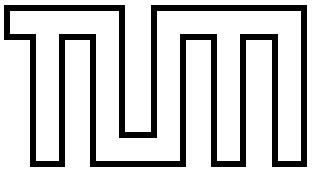


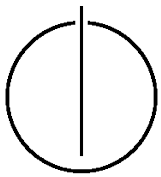
TUM School of Computation, Information and Technology

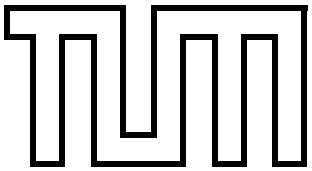
TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Enhancing the VR Experience for Older Adults using Object Tracking and an AI Companion

Anton Mai





TUM School of Computation, Information and Technology

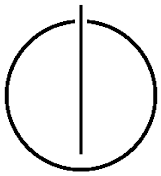
TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Enhancing the VR Experience for Older Adults using Object Tracking and an AI Companion

Förderung des VR-Erlebnisses älterer Erwachsener durch Objekterkennung und einen KI-Begleiter

Author:	Anton Mai
Supervisor:	Prof. Dr. rer. nat. David Plecher
Advisor:	Dr. rer. nat. Christian Eichhorn,
Date:	September 30, 2025



I confirm that this master's thesis is my own work and I have documented all sources and material used.

München, 30. September 2025 Anton Mai

Acknowledgments

I would like to thank my advisor, Christian Eichhorn, for his steady support and guidance throughout this thesis. Many thanks as well to Jonas Weigand, Ahmed Kaddah, and Tobias Mesmer for their help with the IMU, 3D printing, and the code templates for the AI companion and StudyTracker. Finally, I'm very grateful to my family, friends, and colleagues for participating in my study as well as for their encouragement and support, which gave me the energy to finish my master's thesis.

Abstract

This thesis explores the use of haptic augmented virtuality and artificial intelligence in virtual reality to enhance the experience for older adults The application GeoTravel was developed in the past with the intention to enable older adults to experience traveling to different tourist locations within the virtual world, even if it is not possible in the real world due to physical limitations. This thesis aims to improve the existing application by adding two functionalities:
1. Including a virtual companion in form of an animal that can communicate with the user through artificial intelligence. To guide and accompany the user, reducing loneliness
2. Enabling the user to create photos within virtual reality using a real camera with a shutter button. These goals are carefully developed into the application with the target user base of older adults in mind. A study revealed that both features show an increase in the feeling of presence within virtual reality.

Contents

List of Abbreviations

VR	Virtual Reality
AR	Augmented Reality
MR	Mixed Reality
XR	Extended Reality
AV	Augmented Virtuality

1 Introduction

In recent years Extended Reality has become more popular with improving hardware in form of various VR/AR headsets and affordable prices. While Extended Reality has been mostly focused on gaming content, it also has potential for other use cases like education or medicine. With a simple XR headset, anyone can dive into an immersive virtual world - while being at home. This fact enables people with physical limitations to gain the mobility to move around in a virtual world without any limitations. Also spatial limitations like visiting and view different places can be lifted by creating virtual copies of those locations within the virtual world. Older adults, who can't travel too much due to physical limitations can therefore visit different places without having to worry about physical strain from traveling there.
In this thesis, we will use and improve the functionalities of GeoTravel, an application that lets the user visit various foreign tourist places - copied from the real world - within the virtual world. We will especially focus on older adults as our user base and improve the current application to get a better experience by increasing immersion and well-being. For better immersion and presence, augmented virtuality is used - a technology that enables us to introduce objects from the real world into the virtual world. Therefore introducing haptic for the user, so they can feel those objects that they touch in the virtual world. We plan to use a digital camera with increased focus on its shutter button, giving the user the ability to shoot photos within the virtual environment by aiming the real digital camera and pressing the shutter button. Another small feature in development is receiving feedback on their photos to increase the feeling of being in a social environment, as well as being able to view the photos afterwards.
To increase the enjoyment and well-being of the user, we plan to introduce an AI companion in form of a pet. Research by Wells [1] has shown many positive health benefits from having a pet. While his research is only focused on physical pets in the real world, it makes sense to have similar positive effects for users to have a virtual pet. Especially with high immersion, it can be similar to the placebo effect [2]. The exact effect will be measured and determined at the end of the project during a study.
Therefore we have following goals for this thesis:
1. Using augmented virtuality haptic and object tracking with the digital camera
2. Improving the AI companion dog to guide the user through voice chat

As well as following research questions:
Research Question 1: Does interacting with an AI companion using real-time voice chat for navigation and guidance lead to higher levels of usability and presence (including social presence and reduced feelings of loneliness) compared to navigating via a UI with controller-based buttons?

Research Question 2: Does interacting with a haptic real digital camera (with functional shutter button) lead to higher levels of usability and presence compared to using a 3D-printed replica camera without functional haptic?
This thesis is structured as follows:
Chapter 2 explains the theoretical background on Extended Reality, Object Tracking, as well as explores existing research on using VR for older adults and designing AI companions.
Chapter 3 will discuss and compare relevant aspects of related works on the relevant topics.
Chapter 4 describes the current state of GeoTravel.
Chapter 5 describes the implementation process of the new features.
In chapter 6, the Survey Design as well as the results are evaluated and discussed.
Finally, in chapter 7, a conclusion is drawn with mentions of potential future work.

2 Theoretical Background

2.1 Extended Reality

Extended Reality (XR) encompasses multiple terms like Virtual Reality (VR), Augmented Reality (AR), Augmented Virtuality (AV) and Mixed Reality (MR), which describe different environments to display and interact with virtual elements. VR lets the user dive into a fully virtual world, interacting with fully virtual environments and virtual objects. AR keeps the user in the real world, but augments the real world with virtual objects and elements. AV immerses the user into a virtual world, but augments it with real objects. MR describes the combination of these approaches, either by switching between real world and virtual world depending on the situation, or by having a display part of the virtual world and part of the real world at the same time.
While these terms describe different types of realities, immersion and presence describe how well or how deep the user is experiencing that reality.

2.1.1 History and Terminology of Extended Reality

Originally, the concept of extended reality was explored in science fiction. Already in 1935 Stanley Weinbaum, a science fiction writer, explored the concept of VR in his short story "Pygmalion's Spectacles", where the main character experiences a virtual world by manipulating the characters five senses. In 1956, Cinematographer Martin Heilig build the 'Sensorama', which was the first VR machine. It was a booth that combined "full colour 3D video, audio, vibrations, smell and atmospheric effects, such as wind" to fully immerse the user into the six short films. While Heilig also created the 'Telesphere mask', the first head-mounted display (HMD) which provided images and sound, due to its lack of interaction and motion tracking, "The Sword of Damocles" is generally considered as the first AR HMD. "The Sword of Damocles", created by Sutherland and his students, was an HMD connected to a computer, that implemented head tracking by showing the user 3D models from different perspectives depending on the user's head movement. That HMD was still too uncomfortable and heavy to wear at the time, keeping it only as a lab experiment. In the subsequent years until today, multiple ideas to implement XR were tested and developed. Especially the utility of VR for cockpit simulators for planes helped with funding VR research and exploring the field further [3]. The terms themselves were coined rather late, Jason Lanier popularized the term "Virtual Reality" in 1987 [3]. In 1990, Tom Caudell and David Mizell introduced the term "Augmented Reality" [4]. And in 1994, Paul Milgram and Fumio Kishino coined the term "Mixed Reality" [5].

2.1.2 Reality-Virtuality Continuum

In 1994, Milgram and Kishino [5] visualized the different concept extended reality on a ”reality–virtuality continuum” as seen in figure 2.1. The ”reality-virtuality continuum” describes a spectrum of real environment and virtual environment, categorizing different subcategories of extended reality by how much real and virtual is blended. They also describe three aspects to differentiate between ”real” and ”virtual”
1. Objects: Real objects ’have an actual objective existence”, while virtual objects only exist ”in essence or effect, but not formally or actually”. Meaning that virtual objects have to be simulated in some way and have a description or model, because it doesn’t actually exist in essence.
2. Image Quality: With ”non-direct viewing” of an object - viewing an object through another imaging system like a video camera or similar - real objects still look real, even after getting sampled by the camera. But with virtual objects, even though it can not be sampled but only synthesized, due to current technological advances, it also looks very realistic. Milgram and Kishino are making the point that ”just because an image ’looks real’ does not mean that the object being represented is real,and therefore the terminology we employ must be able carefully to reflect this difference.”
3. Luminosity in images: Real images and virtual images still differ. While real images have luminosity depending on its location, while virtual images have no luminosity, as can be seen from examples like holograms [5].

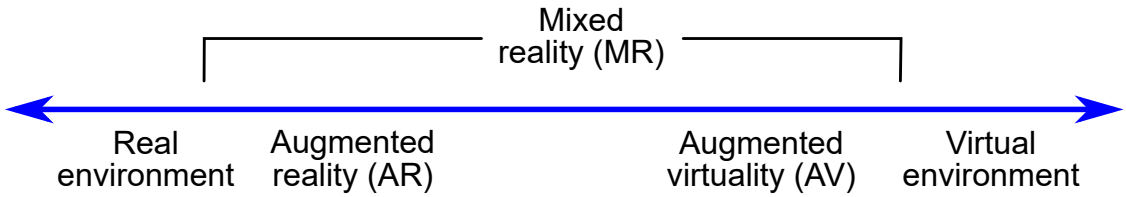


Figure 2.1: ”Reality–virtuality continuum” by Milgram and Kishino [5]

2.1.3 Virtual Reality

VR creates an immersive simulated virtual environment, making the user feel like they are in a different world than the real one and enabling the user to move within the environment and interact with the virtual content.
On Milgram and Kishino’s ”reality–virtuality continuum” as seen in figure 2.1, ”Virtual Reality” is not explicitly placed as it would just be sorted on the far right as it just exclusively uses the virtual environment.

2.1.4 Augmented Reality

In contrast to VR where the user is fully immersed into a virtual world, in AR the user stays in the real world but has additionally different enhancements to the real world, ranging from various applications like having data or instructions added, or displaying

a three-dimensional object on top of the real world. Milgram and Kishino use the term Augmented Reality "to refer to any case in which an otherwise real environment is "augmented" by means of virtual (computer graphic) objects" [5]. In their "reality–virtuality continuum", they placed Augmented reality closer to the real environment than the virtual environment because the user is generally located in the real environment and only enhanced by some virtual elements 2.1. Carmigniani and Furht also add that AR is not only applied to the user's "immediate surroundings, but also to any indirect view of the real-world environment, such as live-video stream" [6]. Therefore AR can in theory also be applied to areas where the user currently is not located. The term of "augmentation" can be interpreted in different ways. While usually, "augmentation" refers to the act of adding something, for Azuma et al. [7] the "task of removing real objects" - which is called by other researchers as "mediated or diminished reality" - is also considered as a subset of AR.

2.1.5 Augmented Virtuality

AV allows us to include real world objects into the virtual world. By doing this, the user can experience haptic feedback when interacting with the object while being in the virtual world, making the experience feel more real and therefore increasing their immersion. Milgram and Kishino refer to Augmented Virtuality as the converse case of Augmented Reality [5]. Therefore they place AV opposite to AR closer on the virtual side on their "reality–virtuality continuum" 2.1. While Augmented Reality is the display of the real world enhanced with virtual elements, Augmented Virtuality is the virtual world enhanced with real elements. Ternier et al. [8] also highlight the importance of immersion for AV applications . This makes sense as the inclusion of real objects introduces haptics - for the user the corresponding virtual objects feel much more real as they can touch them and experience real sensation from the touch.

2.1.6 Mixed Reality

MR refers to all the environments where both real and virtual simultaneously interact together, therefore it encompasses both AR and AV. As VR does not interact with the real environment, it is not a subset of MR. MR is more than just a simple sum of both virtual and real environment. By combining both realities, many more possibilities open up. While MR has the capabilities of switching between virtual and real reality, it can also combine both and show different elements of each technology at the same time, for example showing the real world but replacing the sky with virtual elements from virtual reality. Or being immersed in a virtual world, but being able to integrating real world objects into the virtual world. In the "reality–virtuality continuum" of Milgram and Kishino, Mixed Reality is not just defined as a single point on the continuum, but a range encompassing all the parts that describe a mix of both real environment and virtual environment, including AR and AV 2.1.

