it will be worthwhile to explore this result further to understand whether it was an isolated case or if genuine comprehension of the Visitor by the Controller is not a necessary requirement for consistent behavior towards it.

7.7. The Physical Maze

In this section we review the results concerning the physical experience, it's essential to note that our main research focus was on the sensory translation system so most of our efforts during development were directed to the Virtual Maze. However, adopting a more inclusive approach to the subject and considering that the intended activity inherently involved a physical component, we proactively included this aspect in our study, differently from the past.

7.7.1. Questions about Interaction with the Robot

In general, participants leaned towards appreciating the robot's appearance, as depicted in Figure 7.26, although it wasn't universally liked. This aspect is crucial if we want people to be motivated to interact with it rather than ignore it. Many, if not all, perceived the robot as intelligent and could understand what it was attempting to do. Additionally, there was a widespread perception that the robot was learning how to behave, opening the way for a a layer of more complex interaction to the participants' experience.

Figure 7.26: Interaction with robot survey responses part 1.

In Figure 7.27, we can highlight additional insights. Participants, in general, understood the robot's intentions, although many struggled to comprehend what it was communicating or attempting to communicate. The consistency of the robot's responses to user actions has been previously analyzed but is consolidated here independently of the Human Translations used. The results tend towards reasonably consistent reactions from the robot, albeit not strongly, indicating that there is still work to be done on how much the avatar is capable to express from what the Controller is doing in simulation.

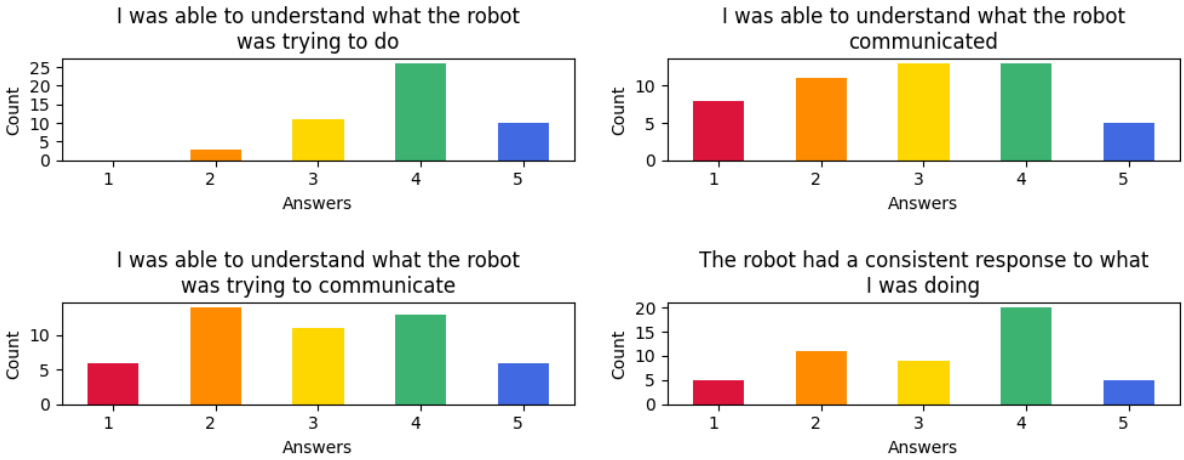


Figure 7.27: Interaction with robot survey responses part 2.