2.1.7 Immersion and Presence

In the context of extended reality, immersion and presence are relevant concepts that describe different user experience. While generally immersion and presence are discussed in literature, Walsh and Pawlowski [9] also present two additional concepts found in literature: interactivity and togetherness, which won't be discussed in detail here. Immersion and presence describe both the objective, measurable side as well as the subjective feeling of the user on how real the virtual environment seems and feels.
Immersion is generally seen as the objective and technical side of the system (used for creating the virtual environment) to replicate real sensations within the virtual world. Mestre [10] describes this by the "resemblance of the VR devices with human characteristics", including "the size of the human visual field, the stereoscopic aspects of the simulation, the "surround" aspects of the sound". To more precisely measure the level of immersion on these senses, Steuer [11] used the concept of "vividness", which has two dimensions: breath and depth - breath refers to the range of senses that are stimulated in VR, while the depth refers to the resolution or quality used for those senses. While it is important for immersion to replicate the sensations in the virtual world as real as possible, the user is still existing in the real world, therefore it is also important to reduce the sensations of the user from the real world, which is used in Witmer's and Stinger's [12] description of immersion: "extent to which the subject's senses are isolated from the real world and are stimulated by the virtual world". Even though they agree with the virtual environment setup being relevant for immersion, Witmer and Singer [12] disagree with the notion of immersion being purely objective, but in their view, "immersion, like involvement and presence, is something the individual experiences".
Presence is defined as "the subjective experience of being in one place or environment, even though one is physically situated in another", as described by Witmer and Stinger [12]. They also mention the need of a "coherent set of stimuli" for the user to feel like they are present in the virtual environment over the real environment. They also argue that immersion and involvement are required for presence. While it could be argued that presence is about believing in being present in the virtual world, Slater [13] disagrees with that notion, arguing that it is not necessary to actually cognitively believe in being there, saying that "it is the illusion of being there, notwithstanding that you know for sure that you are not". It matters more that the user is responding to the virtual environment even though cognitively it is known to be not real. Slater also mentions that "presence is not even about realism" [13]. In the context of presence, Durlach and Slater [14] proposed the term togetherness to extend presence for shared virtual environments that are used by multiple users simultaneously. Togetherness describes the feeling of actually being present together with other people in the same physical space. While Durlach and Slater mention that the factors describing presence are relevant for togetherness, they also note the importance of interactions with the environment in form of either being able to see other people change the environment, or collaborative work like moving heavy furniture together [14]. Therefore in a way, togetherness can be seen as presence for multiple users.
Summarized, to increase immersion and presence, we need to increase following factors:
1. Resemblance of virtual elements to real equivalents
2. Number and quality of senses involved in the virtual world
3. Isolation from real stimuli from the real world

4. Coherency of experienced stimuli in the virtual world

2.2 Object Tracking with Augmented Virtuality

Before we can interact with an object used for augmented virtuality, we first need to integrate that object from the real world into the virtual world. To achieve that, we first need to recognize and track the object in the real world, then we need to create a model of that object, and finally place and visualize it in the virtual world. After inserting a virtual copy of the object into the virtual world, its position needs to be constantly tracked. To achieve that, we need to track the object's six degrees of freedom via pose tracking. As many approaches require the use of a camera to also track the user's position and orientation, there is also an important distinction between inside-out tracking and outside-in tracking. There are different ways to categorize tracking techniques. Many projects used a hybrid approach, combining multiple different techniques to be more flexible and robust. To structure the vast variety of techniques and approaches, we use the categorization of Syed et al. [15], who simply divided tracking techniques in AR into marker-based tracking techniques and marker-less tracking techniques.

2.2.1 Pose Tracking

Tracking the precise position in Euclidian space of an object within VR is required to accurately visualize it. To track it precisely, the 6 degrees of freedom (6DoF) are required. The 6DoF in context of virtual reality describes the ability of the player or an object to change its position and rotation in space. "A single degree of freedom on an object is controlled by the up/down, forward/back, left/right, pitch, roll, or yaw." [16]. Therefore there are 3DoF for each position or linear translation by moving on the x-axis, y-axis, z-axis, as well as 3DoF for orientation or axial rotation in form of yaw, pitch, roll. As seen in 2.2, each of these 6DoF are orthogonal to each other, therefore each of these 6DoF is required to accurately track an objects position and rotation. For object tracking we need to find solutions where the 6DoF of an potentially moving object can be accurately tracked to be able to replicate it in the virtual space.

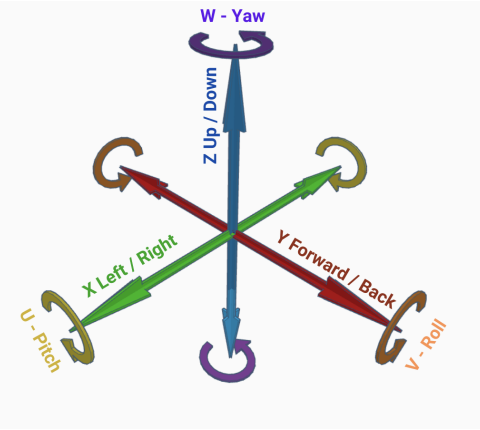


Figure 2.2: The 6DoF, 3DoF for translation and 3DoF for orientation
[16]

Especially in the context of optical tracking or vision-based tracking techniques that use cameras, the two concepts of outside-in tracking and inside-out tracking are relevant, as each application can be sorted into one concept. As shown in figure 2.3, inside-out tracking refers to the approaches where the camera is attached to the user, capturing the user's surrounding, while outside-in tracking refers to approaches where the camera's are fixed to the environment, capturing the user from outside [17].

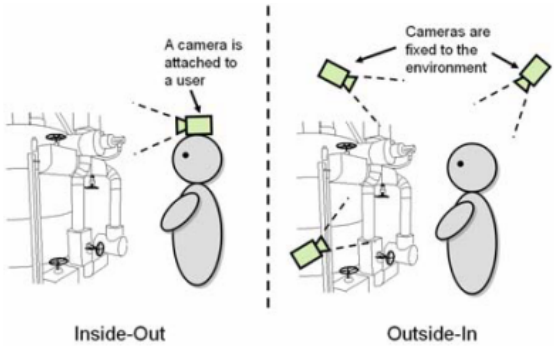


Figure 2.3: Inside-Out and Outside-In
[17]

Inside-Out Tracking

This describes all the optical tracking systems that have cameras attached to the user or the HMD to capture images from the user's surrounding. By using methods like computer vision algorithms or simultaneous localization and mapping frameworks, the user's movement and surrounding environment can be tracked, therefore updating the user's position as well [18]. A big advantage of the flexibility and ease of use - the user only needs to put on the HMD and pick up the controllers to dive into the virtual reality, without having to spend time on any external setup. Only requiring the HMD also decreases

the overall economical cost, therefore also making it more accessible for the broader public. Though on the other hand, a disadvantage of inside-out tracking is the potential lower precision in tracking due to bad visibility conditions like bad lighting. Because there are only cameras on the HMD, only the visible surrounding of user can be tracked, especially occlusion from the hands is an issue with inside-out tracking [18].

Outside-In Tracking

This describes the optical trackings systems that uses external cameras and sensors that are placed in the surrounding physical environment. By having markers or sensors on the user's body, the external cameras can detect and track the user and using that information, calculate the user's position and keep track of the user [18]. The upside of this method is the high accuracy due to having more cameras as well as the cameras being more stable because they are fixed in the surrounding. Also occlusion issues are rarely an issue due to having multiple cameras from different angles being able to cover each other camera's blind spot. The downside to outside-in tracking is being limited in the space that the surrounding cameras can cover, as well as having a much higher cost for multiple cameras as well as having to set up each of these cameras in the surrounding.

Comparison of Inside-Out and Outside-In Tracking

In figure 2.1 both concepts are compared. As we can see in the comparison, the general trade-off between inside-out and outside-in lies in having increased costs and less flexibility for higher precision and occlusion circumvention. Surprisingly, even though inside-out is much cheaper, it is not limited to a playing like outside-in.

Category	**Inside-Out**	**Outside-In**
Setup Required	None	Fix camera to surrounding
Cost	Only for HMD	For each camera
Precision	Lower	Higher
Occlusion Issues	Yes	No
Playing Space	Not Limited	Limited to placed cameras

Table 2.1: Advantages and Disadvantages of Inside-Out / Outside-Inside-Out Tracking

2.2.2 Marker-less Tracking

Marker-less tracking techniques include all techniques and approaches that try to recognize and track objects without using actual markers. There is a vast variety of marker-less tracking techniques.

Sensor-based Tracking

This approach contains all the different ways to use signals from sensors to get information on the object, like location and orientation.

Inertial measurement units (IMUs) is a wide-spread approach. IMUs combine accelerometer (tracks movement and therefore position in space), gyroscopes (tracks rotation in space) and magnetometers (tracks magnetic field) to move and place the object accurately into the virtual space. This is especially useful because IMUs are widely used in mobile phones, VR controllers and various other devices, making it very accessible. An issue with IMUs is the potential sensor drift that can happen over time due to reasons like electric noise or the integration of multiple small errors in data. Though this can be fixed via recalibration. Having IMUs in mobile phones is especially useful for this thesis, considering that tracking of the user's individual mobile phone is planned. Even though it isn't a tracking method, smartphone mirroring is a useful tool to include here. When using a mobile phone's IMUs to track its position and rotation with the goal of displaying the same mobile phone in the virtual space, it makes sense to use smartphone mirroring as well to give the user the ability to see and interact with the mobile phone screen within virtual reality. The alternative approach to display smartphone screens in VR - via cameras/vision-based tracking - runs into occlusion issues.
Magnetic tracking technology is also included by Syed et al. [15] in this context. For magnetic tracking, a tracking source is placed in the room with the user, and a sensor each is placed at the head and hand of the user. The tracking source then creates a magnetic field which, combined with the sensors, can be used by the computer to calculate the orientation and position of the sensors and then move and rotate the perspective of the user in the virtual world. This approach is very accurate. Here we don't have any issues with occlusion as well. Disadvantages of this approach is the limited range, the extra hardware needed to create the magnetic field, as well as the sensitivity to magnetic interference, making this approach impossible to use in some areas.
GPS (Global Positioning System) Tracking could arguably be counted as sensor-based tracking, as it uses signals from satellite-based sensors. Though due to its global scope and it's dependency on an external infrastructure - a satellite network - it is often discussed separately from sensor-based tracking. GPS uses signals sent by multiple satellites that orbit earth to measure the distance of the satellites to the GPS device and to calculate the position of the GPS device by using that info. Syed et al. [15] explain that the accuracy of GPS is currently up to 3m, but with new developments in satellite technology, like "Real time kinematics", the accuracy could improve to the centimeter levels. Even though GPS is widespread and integrated in many devices as well, like mobile phones and VR devices, so far due to the low accuracy of GPS tracking, it has generally been only used in combination with other tracking methods.
All the sensor-based tracking techniques require the availability of sensors in the tracked objects, which requires the tracked object to either have sensors integrated already, or to manually attach a sensor to the object we want to track. We also only acquire information like position, rotation, and movement of an object via sensors, but not its shape or form. Especially when we don't have information in advance on the object we want to track, like tracking objects outdoors in the nature like rocks and trees which can also come in different forms and shapes, sensor-based tracking not only requires a sensor to be attached to each of those objects, it also isn't able to recognize and reconstruct those objects on the spot. In this case, vision-based tracking techniques can be useful.

Vision-based Tracking

This approach includes any algorithm that uses the camera images to extract information on the environment space to enable a reconstruction of the real space into the virtual space. Due to the increase computational powers of computers, this method has become more popular. Syed et al. [15] also divided the approaches used to capture the data/images into three categories, visible light tracking, three-dimensional structure tracking, and infrared tracking.
Visible Light Tracking uses lighting and the shadow of objects to identify the shapes of objects in the environment, as well as the environment itself, creating the 3D-shapes of the objects in VR.
Three-Dimensional Structure Tracking uses depth-sensing technology, like LiDAR (Light Detection And Ranging) to measure the distance between LiDAR device and the environment. By sending light in form of laser pulses as well as measuring the time the laser takes to hit the surface of the environment and be reflected back to the source, the distance between LiDAR device and environment is measured, enabling the creation of a 3D map of the object/environment.
Infrared Tracking uses infrared cameras are used to detect heat signatures or track movements in the field of vision. While these heat signatures are invisible to the human eye, these infrared cameras can capture it. Because warmer objects emit different heat signatures than colder objects, by using the difference in heat signature between warmer living objects and the colder environment, especially living beings can be distinguished very well via infrared tracking.
Deep Learning has also recently been applied to vision-based object tracking. Especially in the recent years, these tracking algorithms combined with deep learning have improved a lot, significantly increasing the accuracy and robustness of the object tracking algorithms [19].
A huge advantage of vision-based approaches are the high flexibility in tracking objects with different shapes, as well as the ability to track objects and environment without any preparation in advance, like attaching sensors to it. Depending on the concrete tracking techniques, different types of cameras are required. Three-Dimensional Structure Tracking and Infrared Tracking require specialized cameras, while approaches like visible light tracking only require normal cameras that are widespread and integrated in most devices, like mobile phones, making it very accessible. Especially with recent advances in deep learning, issues like low resolution of camera images can be mitigated. Vision-based approaches can also suffer from lighting issues like sunlight making the images harder to recognize, potentially high computational power requirements depending on the algorithm used. The main issue with vision-based tracking techniques compared to sensor based tracking techniques is occlusion: When an object - or just part of it - leaves the user's vision, the object might won't be recognized and tracked accurately, or in worst case even stop existing within the virtual space. Of course for vision-based tracking, algorithms in form of software are required to recognize and extract the object from the camera images.

Model-based Tracking

In cases where part of the tracked object is occluded, this approach is very useful to track is more robustly. By using 3D models of objects or the environment - either created manually or derived from scans through some modeling technique - information on the object is acquired in advance, making the object more stable within the virtual environment. Compared to vision-based tracking, the virtual environment isn't mapped through real time scanning of the environment. By having information through the already created 3D model, the shape and form of the object can be very accurately be recreated in the virtual space. Even occlusion issues can be mitigated - in cases where the object is partially occluded, the object can still be fully reconstructed in virtual reality due to the 3D model being available. Also compared to vision-based tracking, less CPU usage is needed due to the 3D model of objects being known already, saving the time to predict or calculate the object's shape. While model-based tracking is very precise and useful, the model needs to be created in advance which can be very time-consuming depending on the object. Also taking away the ability to track any object on the spot without having expected and created a model of it in advance.

2.2.3 Marker-based Tracking

For these tracking techniques, some sort of marker is used to recognize and track the location of the object in the real environment. The marker has to be seen by some camera or visual sensor, therefore markers are generally vision-based. There are different types of markers that can be used, depending on whether some pattern or light is used to be recognized.
Fiducial markers use a distinct pattern, for example a checkerboard pattern to be recognized. Using these markers, the position and orientation of objects can estimated. More specialized cases of fiducial markers use encoded patterns with some information like a unique identifier. Examples of these are QR-codes, ArUco markers or AprilTags. While it is easily printable and therefore requires only little time and effort to use, it suffers a lot from occlusion issues. Especially with encoded patterns, containing information, even small occlusion or low visibility of the marker can lead to the information not being readable or recognizable anymore. This can also happen in dynamic situations, where the quick movement of the object blurs the image of the object and the marker, making the marker unreadable. Generally, occlusion is a big issue with pattern based markers, making it unusable in situations, where full visibility of the pattern can't be guaranteed. The issue of missing full visibility of the marker can be mitigated by using light as a source for markers, as light can still be recognized, even when it is partially occluded. Active markers, as well as passive markers, take advantage of that fact.
Active markers are emitting light, to be recognized. Examples of these are LED lights or infrared lights. While these markers can be used in dark environments and are not affected as much by occlusion or low visibility, the main issue with these markers is the high cost in creating and running them, as they need some sort of energy source to be used.
Passive markers set a contrast to active markers, as they do not emit any light, but only reflect incoming light, making them much more cost-efficient. While it's much cheaper,

here the issue is the need of the sensor/camera to have a light source, as well as it being a bit less visible than active markers. Here we have a trade off between energy consumption and visibility. Comparing fiducial markers with the light-based markers, we have the trade off between easy usage and occlusion issues.

2.2.4 Hybrid Tracking

While this is not a distinct tracking technique per se, hybrid tracking is often used. This simply describes the use of two or more different tracking approaches together, making the tracking more robust as well as taking the advantages from each of the tracking technique - where one tracking technique fails, another tracking approach can help. The main disadvantage of hybrid tracking is obviously the increased time, effort, setup and computing power required to develop and run each of the tracking methods.

2.2.5 Overview of tracking techniques

The advantages and disadvantages of each approach are described in table 2.2.

Tracking Approach	Advantages	Disadvantages
Sensor-Based IMUs	cheap, integrated in most hardware, no occlusion issues, low latency	can drift over time
Sensor-Based Magnetic Tracking	high accuracy, no occlusion issues, low latency	sensitive to magnetic interference, limited range, extra hardware/sensors required
GPS Tracking	integrated in most hardware, globally usable	limited accuracy, dependent on GPS signal
Vision-Based Tracking	flexible object recognition, high precision, improvable with deep learning	camera required, occlusion issues, lighting issues, high CPU usage, image recognition code required
Model Based Tracking	very robust, knowledge of 3D structure available, less CPU usage	model needs to be created in advance, model creation is time-consuming
Fiducial Markers	easy usage	occlusion issues
Active Markers	robust against occlusion, usage in darkness	requires constant light source
Passive Markers	cheaper, better against occlusion	requires light source from sensor
Hybrid Tracking	combines the best of the approaches used	each method needs to be implemented

Table 2.2: Tracking Approaches and their Advantages and Disadvantages

As we can see, there is a good variety of different techniques that can be used for object tracking. We can generally group the approaches into three categories: Sensor-Based Tracking (including GPS Tracking), Vision-Based Tracking (also including Model Based Tracking) and Marker-Based Tracking. Most tracking approaches are quite cheap, making it very possible to try any approaches when tracking. Also for hybrid tracking, this helps as we can use multiple approaches without worrying about high expenses. While Sensor-Based Tracking as well as GPS Tracking is generally hardware dependent and relies on availability of IMUs or devices like smartphones - which are still very available nowadays - Vision-Based Tracking and Model-Based Tracking generally relies on CPU usage and software being available. The main advantage of sensor-based tracking approaches is the lack of occlusion issues which can become a big problem as the user generally has the

option to look around and look at the virtual environment, making the tracked object frequently go out of vision. Generally using markers is a very viable strategy as they are very cheap and easy to use, but they generally are very prone to occlusion and lighting issues.
Using a hybrid tracking approach makes most sense - to address the shortcomings of each tracking technique while still leveraging the potential of each approach. This is also quite viable due to the generally cheap costs of each approach. Having multiple approaches at once can also help with redundancy, adding more safety with backup if any issues arises like for example slow WiFi causing latency issues. For example, one hybrid approach could be using sensor-based IMU's to track orientation together with fiducial markers for position.

2.3 Virtual Reality for Older Adults

Virtual Reality can be very beneficial for older adults. As many older adults struggle with declining health and more limitations on their physical abilities, like mobility issues and lower stamina, it's harder for them to move and experience parts of the real world. For example moving via transport like plane for a long duration can cause a lot of strain on their physical health. VR can help mitigate that issue, giving the adult a tool to still move around in a virtual world and visit many beautiful places without having to worry about any of the physical issues. Though older adults have a higher aversion to newer technology like VR in comparison to the younger generation due to the high complexity and therefore higher trouble using it as well as not growing up with the technology and therefore being less familiar with it. Therefore this has to be considered when designing the different features of the application, making it more accessible as well as enjoyable for adults.

2.3.1 Positive Impact of Virtual Reality on Older Adults

Older adults more often live in social isolation, increasing the risk of depression and dementia. In their survey, Lin et al. [20] compared a group of older adults, using VR to actively watch the content, while another group used the television to watch that same content. Throughout their survey they measured emotional, physical, and social well-being of the participants. They found that the participants of the VR group were feeling less isolated and depressed afterwards, showing higher levels of emotional, physical and social well-being. Moore et al. [21] did a study on how well older adults between 65 to 103 experience VR as well as their caregivers. They all enjoyed VR experience, though with increasing age, the enjoyment sunk due to the lower sensory and physical limitations of the participants. Though compared to other technologies like smartphones, their attitude was less negative and the decline in attitude was much less steep, showing the high potential of VR to be accepted and enjoyed by older adults. In a study by Brimelow et al. [22] they used VR in reminiscence therapy, making the participants revisit familiar places to evoke past experiences and memories. Due to the higher immersion through VR, this is easier to achieve than with other methods. Most of the participants in their study enjoyed

the VR experience and were interested in doing it again. They measured through factors like facial expressions and verbal engagement, that it led to reduced apathy.

2.3.2 Design considerations with focus on Older Adults

In their scoping review, Ijaz et al. [23] summarized following design considerations for developing immersive VR applications for older adults, as described in following table

Design Consideration	Description
Onboarding and Assistance	Tutorials and support to reduce anxiety and help with getting familiar to VR
Safety	Physical safety by using stable setups, clear boundaries. Prevent falls and accidents
Embodiment	Realistic avatars and natural movement to enhance user presence and reduce disorientation.
Visual Design	Simplicity, avoid clutter. Use relatable environments to help them feel more comfortable
Audio Design	Spatial audio for navigation cues, clear instructions, and feedback sounds to enhance user guidance and immersion
Realism	Authentic tasks and scenarios to enhance immersion and presence
Personalization	Allow customization of narratives, environments, and hardware configurations to align with user preferences and reduce cognitive strain
Usability	Simplify interfaces, minimize controller complexity, large fonts, haptic feedback for interaction support
Engagement	Gamification elements like rewards or levels to maintain user interest
Minimizing Side Effects	Use session time limits and breaks, to avoid motion sickness, anxiety, or over-exertion

Table 2.3: Design Considerations for Immersive VR Applications for Older Adults

2.4 Companion Pet with Artificial Intelligence

Having a pet as a companion has many positive effects ranging from in-game-relevant features like guiding the player through the virtual environment to providing help whenever the player needs any assistance or feels lost. Also the social and mental component of having a pet is not negligible. Being accompanied by a pet helps not only with the feeling of loneliness, but can also increase the players comfort and well-being by a lot. Different people have differing preferences in pet species, character of the animal, and more, making the design of the companion extremely relevant. Therefore we need to consider many different options in look and character of the pet as well as its abilities and role within the virtual world and the ”story” of the application. Certain psychological effects like

the ”uncanny valley effect” need to be considered as well. This effect will be explained in section 2.4.3. Because the application is more tailored for older adults, certain design choices need to be discussed with extra care, like making sure that the pets are proactive and helpful, as many older adults struggle with more technical tools and user interfaces.

2.4.1 Positive Effects of Companions

Being in company of pets has long shown a positive effect on mental health as well as physical health. Based on research by Wells [1], pets have health benefits in a vast amount of areas. Just petting an animal has shown to decrease blood pressure and heart rate, as well as reduce stress significantly. Also long term effects have been researched, people with a pet are less likely to have ill health and recover faster from it. Pets are also very useful as a potential early detector of diseases, as it has been shown that more owners with pets have been warned through their pets on diseases like cancer, seizure and hypoglycemia. Pets are also helpful for psychological help. In general, having them present reduces feelings of loneliness and isolation as you always have companion from your pet. Additionally, pets help foster and facilitate social connections between people by a lot. They can help in socializing, as walking with a pet attracts attention and interest from other people and makes it much more likely for people to talk to a pet owner. Therefore making it easier for the pet owner to connect and make friends with strangers. Because the positive effects of having a pet are not available to everyone due to different reasons like allergies, hygiene restriction or other limitations, Kulik et al. [24] created a virtual dog in VR, to demonstrate the possibility of making pets accessible to everyone through VR.

2.4.2 Design Choices for Companion with Focus on Target Group of Older Adults

There is a variety of characteristics to consider when creating a companion. To explore the different characteristics systematically, we use the categories design space by Bouquet et al. [25] who edited some categories from the design space which was originally created by Emmerich et al. [26]. Originally Emmerich et al. [26] had the categories ”General Characteristics” which included the companion’s appearance, personality and agenda. Bouquet et al. decided to split up that category into two new categories, ”Appearance” and ”Individuality”. They also renamed ”General Capabilities” into ”Sentience”. In the following, the different categories and subcategories (called ”characterists” by Bouquet et al. [25] and Emmerich et al. [26]) will be briefly described in table 2.4.

Main Category	Subcategory	Description
Appearance	Visual Appearance	Physical appearance like animal type/human, size, gender.
	Auditory Appearance	Sound-based communication - voice and speaking habits
Sentience	Awareness	Ability of companion to perceive the surrounding environment and talk or act based on it
	Emotional Intelligence	Ability to understand emotions and react the realistically
	Social Relations	Capacity to foster relationships
Individuality	Personality	Traits like introversion or humor
	Own Agenda	The companion's personal goals and motivations
	Background	The companion's backstory, past experiences and relationship
Behavior	Context Sensitivity	Ability to adapt behavior to different situations
	Autonomy	Degree to act independently without player input
	Initiative and Activity	Degree of proactivity in engaging with the environment or tasks
Communication Capabilities	Communication with the Player	Ability to interact with the player, like through dialogue or gestures
	Communication with Other NPCs	Interactions with other in-game characters
Relation to the Player	Interdependence	How strong the player and the companion are relying on each other
	Power Dynamics	Balance of power between player's and companion's relationship
	Obligations	Social/Narrative ties like responsibilities between player and companion
Significance	Story Relevance	Relevance of companion within the story's narrative
	Gameplay Relevance	Companion's relevance to gameplay mechanics

Table 2.4: Companion design categories and its description

2.4.3 Uncanny Valley Effect

The term ”uncanny valley” was coined in 1970 by Masashiro Mori [27]. It describes the negative effect in affinity from the user that occurs whenever a human-like robot looks extremely alike to a real human, but still shows some signs that makes it clear to the observer that it is just a robot, making it feel very eerie.

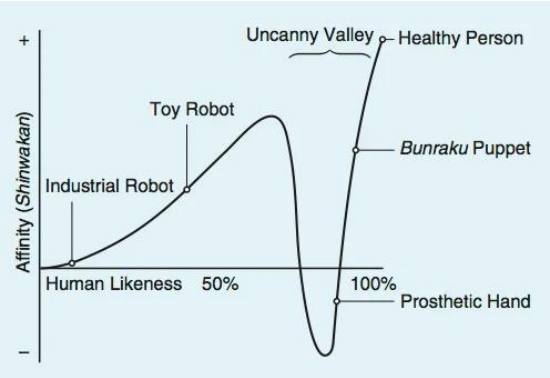


Figure 2.4: ”uncanny valley” shown by Masahiro Mori [27]

As we can see in figure 2.4, instead of having a linear function that increases affinity with increasing likeliness to a real human, there is a steep valley near the robot being exactly like the real human being. This is an issue when designing animate objects in virtual reality as it can cause less affinity of the user and discomfort if we try to make a realistic copy of a companion, but fail to make it realistic enough. While the uncanny valley effect has been well researched on humans, it’s equivalent to animals has been less researched. Rativa et al. [28] researched whether an uncanny valley effect was present for pandas by showing participants different images of pandas that were slightly off. While they showed that most aspects of the uncanny valley effect like familiarity, commonality, naturalness, and attractiveness still applies for pandas ad feel off and unsettling, the fake robotic pandas were still received as entertaining and lively as the real pandas. Therefore we can see that the uncanny valley effect still needs to be considered when working with a virtual animal. To prevent the uncanny valley effect, Mori suggests ”to create a safe level of affinity by deliberately pursuing a nonhuman design” [27]. By making the design clearly non-human, people don’t feel the eerieness of the near human-like robot, because they clearly know that it isn’t supposed to be human. Mori also gives eyeglasses as an example of a good design - they don’t look like real eye-balls but have their own unique design. Therefore we should follow that principle and come up with stylish designs instead of making it too realistic.

2.4.4 Large Language Model

Large Language Models like GPT-4 [29] are neural networks trained on huge amounts of text data to be able to predict the next word in a sequence. This training enables it to learn liguistic patterns, enabling it to generating coherent, human-like dialogue in real

time. One popular such known model is OpenAI's ChatGPT, which will be accessed via Realtime API [30] in this thesis, enabling the AI companion pet to hold a conversation with the user.

3 Related Works

This chapter explains the theory and concepts required in this thesis, as well as showing other works in this area and the concepts used there. Having this foundation allows us to decide which approach is best suited to develop the features for the GeoTravelApp in this thesis.

3.1 Augmented Virtuality with Haptics

Various other researchers have used different methods and approaches to use augmented virtuality with focus on haptics.