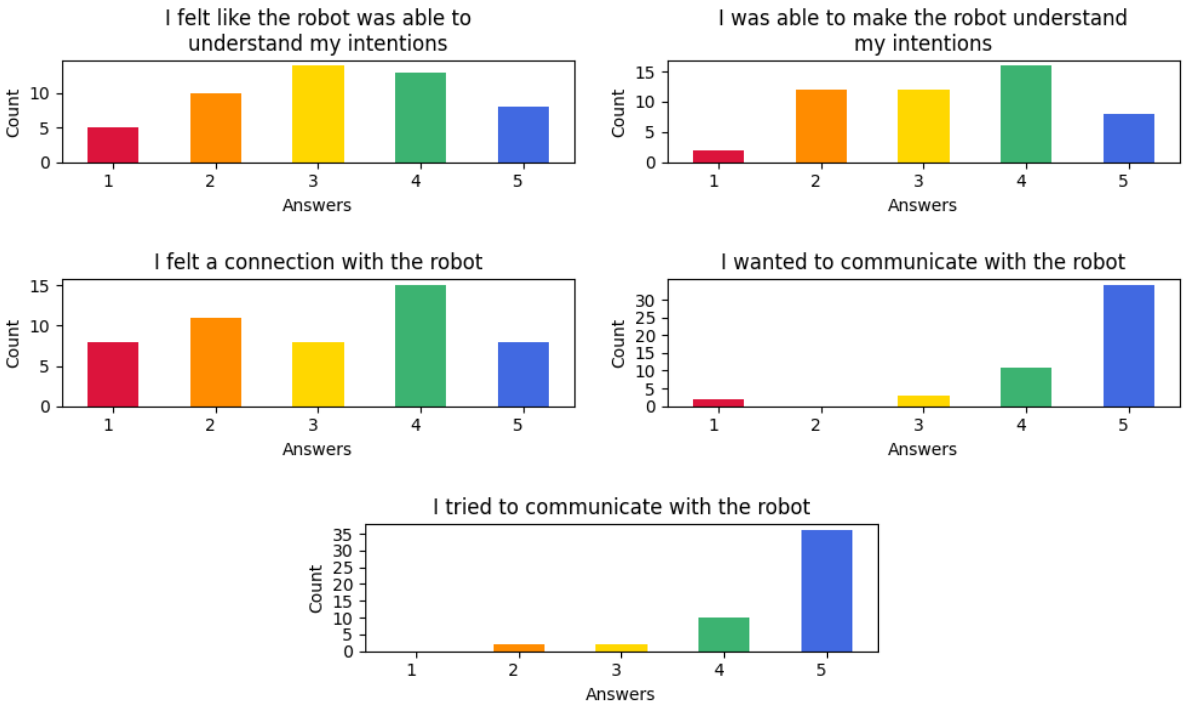


Figure 7.28: Interaction with robot survey responses part 3.

With Figure 7.28, we conclude our analysis of the physical aspect of the experiment. It's evident that the vast majority of participants wanted and attempted to communicate with the robot, highlighting the ample opportunities for interaction in structured experiences of this nature. The results regarding mutual understanding of intentions and the sense of connection with the robot yielded mixed responses. However, we did not expect strongly positive responses as we could not primarily focus our development efforts on this physical side of the experience.

7.8. Extra Questions

Overall, as evident in Figure 7.29, participants enjoyed both experiences. This is a significant outcome as our primary goal fundamentally is to study the interaction between two individuals in a game. Ensuring that both participants enjoy the experience is crucial to encourage active engagement from both parties.

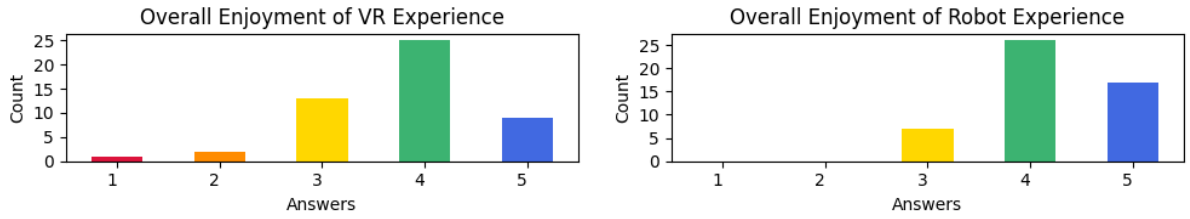


Figure 7.29: Overall Enjoyment of VR Experience and Robot Experience.

7.9. Final Considerations

Let's revisit the questions we set out to answer at the beginning of this chapter:

1. Were the two participants able to understand each other?
2. Were the two participants able to work together?

Based on the results obtained, we can partially affirm both questions. Looking at the participants' responses to questions about mutual understanding for the first question and considering the percentage of total victories for the second question, we find positive indicators. It's important to note that the activity, as it is structured, is challenging to complete solo in the Virtual Maze for a user unfamiliar with the system. Randomly unlocking Stations consumes a lot of time, and navigating the environment requires time to master, particularly without prior knowledge of what to expect during exploration. While there is still much work ahead, it's evident that we are moving in the right direction.