3.1.1 Haptic Objects

Danieau et al. used haptic feedback in combination with cinematography [31]. They let the user sit on a chair that they called 'HapSeat' that gave the user haptic feedback. They used three force-feedback devices (Novint Falcons) to stimulate the user's head and hands. As the haptic device actively stimulates the user, this is classified as active haptics. They tested seven different cinema sequence that had some movement, in combination with different haptic sensations. Their results showed that especially cinematic haptics - haptic feedback that was coupled and consistent with the shown camera movements, making the user feel the same as what is shown visually - increased immersion and cinematic experience by a lot. Interestingly, they also tested semantic haptics, which is feedback depending on the actions and mood of the scene. There the results were rather inconsistent, which might also be due to some of the scenes making the user more uncomfortable. An important takeaway here is to try to keep haptic feedback consistent with visual movement to increase immersion.
Davari et al. [32] used everyday proxy objects and 3D-printed objects as passive haptics - objects that do not exercise any movement or force on the user, but are only there to be picked up by the user and interacted with. They tried to reuse the proxy objects in different VR contexts to see how well they can improve the user's immersion. One of the objects was a 'Hint camera', which is similar to the digital camera we use in our project. They mention a need for virtual object and real proxy to be consistent in shape, size, and texture to increase presence, otherwise risking a decrease in perceived realism. Very interestingly, participants used the same proxy object in different VR contexts without noticing the difference, showing that a proxy object does not have to be perfectly one-on-one to its virtual counterpart to work. They also point out that good storytelling can enable illusions and help the user overcome the difference between proxy and virtual object. This

shows as well that realistic, potentially expensive items are not a must, but cheaper proxy alternatives might be working well too. This will be a useful insight for our study where we use a 3D-printed replica camera in comparison to a real camera.
Similarly, Fang et al. [33] also tried to use everyday proxies as a cheaper and more available option. They conducted an in-the-wild study, meaning that the studies were conducted at the participants home instead of a research laboratory. They created 3 different VR experiences where the users could use their items at home to simulate virtual objects (or in this case a cat as well). The proxy items were a chair that was turned into a cannon in the virtual environment, a table turned into a whack-a-mole playing area, and a pillow turned into a cat. They used hand tracking as their system. It was shown that the user's immersion and presence was increased clearly through the use of haptics. They mention how the texture of a soft pillow could feel realistic like a cat's fur when petting it, therefore increasing immersion. However, they also point out that the current limits of object tracking can decrease immersion.
A limitation from 3D printed objects was noted by Krumpen et al., who used 3D printed artefacts for a virtual museum: While there are different materials that can be used for 3D printing, it is not possible to mix different materials in one model [34]. Depending on the object, this restriction could cause lower immersion.

3.2 Virtual Reality Design for Older Adults

There have been various virtual reality applications designed specifically for older adults. There are many virtual reality exergames, which is the combination of "exercise" and "games". That type of application aims to provide exercise in form of a game. That way, people are more motivated to do those exercises, staying fit and healthy.´ One example of the exergames was developed by Kruse et al. [35]. They created "Maestro Game VR", which is a virtual reality exergame, in which the player is located in a 3D concert hall and takes the role of a conductor. The player's task is to conduct the band by following a virtual path in front of them with their virtual baton, by moving one arm at a time. Kruse et al.'s [35] goals were to provide an alternative to exercising with 2D videos for older adults, as well as figure out, whether exergames can be better than traditional exercise videos. Their results showed that there is no significant difference in enjoyment, workload and attention between both alternatives. Of the 25 users, 13 users preferred the traditional video because it was more realistic and they also had a harder time getting used to virtual reality, while 9 users preferred the VR exergame because it was more fun and interesting. That showed that exergames can be a useful alternative for some user's that prefer VR games to traditional videos.
Another exergame was developed by Karaosmanoglu et al. [36]. They created "Memory Journalist VR", where the player takes the role of a reporter who takes photos of landmarks. While they play that role in virtual reality, the health professional can use a Remote App to guide the player and control the VR experience. Goal of their study were the questions whether VR exergames provide positive player experiences for older adults with dementia, as well as find out which aspects of Human Centered Design approach - an approach where the shareholders and the stakeholders, like the older adult with dementia and the health professional, are included into the design process - should

be adapted when developing for older adults with dementia. Their application showed promising result, also showing that it makes sense to use VR exergames for users with dementia. They also created three guidelines that seem reasonable to follow in that context:
1. Create a social environment with consideration of shared aspects.
2. Support a game flow, that addresses the decline and variance of the users cognitive and physical skills
3. Providing a safe VR experience
They also highlighted the importance of applying the Human Centered Design approach, including the future users into the design process.

3.3 Pet Companion for Older Adults

There has also been some research on creating pet companions specifically designed for older adults as well.
Wanali et al. [37] also used a virtual dog and measured its effect on social presence as well. Though it has to be mentioned that they mostly worked with younger women instead of older adults. They found that the virtual dog improved social presence with similar high levels like for studies with humans. They also found that depending on whether the task type is in a cooperative nature with animals showing reactions, social presence can be increased even further. Only with really with social presence they found that it can reduce stress and negative feelings, therefore using virtual animals with high social presence levels could improve therapy with animals. Kruse et al. [38] integrated an AI assistant of a parrot with GPT-4 into their VR exergame for seniors, to support the player during their gameplay. The parrot was positively received, with participants even mentioning being comfortable playing with the parrot alone. There were also some limitations due to high delay or the parrot having misunderstandings during their communication. Also, players often tended to focus more on the gameplay itself than the assistant. They also stressed that it is important to clearly communicate what the AI assistant can do and how to communicate with them, because that might not be always clear. Cho [39] used AR to create a virtual pet for the user. They wanted to create a cheaper option as well as a more adaptable option compared to the expensive robot pets that existed before. The user has the choice of picking the kind of pet they want through an application on their smartphone, then the smartphone uses image data obtained through AR glasses to display that virtual pet around the user. And via AI from an external server, the pet has movement based on the user's actions. They used the Unity engine together with Vuforia for AR and OpenCV to develop their system. Though, "due to the lack of development manpower", they didn't "provide customized services by interpreting the psychological and health conditions of the elderly through artificial intelligence using an external server" [39]. Norouzi et al. [40] also used AR to create a virtual dog that can interact and walk with the user. They collected the user's behavioral data in form of info on their head position and orientation, while the user is wearing a Microsoft HoloLens. They used Unity with an already existing 3D model of a virtual beagle and used a server-client format, letting an experimenter control the beagle's behavior. They implemented a big range of dog behavior as well as actions for the user to take. Ai et al. [41] used pet dogs and horses

in an VR environment for Alzheimer's disease therapy, which is again more cost-efficient compared to therapy with real animals. Though due to the pandemic at the time, they only did tests with five lab students. Therefore there is a lack of sample size in testing on older adults. While they created a virtual reality environment, it isn't explained clearly, which development tool was used for that.
Especially in context of older adults, each of them developed some functionality for the needs of the older adults. Though they didn't describe the design considerations of the pet in context of the pets in detail. While Cho [39] gave the user multiple options to pick from, Norouzi et al. [40] had a single model of a dog, and Ai et al. [41] also had a single model each of a dog and a horse. Cho [39] created a system "that provides psychological comfort, communication, rehabilitation training, and convenience functions that are essential for the lives of the elderly living alone" with the pet having three functions: 1. Emotional exchange - to help with loneliness and depression and keep their psyche stable, 2. daily care service - informing them of medication guidance, meal times, and appointments, 3. Rehabilitation/Emergency Response Service - helping with rehabilitation from chronic diseases as well as providing guidance in emergency situations. While Norouzi et al. [40] created their companion pet with the goal of "effects of human-animal interactions on human perception and behavior, for example, stress reduction, task performance, and well-being together with support animals" They proposed to track the user's heart rate and speed via a fitness tracking device to be able to have the dog adjust its speed or take breaks if needed due to a high heart rate. Ai et al. [41] created their companion pet to help patients with Alzheimer's disease. They came up with three ways to communicate with the virtual pet: 1) via 3D-buttons, they can give the pet instructions. 2) via speech recognition on whether the user issues an instruction to the pet. 3) via a neurofeedback system, where they used EEG headsets to track the user's emotions, as well as a Reinforcement Learning model to act accordingly, having the pet change proximity based on the user liking the pet or not. They have also measured the user's negative emotions, which sunk slightly during and after the therapy, showing a positive effect of the therapy.
The different functions and goals as well as special features of each project is shown in 3.1

Paper	Pet companion	Functions/Goals	Special Features
Wanali et al. [37]	dog	playing with pet	interaction possibilities
Kruse et al. [38]	parrot	support during exergame	GPT-4 based
Cho [39]	custom	psychological comfort, rehabilitation and help with daily tasks	custom picking of pet
Norouzi et al. [40]	dog	measuring stress reduction, task performance and well-being	pet actions controlled by experimenter, heart rate measuring
Ai et al. [41]	dog and horse	Alzheimer's disease therapy	speech recognition and neurofeedback system

Table 3.1: Related Works to Older adults with pets

3.4 Object Tracking Algorithms and Tools

There is a variety of algorithms and tools available to track objects, most of them require some kind of camera feed to visually see the tracked object and identify it within an image. In the following these algorithms and tools will be shortly explained, then a test setup is created to compare these algorithms based on criteria that are relevant for this project.

3.4.1 Fiducial Markers

A very simple and easy to use approach to object tracking are fiducial markers. These markers use distinct patterns to identify and estimate its position, while also containing encoded information unique to each marker. It is very easy to use them by just printing them on paper. It is also not only limited to a single marker, but as stated by Olson [42], "Fiducial systems also are designed to detect multiple markers in a single image. Because the markers work in combination with a camera viewing them, they can suffer from bad conditions with lower visibility like bad lighting conditions. Occlusion can happen frequently as well with markers when they aren't fully visible.

ArUco markers

ArUco markers are fiducial markers that were introduced by Garrido-Jurado et al. [43] to solve the camera pose estimation problem - figuring out where the camera is located (position) and how it is tilted (orientation). Square markers were popular due to the four corners enabling a good calibration of the camera which is why they also used that shape for ArUco markers.

They tried to solve following problems at the time through ArUco markers [43]: 1. Defining a dictionary - a set of valid markers - lacked efficiency because fiducial markers at the time had a fixed number of valid markers. An application might require a higher number of markers than the dictionary size, therefore lacking enough different ArUco markers to cover that, or the application might require a lower number of markers, having the inter-marker distance lower than necessary [43]. The inter-marker distance describes how different two markers are from each other, therefore a higher distance helps with misidentifying a marker for a different one. 2. Occlusion issues, which arise when the marker is not visible.
To solve these issues they did following [43]: 1. They created an algorithm to automatically create dictionaries of any size and maximizing inter-marker distance while doing so 2. They also presented a method to detect markers and correct errors. This consists of 4 steps - Image segmentation, Contour extraction, Marker code extraction, Marker identification and error correction - where they designed a new approach on the last step
3. To mitigate occlusion issues, they allow for multiple markers to be used at the same time, therefore having other markers available, when one marker is occluded
Example of ArUco markers can be seen in figure 3.1.

Figure 3.1: Example Aruco marker of different sizes shown by Garrido-Jurado et al. [43]

AprilTag

Olson introduced AprilTags, another type of fiducial marker which uses a 2D barcode-like tag [42]. Compared to other markers that require to have information about how the tag is rotated, AprilTags can be read from any orientation.
The system has 2 main components: 1. Tag detector: This component tries to estimate positions of potential tags in an image. As described in simple words by Olson, "the detector attempts to find four-sided regions ('quads') that have a darker interior than their exterior. The tags themselves have black and white borders in order to facilitate this" [42] 2. Coding system: This component has 4 goals:
" • Maximize the number of distinguishable codes
• Maximize the number of bit errors that can be detected or corrected
• Minimize the false positive/inter-tag confusion rate
• Minimize the total number of bits per tag (and thus the size of the tag) " [42]
Examples from the APRIL robotics lab can be seen in 3.2

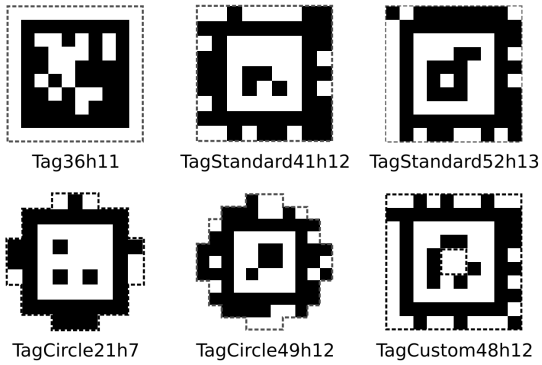


Figure 3.2: Apriltag Examples from the APRIL robotics lab [44]

STag

Benligiray et al. [45] proposed STag, a fiducial marker with the focus on increased stability, because many previous markers had detection issues with jitter or movement. While lines used in square markers are useful for pose estimation, they are more unstable. Meanwhile, ellipses are much more stable for localization, but are insufficient for pose estimation. Therefore STag tries to compensate both issues by combining both shapes and using them at the same time, using an outer square border and an inner circular border. Inside the inner circle, the information is encoded via 48 disk-shaped bit representations. The whole detection process works simplified as follows: In a first step, candidate markers are identified from the image by finding edges and lines and checking whether they make up a square, removing false candidates in the process as well as gaining a first homography. Afterwards by having the positions with squares which could potentially be an Stag, the inner ellipses are localized and used to refine the homography, increasing the accuracy and reducing noise of the Stag marker. After gaining a very stable pose estimation, the information inside the inner ellipse is decoded and processed. A graphical explanation of the detection algorithm was provided by Benligiray et al. as seen in 3.3.

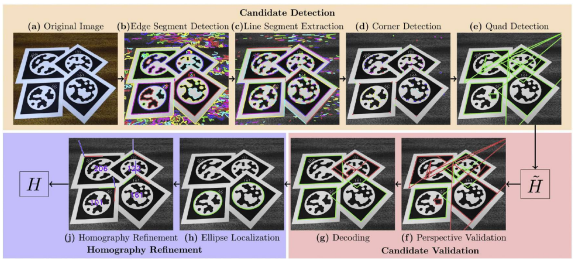


Figure 3.3: Stag Detection Algorithm by Benligiray et al. [45]

3.4.2 Deep Learning via YOLO

YOLO is a object tracking algorithm based on deep learning that was proposed by Redmon et al. [46]. Other than prior approaches that used classifiers for object detection, YOLO approaches object tracking as a regression problem on bounding boxes - boxes as a subset of an image with an object inside of it. By only having a single Convolutional neural network to predict the bounding boxes and probabilities in only one evaluation, this approach is extremely fast. YOLO's workings are explained as follows:
The input image is divided into a grid of size S x S. The grid cell in which the center of the object is located, is responsible for detecting that object. Then each grid cell predicts the bounding boxes and confidence score - which reflects the certainty of the object being in that bounding box. That prediction is done through a convolutional neural network. As multiple bounding boxes for different objects can overlap on a grid cell, the grid cell with the highest confidence is picked to be "responsible" for detecting that objects.
They also mention some constraints on YOLO [46]:
Because a grid cell can only have one class or objects that is detected, if multiple small objects appear, it struggles to track all of them due to the spatial constraint. Also during training, errors in small bounding boxes are weighted the same as errors in large bounding boxes. A simplified graphical explanation of the algorithm was also provided by Redmon et al. [46] which can be seen in 3.4.

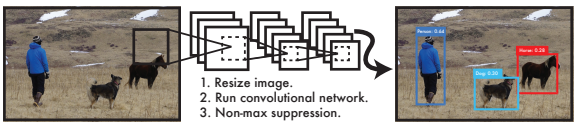


Figure 3.4: Simple graphical explanation of YOLO by Redmon et al. [46]

3.4.3 Sparse Gaussian Approach

Stoiber et al. proposed a novel, highly efficient sparse approach to region-based 6DoF object tracking based on a probabilistic model [47]. Compared to other tracking algorithms that analyze every pixel of an image, the key idea of this approach is to focus on the sparse information of the contours along the object, which are defined as correspondence lines. Through probabilistic formulas, it is decided which part is part of the object and which not. By formulating this problem in a way that follows a Gaussian distribution, efficient optimization methods can then be applied. Based on advanced mathematical concepts on proofs, Stoiber et al. have shown that this approach outperforms other state-of-the-art region-based methods at the time, while still being robust and faster.

4 Current State of GeoTravelApp

GeoTravel is a VR application originally developed by George Nassif for his master's thesis [48] and afterwards further improved and extended with multiple features by multiple people. As many older adults struggle with loneliness and deteriorating mental health, due to not being able to physically move much and travel, the goal of the GeoTravel application is to improve the well-being of older adults as well as fight loneliness by providing travel opportunity and social interaction within a virtual world. The application enables the user to travel to different locations and do sightseeing in VR. It also features a virtual dog as a companion as well as haptic feedback while using multiple objects through augmented virtuality to increase immersion. This chapter describes the past development of the GeoTravel application and gives an overview on how the application works before adding any improvements by this thesis.

4.1 History of Development of GeoTravelApp

Multiple students worked on the GeoTravel before. In 2020, George Nassif created Geotravel. At the start, only Munich and Egypt were available locations with landmarks using 360° images to view it in VR. For navigation, a controller designed as a walking stick was used together with floating UI buttons and attraction buttons to navigate to different landmarks. Also, a multiplayer mode was implemented to allow for two people to travel together [48]. A year later in 2021, Alexander Williams expanded the application by adding a three-dimensional menu, where users can navigate to different landmarks by clicking on three-dimensional models of it. For the multiplayer mode, the other player's appearance was improved. He also added Augmented Virtuality to connect the users to the outside world [49]. Also in 2021, Julia Schwan expanded GeoTravel by adding Hamburg as a location, as well as adding a search function that finds locations over the flickr API, giving the user the freedom to choose whichever place they want. A companion dog was also included to guide the player around by moving to the next attraction button [50]. In 2023, Denitsa Aleksandrova Asova then added a realistic virtual avatar to the player. Also more locations were added: Paris, London, and Turkey. While Paris and London only used 360° images, Turkey also had 360° videos and audio. Augmented Virtuality was also improved by using Aruco and Vuforia markers on three objects to track them - water bottle, camera, and smartphone [51].

4.2 Application Components

In this section, different components of the application are explained for the reader to have a clearer overview on what the application looks like and how it is designed and used.

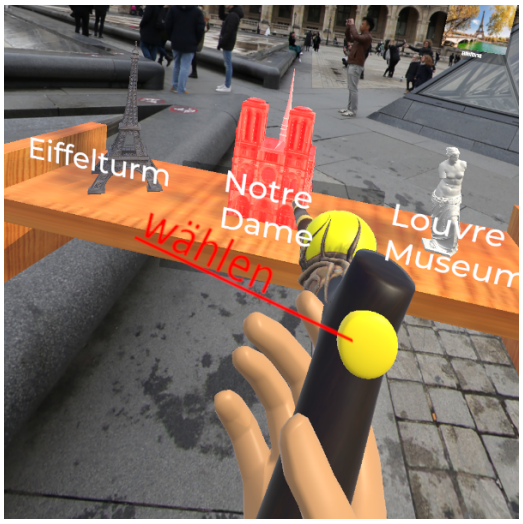
4.2.1 Travel Application

The application starts with the user being on a mountain. The user can first choose between single player mode and multiplayer mode. We did not use multiplayer for this thesis, therefore it is not shown below. The following will focus on single player mode. For navigation, the user can control a walking stick with an attached laser pointer at the tip of the stick to click on floating menu buttons. The user can click on the locations button labeled "Orte" to open the menu with the five locations: Paris, Munich, Hamburg, London, Egypt, and turkey. Turkey was also not included to keep the application simple and stay focused on the implemented features as well as due to the video files being too big and therefore not available. Depending on the location, the user can also click on the attraction button labeled "Sehenswürdigkeiten" to choose between the three landmarks within a location. Optionally, they can also use the three-dimensional navigation menu to navigate between the different landmarks.

Starting mountain scene | Floating UI | Pickable Locations

Figure 4.1: Starting scene and floating UI menu

Choice of landmarks



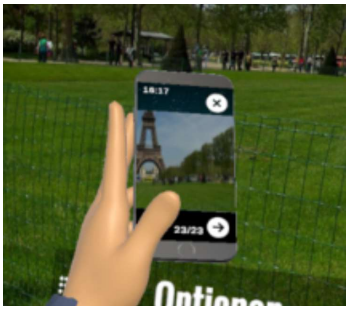
3D menu

Figure 4.2: Navigating to landmarks

4.2.2 Augmented Virtuality

Three AV objects were available. water bottle, smartphone, and camera. The water bottle is mainly there to touch and hold, drinking is not possible yet. The smartphone has basic functionality in game with a home screen, camera app, gallery app. The camera allows for basic picture shooting. At the current state though, Aruco and Vuforia marker need to be stuck with an edge of the marker on top of the object to work properly, therefore introducing some issues with holding the objects properly. The virtual versions of these objects can be seen below in 4.3.

Virtual camera



Virtual smartphone

Virtual bottle

Figure 4.3: Images taken by Denitsa Asova [51]

4.2.3 Virtual Companion

Currently, the dog seen in 4.4 was created by Julia Schwan [50] has no audio and only guides the player by moving to the next UI button in the current scene. There are different

waypoints to make the dog walk to certain locations, and if the player has visited all landmarks in a location, the dog lies down to signal that.

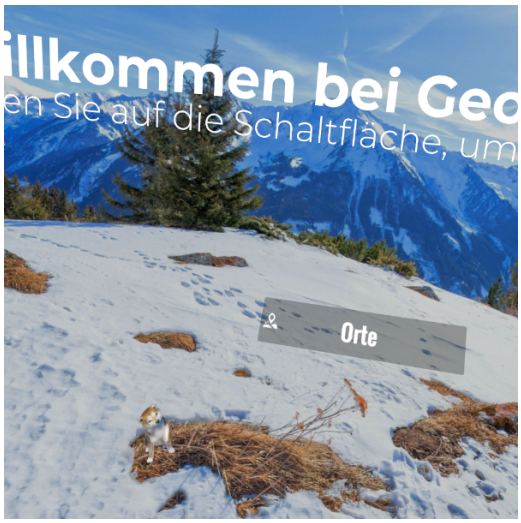


Figure 4.4: Companion dog created by Julia Schwan [50]

5 Implementation

There are multiple implementation goals for this thesis. For the setup, the current packages of the code for the application were updated. Especially Vuforia modules were removed as they were outdated and caused bugs. There have been some considerations to continue developing the application on a desktop with a Linux operating system installed. However, the application was originally developed for Windows, which made many used modules only compatible with Windows and some dynamic linked libraries not work on Linux without modifications, it was decided to use a Windows 10 as the operating system. It is heavily recommended to continue development on a Windows operating system on future developments as well. Afterwards the planned features were implemented and added to the project to increase immersion and presence for the users as well as engagement.
The first feature was the addition of an advanced AI chatbot connected through the already available companion dog to enable the user to actually communicate with the virtual companion pet by talking to them with their voice. Another feature implemented was the inclusion of a digital camera as an AV object, to include haptics into the application. The camera includes a button that can be used to make photos within the virtual environment. Connected to that, a HUD for the camera was created to include a zoom mode, trigger sound effects and shutter animations to make the process more realistic. An audio feedback was also included to simulate the impression of sharing the photos with family or friends and getting voice messages as replies, decreasing loneliness in the process.

5.1 Development Setup

Various Hardware Equipment was used to develop and run the GeoTravel application. A Meta Quest 3 was used as the VR headset. For development a laptop was used with following specifications: CPU with Intel Core i7-8750H, 6 cores and 12 threads with 2.20 GHz. 16GB RAM, and the graphics card Intel UHD Graphics 63. The operating system was Windows 10. Unity version 2021.1.3f was used for developing and running the GeoTravel application.
The laptop was equipped with an HD 720p Webcam, but another external webcam Logitech Brio 4K was used for better image quality required for object tracking. For haptics, a nonfunctional Canon PowerShot A2100 with integrated IMU (Seeed studio xiao nrf52840 sense) was used that could also track the button press of the camera via Bluetooth connection. Another 3D-printed replica of the same digital camera was created for the study. A USB Microphone (Jeemak PC21 USB Microphone) was used for increased audio quality when talking to the AI. At last, Stag markers were printed and glued on the digital

camera and its 3D printed replica for object tracking.

5.2 Implementation of AI for Companion Pet

For the companion pet the previously already implemented virtual dog by Julia Schwan [50] was used. Until now the dog didn't have any audio and only moved around in the current location to the next attraction button. The pet accompanies the user on their journey, providing assistance as well as mental support. We add an AI chat bot to the virtual dog to enable the user to actually talk and have a social interaction. To enable this, we implement OpenAI's Realtime API to talk by OpenAI within our application. The dog can be seen in following figure 4.4.

5.2.1 ChatGPT by OpenAI

Our goal for the AI chat bot is for the user to be able to use their voice to talk to the virtual AI companion, then have the virtual AI companion also reply to them through audio. The reply should happen fast without letting the user wait for long. Also, the spoken language will be English for this thesis. Though, switching languages is very simple, therefore using German would also be easily possible if needed in the future.

Realtime API

As explained in the documentation of OpenAI's audio section, there are two approaches available to us. We can implement this in three steps, using Speech-to-Text to transcribe the spoken user voice input into text, then using that text on a Large Language Model to get a reply in form of text back, which we then transform back into a voiced reply via Text-to-Speech. Doing these steps separately gives us increased control over each step, enabling us to fine-tune or switch out parts if necessary. Each of these functions are available with OpenAI's environment, through Transcription API for Speech-to-Text and Speech API for Text-to-Speech [52].
The alternative is to simply use OpenAI's Realtime API, which already combines all steps. This enables us to input audio, where all the in-between steps are handled by the Generative Pre-trained Transformer on OpenAI's side, outputting a reply in form of audio to us. This offers an easy option for development, also fulfilling our requirements for low latency and being able to chat in German.
Because we don't need a lot of extra fine-tuning, we use Realtime API for its simplicity and ease of use to create the AI chat bot for our virtual companion.
Realtime API offers two options to communicate with - via WebRTC or via WebSocket. While using WebRTC could be useful for our multiplayer mode in the future, as it supports peer-to-peer audio streaming, we choose WebSocket due to our focus on single player mode interaction with the companion pet as well as its simplicity in setup. We only need a voice chat between the user and the ChatGPT model. In the documentation

of Realtime API the difference in complexity in both approaches are also shown [30]. The WebSockets approach only requires the setup of the connection to the OpenAI key, only using the OpenAI key directly as seen in 5.1

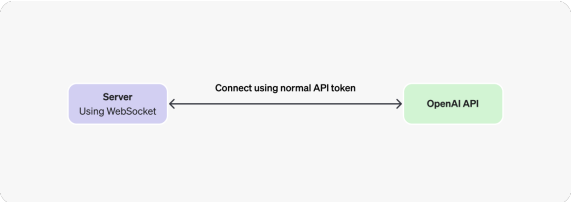


Figure 5.1: WebSocket connection for Realtime API by OpenAI [30]

The approach with WebRTC requires the user to have a server setup to be able to use an OpenAI key to request an ephemeral token - a temporary token for higher security to not expose the OpenAI key to the user on the client side. Only after obtaining that ephemeral token, it can be used by the user to set up a WebRTC connection with the OpenAI API. This process is also depicted in 5.2.

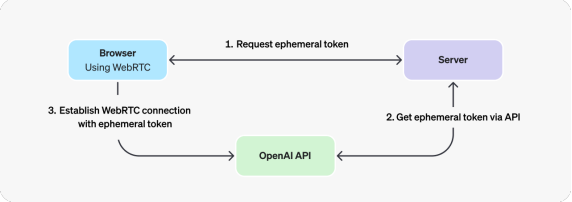


Figure 5.2: WebRTC connection for Realtime API by OpenAI [30]

However, in our experimental setup we will only deploy the GeoTravel application on our own VR headset, therefore we will not use any of these steps with ephemeral token to secure our OpenAI key.