8 | Conclusions and Future Developments

This project has successfully demonstrated the feasibility and potential of integrating advanced technologies like virtual reality, robotics, and sensory translation in the context of the Physical Metaverse. The experiments conducted, particularly the final phase during the X-Cities exhibition, have provided invaluable insights into the dynamics of human-avatar interaction and the effectiveness of our sensory translation system.

Our findings indicate that participants were able to engage meaningfully with both the virtual and physical aspects of the project. The results from the Virtual Maze experiment, especially, highlight the capability of participants to navigate and interact within a virtual environment effectively with our system. The presence of a Physical Avatar, which participants controlled remotely, added a layer of complexity and realism to the virtual experience. This blend of virtual and physical elements showcases the immense potential of the Physical Metaverse for creating immersive, interactive experiences.

Moreover, the project has made significant strides in understanding nonverbal communication in a virtual context. By allowing participants to embody non-humanoid avatars, we have opened new avenues for exploring how nonverbal cues and interactions can be translated and understood across different forms. This has implications for a wide range of fields, from telepresence and remote collaboration to entertainment and gaming.

8.1. Future Developments

Looking ahead, several areas have been identified for future development to enhance the capabilities and applications of this project:

8.1.1. SLAM vs LIDAR Algorithms

Exploring the integration of Simultaneous Localization and Mapping (SLAM) technology, as opposed to relying solely on handcrafted LIDAR algorithms, could significantly

improve the system's capability to distinguish humans and environment affordances from simple obstacles. The integration of this technology into the system will require a substantial amount of refactoring, but it promises great advantages. It will likely lead to the abandonment of algorithms developed for LIDAR, but in any case, the part concerning camera calibration will remain a fundamental component. The key is to have conducted significant experiments and paved the way for future developments, demonstrating that the designed system is indeed feasible. Looking back after completing the work, we can state that without the succession of tight deadlines for public exhibitions, which guided us to avoid risks and stay the course on the already consolidated work, we would probably have explored the SLAM alternative earlier.

8.1.2. Kalman filtering of the Entity's location

The suggestion to implement a Kalman filter for the Entity's location comes from Professor Bonarini while he was observing the Digital Week's virtual experiences. While at that time we found more immediate points for improvement in our work, we strongly recommend considering this strategy to those who will take over the project. Implementing a Kalman filter for human pose estimation may enhance the accuracy and fluidity of the avatar's movements. This would result in more natural and intuitive interactions, making the virtual experience more immersive and engaging.

8.1.3. New Human Visualizations

The primary goal of this thesis was to overcome technical challenges in realizing the system theorized in the Physical Metaverse research, focusing on providing engineering solutions. Now that we can confidently state that many of the challenges have been overcome, we feel it's time to hand over the baton to designers who can finally bring out the best in this project, enriching it with new and intriguing Human Translations to push the interaction between human and Physical Avatar to the limit.

8.1.4. Beyond QR, ARUCO Tags

The choice of QR codes was immediate insofar as we needed a quick solution to implement an activity for the Digital Week. We then found that this solution was indeed stable and versatile, but it has a non-negligible computational cost on the robot's onboard computer. We suggest exploring alternatives for the future, such as ARUCO tags, which could introduce additional stability in Station detection, possibly even at greater distances and eliminating the need for small versions of QR codes.

8.1.5. Proprioceptive Board Movement Control

A solution we would have liked to integrate for avatar control is movement through the Proprioceptive Board. This would recall the work of another student who contributed to the Physical Metaverse [3]. This control method consists of a wooden board suspended by cushions with a central fulcrum composed of a wooden hemisphere. Users, by stepping onto this board, can command the robots simply by leaning in a direction, thanks to an IMU connected to an ESP32 located at the center of the board. This control system would likely solve the problem of the separation between the direction of movement and the direction of observation with the VR headset because users would have a physical intuition of their frontal direction, thereby increasing the sense of embodiment.