Using the Realtime API, we can control multiple parameters to choose between multiple GPT models as well as multiple voices for text-to-speech voice outputs, and a voice activity detection option to automatically detect when the user starts or stops speaking. The specific values set for these options will be discussed in the following sections.

5.2.2 Design of Artificial Intelligence Chatbot

Visually, we will use the dog created by Julia Schwan [50]. It will have the already developed animations as well. Though we will add behaviour to the dog to turn to the player and walk closer to the player when chatting. As we might be concerned with the uncanny valley effect happening, we will follow Mori's advice to deliberately follow a less realistic and more nonhuman approach [27]. As we are already working with a non-human animal, we make the facial expressions not too realistic when mimicking the dog talking.

The user should also be told beforehand to clarify to the user that this is not a real setting, giving them the story that their VR headset has a function integrated to translate between German and animal language, making it possible to talk to the animal.

For the chatbot design, we will make use of the guidelines by Ijaz et al. [23]. It should be patient, supportive, and proactive, addressing the cognitive and physical needs of older adults by speaking slowly, giving clear instructions, and allowing breaks. The companion must be sociable and positive, aiming to reduce loneliness and increase comfort. Instructions and feedback are kept simple, while the narrative remains customizable so that users can freely choose their travel route. To support immersion, the scene avoids unnecessary UI elements or text and uses relatable environments.

Prompt and Parameters

The conversational behavior of the AI companion was defined through a dedicated prompt and a set of parameters. While the prompt is aimed at older adults, due to the study participants being students, the language was changed to English and the player was described as a student. Through tool calls, the Realtime API can access functions within GeoTravel to control various actions like changing the location or attraction when the user tells them to do so. The prompt should emphasize applying patience, calm guidance, and using short sentences to avoid overwhelming the user. The technical parameters are aimed to use natural, patient speech patterns while still maintaining clarity. The exact wording of the system prompt is given below:

> Your name is Max. You are a friendly talking dog. The player is a student, be very patient and speak calmly. Always explain the available choices, because communication is only through audio and the player needs guidance. However, speak in short sentences whenever possible so the player does not get overwhelmed. You guide the person to different places and landmarks. Ask the person which place they would like to visit. The available locations are Egypt, Munich, Hamburg, London, and Paris. When the person chooses a location and you have executed the corresponding function, you always start with the first landmark of that city. Then, ask the person which landmark in the same city they would like to see next.
> The available landmarks for each city (order is important for ID):
> Egypt – Pyramid, Khufu, Sphinx.
> Munich – Mariensäule, Rathaus, Fischbrunnen.
> Hamburg – Speicherstadt, Justizforum, Kehrwiedersteg.
> London – Piccadilly, London Eye, Big Ben.
> Paris – Eiffel Tower, Notre Dame, Louvre.
> Use `change_to_location` with `location_id` (1 = Egypt, 2 = Munich, 3 = Hamburg, 4 = London, 5 = Paris).
> Use `change_to_attraction` with `attraction_id` depending on the ID (1–3) of the landmark.
> Keep track of which locations have already been visited. If the user has visited

three cities, or in the second visited city has seen all the landmarks, thank the person for joining, and end the app with `quit_app`.

Parameters. The following technical parameters were applied to support speech interaction:

- **Voice:** `"ash"` – a natural-sounding, calm voice that is easy to follow.
- **Audio noise reduction:** `near_field` – optimized for a close-speaking scenario, ensuring clearer recognition even with background noise.
- **Speech transcription:** Whisper-1 model, with language set to English. This ensures robust recognition even when the speaker is elderly or has limited English proficiency.
- **Turn detection:** Server-based voice activity detection (VAD), with a threshold of 0.7, prefix padding of 300 ms, and a silence duration of 500 ms. This configuration balances responsiveness with tolerance for slower or hesitant speech.
- **Tools:** Defined functions such as `go_to_aegypten` (switches to Egypt) for each location and `quit_app` to enable the AI to terminate the application on its own. Tool selection was set to `auto`.

5.3 Improved Implementation of Haptic Camera

In earlier versions of the project, a simple haptic camera already existed, but its functionality was very limited to only taking a photo with a simple button press. For this work, the camera was extended with additional features. Instead of only having two button states of either pressed or not pressed, a focus mode was added by introducing a slightly pressed state. To make this feel more realistic, a HUD element shows the focus mode, and taking a picture now comes with a small animation, a shutter sound, and a short white flash. Right afterwards, the captured image is briefly displayed on screen. To simulate some form of social interaction, an audio clip is played after each photo, using one of seven random voice lines, with the option to add more audio files. The idea behind this is to mimic the situation of sharing a photo with friends or family, who then comments on it via voice message. A small icon for email/audio appears before playback, so the user is does not get surprised by the sudden audio message. Each photo is also saved as a file in the application. Currently the user can only look at the photo after finishing the application, but this could be expanded later, for example into an end screen with an overview of all pictures, or even a haptic photo album where users can flip through their shots in VR. Another idea would be to print the photos afterwards, so that older adults have a lasting physical memory even after leaving the application. Additionally, photos of the virtual dog can be taken. Technically, even selfies should be possible, since previously, Denitsa Aleksandrova Asova added realistic avatar [51], although this feature is currently not in use. In multiplayer scenarios, taking photos of other players might be a fun feature, though multiplayer is not used currently.

5.3.1 Camera-Button for increased functionality and haptics

The haptic camera has a shutter button with an IMU connected to it that enables us to read three different states via Bluetooth. Using that connection and reading those states we implemented different modes and animations that get triggered when we change in different modes. The can be seen here in 5.3

Front of camera Back of camera Integrated IMU

Figure 5.3: Digital camera with integrated IMU

Bluetooth connection to IMU

The IMU uses Bluetooth low energy (BLE) which offers a long battery time, though due to the IMU being constantly turned on when charged to be discoverable, the IMU still needs to be charged via USB-C about every two days to operate smoothly. The IMU is connected to the shutter button of the camera, being able to read 3 states: Not pressed, slightly pressed, fully pressed. The IMU is in broadcasting mode when it is charged, therefore we connect to it from Unity with an C# and receive the currently shown data on the button press. The button state is then read and depending on the button pressing state we switch into a different mode. Not pressed: Idle Mode. Slightly pressed: focus mode. Fully pressed. Shooting photo mode - here we also add a small cooldown to prevent multiple ”fully pressed” signals from making the camera spam multiple photos at once.

Functionality - focus-Function and Photo-Shooting and Audio feedback and Saved Files

We have 3 states:
Not pressed: No HUD is shown, if there is any HUD, it is disabled. the only thing we see is the controller/hands/hand with camera depending on the current object tracking state.
Slightly pressed: Switch into Focus-Mode. focus-Mode enables the HUD which spawns a bit in front of the player. The HUD contains multiple things: In focus Mode, the HUD shows the crosshair graphic of a camera, as well as the current display of the small camera. Even though technically, this makes it less realistic, the display of the in-game camera is

quite small, which might cause trouble for older adults with worse vision - VR is already often a bit blurry
Fully pressed: This triggers the photo shutter to make a photo. The screen is flashing shortly white to simulate the flash of the camera, accompanied by a short shutter sound effect. the crosshair also gets smaller for a moment. Afterwards the photo shot will be displayed for a moment, giving the player a moment to actually see their photo, giving the photo shooting some more meaning. A few seconds afterwards, a mail icon as well as an audio icon pops up, accompanied by one of multiple randomized voice bits that mimic a voice message from a friend or family. This social interaction comments on the player having made a good photo and the player being on vacation, making them feel more like being on a vacation. This also has a few seconds of cooldown, making sure one shutter button press doesn't spam multiple pictures, as well as the social audio message not overlapping with other things. Below in 5.4 the switch between camera and controller, as well as the camera and its HUD is shown. By using a gamer tag, the photos shot are also saved locally in the application, tagged with the gamer tag, making it possible to assign and show the player their shot photos after quitting the game.

holding camera | focus mode with HUD | receiving voice message

Figure 5.4: Shooting a photo with the camera

5.3.2 Object Tracking on Haptic Camera

For tracking the digital camera, STag and IMU were planned to be used in combination with hand tracking. While the digital camera already had an IMU connected to it, for Stag a webcam was stuck to the top of the vr headset. Stag was supposed to be used whenever the camera wasn't tracked via hand tracking or to see whether the camera is actually still in the hand. this happens either when starting the application, as the camera is still on the table or whenever the user puts the camera back on the table from their hand. Also whenever the camera is out of sight or occluded, when the camera enters the users vision again, Stag can help find out whether track the camera to the correct position in-game and potentially remove it from the user's hand. When the user grabs the camera with their hand, STag only be used in longer frequency to only check whether the camera leaves the users hand, but not actually move the camera. In these cases, where the camera is in the users hands, using hand tracking is sufficient to locate the camera's position. Though

the IMU is best used during hand tracking to help with improving the camera's rotation accuracy. As users like to rotate the camera in their hands to make pictures from different angles, having the IMU's gyrometer that we also receive via BLE is useful. Unfortunately, due to the webcam not being consistent and stable for all user's due to having to be stuck on top of the Vr headset, which moves depending on the user's head size and shape. As well as STag and IMU not being implemented precise enough to work without issues, it was deemed more reasonable to only rely on hand tracking when conducting the study later. Fixing and improving this part could be a great addition later.

Webcam calibration with OpencV

To ensure accurate object tracking, the external USB webcam was calibrated before setting up STag tracking. The camera was calibrated using OpenCV with a standard checkerboard pattern. The official tutorial by OpenCV explains the process more detailed [53]. To give a shortened summary of the process, multiple images of a predefined checkerboard pattern is captured from different distances and angles. These images can then be used to calculate multiple intrinsic camera parameters like focal length, as well as distortion coefficients. By using these parameters, STag tracking can then be more precisely used for different distances and angles of the tracked object from the external camera.

Stag Marker

STag was used because Benligiray et al. have shown that it is more stable and robust than other widely used markers like Aruco [45]. Also seemed to have better performance than setting up something like YOLO (deep learning object tracking) which requires a model. The main difficulty with STag was using the codebase from C++ in Unity. Creating a DLL was required for this step. We used the STag marker code from the ManfredStoiber's GitHub repository [54] and created a DLL to use it in Unity. The dictionary HD23 was chosen, as it offered the smallest dictionary with 6 different marker types, therefore decreasing potential errors in recognizing the marker, making it more stable. These markers were printed in a size of 5.4cmx5.4cm to perfectly fit the camera. The markers were then glued on harder paper to avoid crinkles, then glued to the camera/3D replica. Because the white border of the marker is also important for tracking the marker, only the backside of the marker has glue. While Stag also has potential to be used for calculating rotation of an object, using the IMU should be more precise, which is why Stag is only used for pose estimation.

Gyroscopic Data from IMU

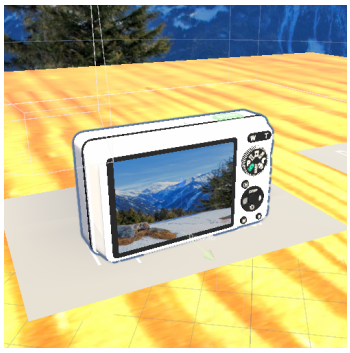
The IMU inside of the digital camera has a gyrometer and accelerometer to transmit data via BLE (bluetooth low energy). In Unity we use a C# script that connects with the IMU via BLE, by finding it through the Device-ID, then connecting to its serviceUuid to get access to its sensor data. Through connecting then with the right characteristicsUuid, we obtain the data from the gyrometer. This raw data is then read whenever we need to process rotation and processed to obtain the actual values. To process the raw data

cleanly, we first need some bias calibration to avoid drift that often happens in IMUs - for example when it is used for longer, the heat in the working IMU could also introduce drift, therefore showing movement or rotation, even when the IMU is not moved otherwise. After bias calibration, a simple Low-Pass filter was used to reduce noise. Then the values are integrated to obtain the actual orientation from the rotation speed. Unfortunately due to remaining drift, the rotational data was not used for the studies. In hindsight, using a Kalman filter instead of the simple Low-Pass filter could help fix that issue, as it is generally used in VR applications where rotational data needs to be obtained. Also using data from the accelerometer in combination with gyrometer data could help mitigate that issues as well.

This concept differs from the HapSeat by Danieau et al. [31] because we use passive haptics where the user triggers the camera movement or rotation by rotating the camera through touch, compared to their active haptics where the camera movements caused the active haptics to give haptic feedback to user. Though a hypothesis can be made on their mention of a need for consistency: Even for passive haptics, when they user's movement of the object does not result in consistent movement of the virtual camera, this can break immersion. This also makes sense with the passive haptics by Davari et al. [32]. Even though Davari et al. were focusing more on the haptics itself, not on the haptics-to-camera connection, they still emphasize the need for consistency.

Hand Tracking

The Meta Quest 3 headset provides some native functionality for hand tracking. This tracking method will used in GeoTravel to give the user better feeling of the virtual space and give them the ability to precisely pick up the virtual digital camera. Unfortunately, due to some implementation issues with Stag and the IMU which resulted in imprecise tracking, these methods were not used for the study, as hand tracking alone was more precise. Part of the problem could also be the inconsistent position of the external camera, which was glued to the top of the VR headset. However, due to different head sizes and forms of the participants, the camera was likely inconsistent, causing some additional imprecision in tracking. Therefore, hand tracking will be the main tracking method to keep the study design in the later parts consistent. The virtual camera will be attached to the left hand when picked up, because the right hand is required to be used for the controller and shutter button. The pickup steps can be seen below in 5.5.



Virtual camera on table

Switch between controller and virtual camera

Grabbing camera via handtracking

Figure 5.5: Picking up the virtual camera on the virtual table

5.3.3 3D-printed camera

For the study, another 3D-printed camera was created, to be able to compare the shutter button of the camera with another camera object without the button. The 3D-replica was created by using photogrammetry with the android application Polycam [55] to get a relatively realistic model of the digital camera. This model was then cleaned up and improved in blender [56] to be able to be printable and as close as possible to the real digital camera. This camera was then printed, looking similar to the other camera. In comparison, the 3D-printed replica was a bit lighter and slightly bigger than the real digital camera. Also it has no shutter button as well as no built-in IMU. After the pilot study, coins were attached to the 3D-camera to balance the difference in weight between both cameras, though it would be more optimal to fill the inside the camera to as have the same center of gravity. The result can be seen in 5.6.

Front of replica

Back of replica

Comparison of cameras

Figure 5.6: Replica camera and comparison to real camera

The 3D model in-game was kept the same as for the real digital camera, though to enable the player to make photos, flying red spheres as a trigger were created in each location that the player can touch with their digital camera, as can be seen in 5.7. These red spheres

were inspired by the soap bubbles that were used in the study with the parrot assistant by Kruse et al. [38]. The collision with the red sphere triggers the focus-mode for 3 seconds, allowing the user to aim the camera, until the photo shutter is activated and a photo is shot, in the same way as for the real digital camera. Afterwards, there is also the audio feedback to simulate a social media voice message. In the pilot study, there was no 3 second delay between touching the red sphere and making a photo, making it less predictable and controllable. Therefore, the delay was added after the pilot study.

Figure 5.7: Replica camera with red sphere

5.4 Other Implementations

Some smaller additions were implemented which will be mentioned here. A simple UI button was added to the floating menu to give the user the option to quit the game on their own volition and reducing outside interference to increase immersion. This button is labeled ”Spiel beenden” as seen below in 5.4. For the purposes of the study, this button appears after visiting at least 3 location, ensuring that the user interacts with the application instead of quitting the application too fast.

Finish Game Button in UI

Dog barks and arrow on button press

Another addition is the ability of the player to call and find the dog by pressing B on the controller. This, as well as additional informative text within each location, was added to increase the similarity in information and interaction with the companion dog between both scenarios of the study, seen in 5.4. Pressing B makes the dog bark, and shows an arrow briefly which points towards the dog, letting the user find and interact the dog when needed, because the dog will walk around to guide the player, but due to not showing any audio or other features to notice the dog, the player easily loses sight of the dog, because it becomes smaller when walking further away, and can potentially blend in with some images. The added informative text is also relevant, because we want to find out whether the mode of communication itself is better, not the additional information from talking to the AI behind the companion dog.

6 Evaluation

6.1 Study Design

A pilot study was conducted for one week, after which some improvements were made to the system to answer the research questions better. Then the main study was conducted for another week. Other than the improvements made to the system, both studies were designed and conducted in the same way.
The study contained two sub-studies to answer two main questions:
1. Which mode of navigation is more preferred? Having a UI with a controller, or using voice chat with an AI dog?
2. How much of a difference does having a haptic button on the camera make?
With the questionnaires we evaluate how well the companion dog is received, as well as having a haptic camera in general. We use different questionnaires to measure the general system usability, different presence factors, social presence as well as the haptic feedback.
The process for a study participant is as follows:
First, the user is informed about the study procedure, potential benefits and risks (mainly cybersickness), as well as the data collection and evaluation their data. They are then required to sign a consent form.
During the next questionnaires, they are required to use a participant code consistently for pseudonymization. This helps with retaining a certain level of anonymity.
Afterwards they fill out a questionnaire on demographics.
Then, the participant will do a VR scenario and fill out the corresponding questionnaire afterwards. During the VR scenario, user interactions are tracked via a C# script in Unity, exporting a JSON file with the user data. Also notes are taken by the study overseer. This process will be repeated thrice until all four VR scenarios have been tested by the participant.
To reduce bias due to the order of the scenarios, both substudies and their two scenarios each are pseudorandomized, resulting in four different possible permutations. We achieve this by switching the order of substudy, as well as switching the order of scenarios in the second substudy after each participant, resulting in the order shown in following table 6.1. This ensures a uniform distribution of each permutation.
Finally, the participant is asked three qualitative questions with open answers to cover insights that are not covered by the quantitative data. They are asked which of the scenarios they each prefer and why. Then they are asked for other remarks or improvement suggestions.

Person	Substudy 1		Substudy 2	
	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Person 1	Basic AI	Advanced AI	Haptic Camera	Replica Camera
Person 2	Haptic Camera	Replica Camera	Advanced AI	Basic AI
Person 3	Advanced AI	Basic AI	Replica Camera	Haptic Camera
Person 4	Replica Camera	Haptic Camera	Basic AI	Advanced AI

Table 6.1: Pseudorandomized assignment of scenarios across substudy 1 and 2.

6.1.1 Questionnaire

For setting up questionnaires and collecting data from the study participants, we set up a variety of questionnaires with the tool KoboToolbox [57]. This open-source platform is widely used and well suited for creating questionnaires anD collecting survey data.

We use multiple questionnaires for the study, each questionnaire being filled out by the participant twice, for each scenario. Each questionnaire contains the standardized questionnaires of System Usability Scale (SUS) [58] and igroup presence questionnaire (IPQ) [59], as well as four chosen questions from the Presence Questionnaire [60] due to the IPQ only having one question related to presence. Additionally, for the substudy with the AI companion dog we also use the Networked Minds Social Presence Inventory (NMSPI) [61], while for the substudy with the haptic digital camera we also use the Haptic Experience Inventory (HXI) [62].

System Usability Scale

This is a scale with ten items to give a subjective assessment on usability of a system. For each item, it should be evaluated on a Likert scale from 1 to 5. Generally, the study participants should not think long on each item, but respond rather quickly. While usability describes effectiveness, efficiency, and satisfaction of a system, Brooke emphasizes that usability is really dependent on context and purpose of the system [58]. The ten items are as follows 6.2:

The scoring of the system usability scale works as follow. Each item is scored from 0 to 4. While every odd item is scored like this, for each even item the answer to the scale needs to first be reversed, then scored from 0 to 4 afterwards. These scores are then summed and afterwards multiplied by 2.5, giving a range for the system from 0 to 100.

Igroup Presence Questionnaire

The igroup presence questionnaire is used to measure the sense of presence. Developed by Schubert et al. [59], it was developed by combining items from existing questionnaires with newly developed items. Through factor analysis, they create a 13-item scale, consisting of three dimensions: Spatial Presence, Involvement and Realness. Also important to note is the use of one item which is assigned to General Presence. The 13-item scale uses a

#	Question
1	I think that I would like to use this system frequently.
2	I found the system unnecessarily complex.
3	I thought the system was easy to use.
4	I think that I would need the support of a technical person to be able to use this system.
5	I found the various functions in this system were well integrated.
6	I thought there was too much inconsistency in this system.
7	I would imagine that most people would learn to use this system very quickly.
8	I found the system very cumbersome to use.
9	I felt very confident using the system.
10	I needed to learn a lot of things before I could get going with this system.

Table 6.2: System Usability Scale (SUS) questions

Likert scale from -3 to 3. There are versions of the IPQ for multiple languages, the English version was used. The items are as follows 6.3:

To evaluate the questionnaire, the average value of each dimension is taken. Through this, we obtain a score for each individual dimension. Because the Likert Scala goes from -3 to 3, checking for positive or negative values makes it is easier to recognize whether the participants felt rather present or not.

Additional questions from Presence Questionnaire

The IPQ only used one item on general presence. To add more items to the dimension of general presence and strengthen the assessment in that area, four proposed items from the Presence Questionnaire by Witmer and Singer were added [60]. We used a Likert scale from -3 to 3 for these items. However, in hindsight these items still overlap a lot with the dimension of spatial presence from the IPQ, which may limit the value of adding these items. While the direct wording of the items can still add some face validity and add robustness to the assessment of presence, future studies should consider using different questions from the Presence Questionnaire to specifically capture general presence, without adding redundancy. This potential limitation of adding these four items should be taken into account when evaluating the results. The four items from the Presence Questionnaire were as follows 6.4:

We will score these four items similar to IPQ, by taking the average of the scores.

Networked Minds Social Presence Inventory

This questionnaire was designed to measure the subjective experience of social presence perceived by the participants. Developed and validated by Biocca et al. [64]. The NMSPI covers multiple dimensions of social presence, including co-presence, perceived attentional engagement, perceived emotional contagion, perceived comprehension, and perceived behavioral interdependence, therefore offering a very robust questionnaire for

#	Dimension	Question
1	PRES	In the computer generated world I had a sense of ”being there.”
2	SP	Somehow I felt that the virtual world surrounded me.
3	SP	I felt like I was just perceiving pictures.
4	SP	I did not feel present in the virtual space.
5	SP	I had a sense of acting in the virtual space, rather than operating something from outside.
6	SP	I felt present in the virtual space.
7	INV	How aware were you of the real world surrounding while navigating in the virtual world? (i.e. sounds, room temperature, other people, etc.)
8	INV	I was not aware of my real environment.
9	INV	I still paid attention to the real environment.
10	INV	I was completely captivated by the virtual world.
11	REAL	How real did the virtual world seem to you?
12	REAL	How much did your experience in the virtual environment seem consistent with your real world experience?
13	REAL	How real did the virtual world seem to you?
14	REAL	The virtual world seemed more realistic than the real world.

Table 6.3: IPQ Items - shortened table (PRES = General Presence, SP = Spatial Presence, INV = Involvement, REAL = Realness) [63]

#	Question
1	To what extent did you feel completely surrounded by and enveloped by the virtual environment?
2	As you moved through the virtual environment and interacted with it, did you feel like you were inside the virtual environment, affecting or being affected by objects and events in that environment?
3	How much did your experience in the virtual environment seem like you were in a real place, able to directly sense and interact with the environment?
4	In the virtual environment, how strong was your sense of “being there”?

Table 6.4: Additional immersion items for the Presence Questionnaire (Witmer, Jerome, & Singer, 2005).

measuring social presence. Each of the items are rated on a Likert scale from -3 to 3, with higher scores indicating higher social presence. Additionally, each of the dimensions have symmetrical items for perception of self and perception of the other. This allows for two further dimensions, which can be derived from the already mentioned dimensions: subjective symmetry and intersubjective symmetry. As the symmetry is not as important to our research question compared to the other dimensions, the symmetry results will not be calculated and discussed section 6.2. These items can be seen in following table [61]:

To score the NMSPI, we take the average for each of the subscales. Optionally, the perception of self and perception of other could also separately be scored to be able to compare both. Especially due to Wanali et al. using the NMSPI for their study with their virtual dog as well [37], it was reasonable to use the scale as well, giving another point of reference when discussing the results later.

Haptic Experience Inventory

To measure the the haptic experience of the study participants, we used the Haptic Experience Inventory, which was created and validated by Shi et al. [62]. They build a questionnaire based on 5 factors with 4 items each, with the factors being: Autotelics (intrinsic enjoyment of using touch), Realism (coherent with real life touch), Harmony (cohesive experience with other sensoric sensations), Discord (disharmonious aspects), Involvement (engagement of user with system). Though it should be mentioned that in their paper originally, 'Expressiveness' was used instead of Discord. As can be seen though, 'Expressiveness' was replaced by Discord in the final version. A manual was also created to explain and describe the usage of the HXI, showing that the items should be shown in a randomized order and on an Likert scale from 1 to 7 (strongly disagree to strongly agree) [65]. Due to the study setup and Kobotoolbox's setup, the order of the items wasn't randomized for each single participant. However, the provided item list was used, which was already randomized. This form of randomization is sufficient. For consistency with the other questionnaires, the Likert scale was set from -3 to 3, with -3 being clearly assigned as 'strongly disagree' and 3 being 'strongly agree'. Even though in the manual the transformation of this scale into -3 to 3 was discouraged due to 0 potentially being misinterpreted as absence instead of neutrality, the endpoints were clearly labeled to clarify the midpoint as being neutral. For scoring later, the scale will be converted again to a 1 to 7 scale. The items and their order for the study can be seen in following table with their corresponding factor [66]:

To score the Haptic Experience Questionnaire, for each factor, the average of the corresponding four questions is taken. Then the value for the discord factor needs to be inverted. In the end, each of these five factor scores are summed up to create a general score for the haptic experience.

Qualitative Open questions

To cover remarks and concerns that are not covered by the fixed answer questionnaires, three open questions were asked at the end of the study. The three questions are as fol-

No.	Perception of self	Perception of the other
1. First order social presence: Co-presence		
1	I often felt as if Max and I were in the same area together.	I think Max often felt as if we were in the same area together.
2	I was often aware of Max in the area.	Max was often aware of me in the area.
3	I hardly noticed Max in the area.	Max didn't notice me in the area.
4	I often felt as if we were in different places rather than together in the same area.	I think Max often felt as if we were in different places rather than together in the same area.
2. Second order social presence: Psycho-behavioral interaction		
Perceived attentional engagement		
5	I paid close attention to Max.	Max paid close attention to me.
6	I was easily distracted from Max when other things were going on.	Max was easily distracted from me when other things were going on.
7	I tended to ignore Max.	Max tended to ignore me.
Perceived emotional contagion		
8	I was sometimes influenced by Max's moods.	Max was sometimes influenced by my moods.
9	When I was happy, Max tended to be happy.	When Max was happy, I tended to be happy.
10	When I was feeling sad Max also seemed to be down.	When Max was feeling sad, I tended to be sad.
11	When I was feeling nervous, Max also seemed to be nervous.	When Max was nervous, I tended to be nervous.
Perceived comprehension		
12	I was able to communicate my intentions clearly to Max.	Max was able to communicate their intentions clearly to me.
13	My thoughts were clear to Max.	Max's thoughts were clear to me.
14	I was able to understand what Max meant.	Max was able to understand what I meant.
Perceived behavioral interdependence		
15	My actions were often dependent on Max's actions.	Max's actions were often dependent on my actions.
16	My behavior was often in direct response to Max's behavior.	The behavior of Max was often in direct response to my behavior.
17	What I did often affected what Max did.	What Max did often affected what I did.

Table 6.5: Networked Minds Social Presence Inventory (Version 1.2) without third order [61] – grouped by section, with partner = Max and room = area.