8.2. Final Thoughts

In conclusion, this project represents a significant step forward in the exploration of human-avatar interaction within the Physical Metaverse. The insights gained from the experiments conducted provide a solid foundation for future research and development in this field. The potential applications of this technology are vast, ranging from enhanced remote collaboration to new forms of entertainment and education. As we continue to explore and develop these technologies, we open the door to a future where the boundaries between the physical and virtual worlds become increasingly blurred, leading to new possibilities and experiences.

Bibliography

[1] A. Bonarini and S. Besio. *Robot Play for All: Developing Toys and Games for Disability*. Springer Nature, CH, 2017.

[2] M. M. Botvinick and J. D. Cohen. Rubber hands 'feel' touch that eyes see. *Nature*, 391:756–756, 1998.

[3] G. M. Carri. Control devices for physical metaverse. Master's thesis, Department of Computer Science and Engineering - Ingegneria Informatica, Politecnico di Milano, 2023.

[4] M. Clarkson, S. Thompson, E. Bonmati, A. Blandford, T. Dowrick, and Y. Hu. Camera calibration. `https://mphy0026.readthedocs.io/en/latest/calibration/camera_calibration.html`, ongoing. Documentation.

[5] N. Diolaiti and C. Melchiorri. Teleoperation of a mobile robot through haptic feedback. In *IEEE International Workshop HAVE Haptic Virtual Environments and Their*, pages 67–72. IEEE, 2002.

[6] G. Epifani. Non-verbal communication through a robotic physical avatar: A study using minimal sensor information. Master's thesis, School of Computer Science and Engineering - Ingegneria Informatica, Politecnico di Milano, 2022. Student ID: 10578541.

[7] F. Espositi. First contact crafting authentic 'otherness' with embodied narratives through play and theatre. *To be Published*, 2023.

[8] F. Espositi and A. Bonarini. Towards a framework for embodying anybody through sensory translation and proprioceptive remapping: a pilot study. *To be Published*, 2023.

[9] Geaxgx. Depthai-blazepose. `https://github.com/geaxgx/depthai_blazepose`, 2021. GitHub repository.

[10] D. M. Hilty, K. Randhawa, M. M. Maheu, A. J. McKean, R. Pantera, and M. C.

Mishkind. A review of telepresence, virtual reality, and augmented reality applied to clinical care. *Journal of Technology in Behavioral Science*, 5:178–205, 2020.

[11] L. Hudson. Pyzbar. `https://pypi.org/project/pyzbar/`, ongoing. Python Package Index (PyPI).

[12] J. Jiang, G. Cao, J. Deng, T.-T. Do, and S. Luo. Robotic perception of transparent objects: A review. *arXiv:2304.00157 [cs.RO]*, 2023. doi: 10.48550/arXiv.2304.00157. URL `https://arxiv.org/abs/2304.00157`. Accepted by IEEE Transactions on Artificial Intelligence.

[13] A. J. Kok and R. Van Liere. A multimodal virtual reality interface for 3d interaction with vtk. *Knowledge and Information Systems*, 13:197–219, 2007.

[14] J. Laroche, L. Vuarnesson, A. Endaltseva, J. Dumit, and A. Bachrach. [re]moving bodies – a shared diminished reality installation for exploring relational movement. *Frontiers in Psychology*, 2021.

[15] B. T. Naik, M. F. Hashmi, and N. D. Bokde. A comprehensive review of computer vision in sports: Open issues, future trends and research directions. *To be Published*, 2022. Affiliations: [a]Department of Electronics and Communication Engineering, National Institute of Technology, Warangal, India; [b]Department of Engineering - Renewable Energy and Thermodynamics, Aarhus University, 8000, Denmark.

[16] V. Nandini, R. D. Vishal, C. A. Prakash, and S. Aishwarya. A review on applications of machine vision systems in industries. *Indian Journal of Science and Technology*, 9:1–5, 2016. doi: 10.17485/ijst/2016/v9i48/108433. Original Article.

[17] S. Nayar. First principles of computer vision. Lecture series, ongoing. URL `https://www.youtube.com/@firstprinciplesofcomputerv3258`. YouTube channel.

[18] E. Panelli. Odile: An expressive robotic agent for emotional exchange and information sharing. Master's thesis, School of Computer Science and Engineering - Ingegneria Informatica, Politecnico di Milano, 2023. Student ID: 953584.

[19] D. Rakita, B. Mutlu, and M. Gleicher. Effects of onset latency and robot speed delays on mimicry-control teleoperation. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, pages 519–527, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367462. doi: 10.1145/3319502.3374838. URL `https://doi.org/10.1145/3319502.3374838`.

[20] G. Riva. Virtual reality and telepresence. *Science*, 318(5854):1240–1242, 2007.

[21] M. Samad, A. Chung, and L. Shams. Perception of body ownership is driven by bayesian sensory inference. *PLOS ONE*, 10:e0117178, 2015.

[22] I. E. Sutherland. The ultimate display. In *Proceedings of the Congress of the International Federation of Information Processing (IFIP)*, volume 2, pages 506–508, 1965.

[23] K. Takeuchi, Y. Yamazaki, and K. Yoshifuji. Avatar work: Telework for disabled people unable to go outside by using avatar robots. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 53–60, 2020.

[24] A. Toet, I. A. Kuling, B. N. Krom, and J. B. Van Erp. Toward enhanced teleoperation through embodiment. *Frontiers in Robotics and AI*, 7:14, 2020.

[25] J. Wang, J. Li, S. Wang, T. Yu, Z. Rong, X. He, Y. You, Q. Zou, W. Wan, Y. Wang, S. Gou, B. Liu, M. Peng, K. Di, Z. Liu, M. Jia, X. Xin, Y. Chen, X. Cheng, X. Feng, C. Liu, S. Han, and X. Liu. Computer vision in the teleoperation of the yutu-2 rover. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, V-3-2020:595–602, 2020.

[26] A. Wellerdiek, M. Leyrer, E. Volkova, D.-S. Chang, and B. Mohler. Recognizing your own motions on virtual avatars: Is it me or not? In *Proceedings of the conference (if available)*, pages 138–138, 2013.

[27] Wikipedia contributors. Git. `https://en.wikipedia.org/wiki/Git`, ongoing. Wikipedia.

A | Appendix A

Link to the GitHub repository: https://github.com/AIRLab-POLIMI/PhysicalMetaverse

List of Figures

Ringraziamenti

Ed eccoci giunti al termine di questo viaggio. Numerose sono le persone che mi hanno sostenuto e permesso di arrivare fino a questo punto, a cominciare dal professor Bonarini, grazie al quale ho potuto coltivare la mia passione per la robotica anche in vari altri progetti sempre nel laboratorio Airlab. Questo posto sarà sempre un luogo dall'aura mistica per me, dove poter dare spazio alla creatività. Ringrazio poi Federico, il mio correlatore, con il quale ho condiviso interminabili giornate di programmazione, colorite imprecazioni e riferimenti alla serie tv Boris.

Grazie papà, che mi hai sempre supportato nonostante il percorso che ho scelto per arrivare fin qui non fosse sempre il più breve, permettendomi di arricchiare il mio percorso di studi con esperienze indimenticabili, prima fra tutte l'Erasmus.
Grazie mamma, che sei sempre stata al mio fianco anche nei momenti in cui sentivo di aver perso la rotta, sostenendomi sempre incondizionatamente.
Grazie Alessandro e grazie Valeria, che ci siete sempre stati e avete assistito con pazienza e a volte persino interesse alle presentazioni delle mie folli invenzioni.

Grazie, gracias e thank you a tutti gli amici, le amiche (e si, anche gli amori ormai passati) che ho incontrato durante questo lungo viaggio, con molti dei e delle quali ho ormai perso i contatti, ma che hanno contribuito a rendere indimenticabili questi anni fra mercoledì Open Wine al Serendepico, notti a girovagare fra le vie di Cádiz, serate Mario Kart e moltissimo altro.

Спасибо Таня, che mi hai accompagnato alla meta, motivandomi a dare il meglio.