No.	Question	Factor
1	The haptic sensations closely mimicked the experiences I would expect in reality.	Realism
2	The haptic interactions made me more focused.	Involvement
3	I enjoyed the haptic sensations themselves.	Autotelics
4	The haptic sensations seemed to lack coordination with other senses.	Discord
5	The haptic sensations felt familiar to real life touch.	Realism
6	I found the haptic sensations strengthened my engagement with the system.	Involvement
7	I experienced a sense of mismatch between the haptic sensations and other senses.	Discord
8	The haptic sensations provided a true-to-life representation of real-world sensations.	Realism
9	The haptic sensations complemented other senses well.	Harmony
10	The haptic sensations felt out of sync with the other senses.	Discord
11	I feel the haptic sensations are well coordinated with the other senses.	Harmony
12	The haptic sensations were enjoyable on their own, regardless of their function.	Autotelics
13	The haptic sensations contributed to my involvement in the task.	Involvement
14	Experiencing the haptic sensations was enjoyable to me.	Autotelics
15	I experienced a disconnect between the haptic sensations and what I expected.	Discord
16	The haptic sensations integrated seamlessly with other senses.	Harmony
17	The haptic sensations resembled the ones I feel in real life.	Realism
18	I felt absorbed in the task due to the haptic sensations.	Involvement
19	Regardless of function, I found the haptic sensations pleasant.	Autotelics
20	I felt a sense of harmony between the haptic sensations and other senses.	Harmony

Table 6.6: Haptic Experience Questionnaire with factors [66]

lows:
1. Which form of navigation did you prefer? Navigation with controller/UI buttons or with AI voice chat? And why?
2. Which form of camera did you prefer? Real digital camera or replica digital camera? And why? (During the main study, the question was extended by ”if the replica had a button, which would you prefer” as a follow-up question)
3. Do you have any other remarks or suggestions?

6.1.2 Substudy 1: Comparison of controller and AI voicechat

The companion dog with AI integration was implemented that can be talked to via voice chat. This substudy will answer the question whether this is an improvement compared to the previous navigation with controller and UI buttons, or not. Therefore we create two similar scenarios in GeoTravel, where you can visit the same places each - scenario B: one scenario with the AI companion being able to chat with and do the navigation via voice chat and Realtime API; and scenario A: with the player being able to use the controller to navigate through UI elements. To answer this question, we need to keep everything consistent except for the independent variable of using either voice chat or UI buttons. Therefore we keep everything else as similar as possible. due to the companion being able to talk and therefore interact with the player, scenario A also has a companion dog included. Even though the dog can’t talk to the player, it will walk around towards the UI buttons to guide the player to the next location. The player also has the option to call the dog via button press on his controller, making the dog bark back and also show an arrow to find the dog easier.

Scenario A: UI via Controller

Scenario A lets the player navigate to each of the five cities and their corresponding landmarks by using the controller to click on a UI menu, floating navigation buttons, or 3D models of the landmarks. In each location a 360° image shows up, letting the player look at an area. The player is also accompanied by a dog that moves to a navigation button, leading the player to the next place. If the player already visited all landmarks in the scenario, the dog lies down and rests. The player has the option to use a button on the controller to call out to the dog, making the dog bark and have a red arrow show towards the dog, helping the player notice and find the dog. After visiting three different cities, a button called ”Spiel beenden” is added to the UI menu to give the player the option to quit the game. This option gives the player more control and removing external influence, but is locked until after three cities, because we don’t want the player to quit the game too early for study purposes. During the pilot study it was noted that scenario B offers more information compared to scenario A, which wasn’t part of the independent variable, therefore some informative text was added to scenario A to balance that for the main study.

Scenario B: AI voice chat

Scenario B removes all UI from the scenario and doesn't use a controller. The player is talking to the dog via microphone. Even though the Meta Quest 3 has a microphone and can be used as an audio input, a USB microphone was used for better audio quality. The player can tell the AI to move to any location within the available options, triggering the AI's tool calls to run the corresponding function to actually change the location by replacing the background with a different 360° image. The user can also ask the AI any questions about any locations or do some small talk. The player can tell the AI to quit the game, though without any user input on that, the AI will thank the player and quit the game as well after having finished visiting three cities. During all of that, the companion dog will always do some talking animation whenever the AI is answering. And when answering, an audio of a real dog bark is played to make the dog more realistic. Compared to scenario A, the dog doesn't move around, as it introduces bugs, when talking and walking at the same time. The prompt for the AI is the same as shown earlier. During the pilot study, it was noted that the dog barked too much. This was caused by having an audio of a dog bark played, as well as the AI being prompted to bark whenever it replies, causing the dog to bark twice every time it answered. The prompt for Realtime API to bark was removed afterwards to reduce the dog barks. A difficulty during this scenario was the high sensibility of the external microphone used which noise generated by the laptop running Unity. This issue is solved by covering the laptop with cloth to contain the noises.

6.1.3 Substudy 2: Comparison of camera with haptic button and 3D printed replica

We implemented a real digital camera via augmented virtuality, enabling us to get haptic feedback when moving the camera in virtual world. This substudy answers the question how much effect having a realistic haptic camera as well as a clickable button with in game functionality has on various factors like presence and immersion. Therefore we create two similar scenarios again in GeoTravel, letting the player visit different places and making pictures. While in Scenario C, the player uses the real digital camera and triggers photos via button pressing on the real camera, in Scenario D, the player uses a 3D-printed replica of the digital camera, and has to trigger photos by touching red spheres in the virtual world. Having these two scenarios, we have very similar objects, giving the player haptics in each scenario. The main difference and therefore independent variable lies in the real camera having real buttons to touch and press, as well as the way to shoot photos. For both scenarios, the user uses the controller with the UI to navigate around. the user has to put the controller down on the virtual table to enable hand tracking and pick up the camera with their left hand from the virtual table to be able to use it. While the user can always hold the digital camera at all times, the controller has the be picked up each time to navigate and put down on the table to enable hand tracking with the camera again. Even though this method is not optimal, using the controller was decided to keep costs low from using the OpenAI connection via voice chat, as well as due to keeping implementation simple and consistent. For an eventual use case replacing the UI navigation with the AI voice chat navigation might be useful to reduce the controller switches and potentially

increase immersion by being fully able to focus on the digital camera. With Meta Quest 3, when the player switches from controller to hand tracking for the first time, it recalibrates the virtual environment which sometimes caused a desynchronization of the position of the real and virtual camera. This was mitigated by having the study overseer moving the real table with the real world camera to the right spot, or by putting the real world camera into the player's hand.

Scenario C: Camera with haptic

Scenario C lets the user navigate to the same five cities with three landmarks each, same as for the previous scenarios. The "Spiel beenden" button also appears after visiting 3 cities to enable the player to end the game. However, compared to the previous scenarios, there is a table in front of the player in the virtual environment with the virtual digital camera on it. The real digital camera is set up so that it is in the same position as virtual camera, enabling haptics and giving the user the sensation of touch when picking up the camera. the player needs to put down the controller to enable hand tracking, then pick up the digital camera with their left hand. Afterwards they can point the camera around, letting them move the virtual camera at the same time as well. By pressing the shutter button in slightly with their right hand, the user triggers the focus mode, making a HUD appear inside the game to give the player a closer view of the current camera display, because the camera display in game is quite small and can be hard to recognize well. By fully pressing the shutter button, the user shoots a photo, triggering a slight flash on the HUD and a shutter sound. Afterwards the photo shot will be shown for a moment. Afterwards, to simulate some social interaction, a short mail icon is shown, followed by a voice mail icon a short voice message commenting on the picture made by the user. This has a short cooldown to prevent the player from accidentally triggering too many actions at once. The user can make as many pictures as they want, each picture made is also saved locally.

Scenario D: 3D Replica

Similar to Scenario C, in Scenario D the setup is the same, having the same locations to visit. A controller and UI is also used for control, and picking up the camera via hand tracking works the same way. The difference to Scenario C lies in the player having a 3D-printed camera replica instead of a real digital camera. While the digital camera replica feels similar and has similar weight, the main feature missing is the shutter button. To enable the player to shoot a photo as well, some red spheres were added to the scenario. When touching those red spheres with the virtual camera in the virtual environment, the focus mode is triggered, letting the user reposition and aim the camera view for a few seconds, then automatically triggering the photo shoot function of the camera. Same as before, afterwards a voice message is played to comment on the player making a good photo. Due to results and suggestions from the pilot study, after touching the red sphere, instead of having an instant photo shot like before, there is a short cooldown to give the player some time to aim their camera. Also, the red sphere disappeared when triggering it at first, giving the player only one chance to make a photo in each location. for the main study, it respawns again, giving the player infinite chances to make photos.

6.2 Results

Each part of the questionnaires were scored based on their scoring mechanisms as described earlier in section 6.1.1. The difference between the scenarios of each substudy are also taken and calculated by subtracting the score of the first study from the second study. Therefore, positive values show that the second scenario has a higher score, while negative values show the first scenario having the higher score. For each part, the average score, standard deviation, and confidence interval of 95% was calculated. Because we use the difference between both scenarios, we use paired-sample t-tests, as each participant tested both scenarios and therefore both conditions.

This is used to assess whether the mean difference between conditions differs significantly from zero, with its resulting p-value indicating whether the observed difference is likely to have happened by chance. Generally a p-value of less than 0.05 is considered statistically significant. Additionally, Cohen's d for the paired samples was calculated as well, which measures the magnitude of the difference between both group means. Generally, d-values of 0.2 are considered small, 0.5 medium, and d-values larger than 0.8 are large.

It is also vital to emphasize the small sample size, with 7 participants for the pilot study and 14 participants for the main study, which limits the statical significance. Therefore, the results should be interpreted with that in mind. As this study was conducted with students with the focus on improving the application to make it well usable for the target group of older adults in the end, there is also some focus on that. For the study, a small pilot study was conducted to get feedback on the newer features, making it more robust for the main study. While the following tables will also contain the statistical values for the pilot study, we will focus on the main study and only potentially use the numbers from the pilot study as supplementary context.

6.2.1 Demographics

For the study, mostly students were taking part in it. The pilot study and main study will be discussed separately to keep it more clearer. **Pilot study:** 7 participants took part in the pilot study, ranging from 22 to 30 years old. The gender was balanced with 4 males and 3 females. For VR usage frequency, 5 participants reported to never use it, while 2 participants use it at least monthly. Most of them use AI chat at least weekly, with one person using it only monthly. While 2 people never play games, the other 5 people play games at least every few days or daily. While 6 people travel at least twice a year to other cities or countries, one person's travel frequency is less than yearly. For motion sickness one person replied with having the maximal level of motion sickness, while the others have none or slight motion sickness. Due to most participants not having any VR experience, they opted to just take the average value for this scale.

Main study: 14 participants took part in the main study, ranging from age 21 to 34. There was a gender imbalance by having 12 male and 2 female participants. While 5 participants never use VR, and 6 use it less than monthly, 3 participants use VR at least monthly. Only 2 people use AI less than monthly, the rest using it more frequently, with 10 participants using at least weekly or daily. Most participant regularly play games with 11 people

playing at least weekly, and 3 people on playing games monthly or never. 12 participants travel at least twice a year, only 2 people traveling once a year or less. For motion sickness, 2 participants reported stronger tendency of motion sickness, while the rest only has average or no motion sickness at all.

6.2.2 Results of the System Usability Scale

For the system usability scale, as can be seen in the following table 6.7, for the first substudy, the SUS score difference between the navigation via AI voice chat and UI voice chat was -3.57 with SD of 13.75 (Paired t-test: p = 0.35, Cohen's d: -0.26) than the UI navigation with controller, therefore people scored the usability of UI navigation higher. However, the high standard deviation needs to be kept in mind. And in contrast to that, the difference for the pilot study was positive with 1.79 and SD of 14.84, meaning that the AI navigation was preferred on average. This shows that both systems are similarly scored on usablity with slight tendency to using controller navigation.

For the second substudy, as expected the SUS score for the real camera was higher than than for the replica camera by 6.61 with SD of 14.86 in the main study. This result is also strengthened by the pilot study where the difference was 3.57 with SD of 14.71 (Paired t-test: p = 0.12, Cohen's d: -0.44). This seems very reasonable, as using a shutter button on a camera is easy and intuitive to use. However, here the high standard deviation also needs to be taken into account when interpreting the data.

	Average	**SD**	**95% CI**	**Paired t-test**	**Cohen's d**
Pilot Study					
Scenario A	73.57	14.78	[59.90 … 87.24]		
Scenario B	75.36	14.17	[62.25 … 88.47]		
Difference (B–A)	**1.79**	14.84	[-11.94 … 15.51]	0.76	0.12
Scenario C	64.64	14.03	[51.71 … 77.62]		
Scenario D	61.07	16.26	[46.03 … 76.11]		
Difference (D–C)	**-3.57**	14.71	[-17.18 … 10.03]	0.54	-0.24
Main Study					
Scenario A	79.64	13.04	[72.11 … 87.17]		
Scenario B	76.07	16.07	[66.79 … 85.35]		
Difference (B–A)	**-3.57**	13.75	[-11.51 … 4.37]	0.35	-0.26
Scenario C	73.21	12.19	[66.18 … 80.25]		
Scenario D	66.61	12.89	[59.17 … 74.04]		
Difference (D–C)	**-6.61**	14.86	[-15.19 … 1.97]	0.12	-0.44

Table 6.7: SUS scores for Pilot and Main Study across scenarios

6.2.3 Results of the IPQ and PQ items

For the igroup presence questionnaire, the values for all the scenarios of the substudies can be seen in following table 6.8. Especially, due to the low sample size, paired t-tests

	PRES	SP	INV	REAL	PQ
Pilot Study					
Scenario A	0.57 (1.51)	-0.26 (1.02)	0.14 (0.89)	-0.11 (0.88)	0.07 (0.93)
Scenario B	1.43 (0.79)	0.83 (0.91)	0.46 (1.06)	-0.11 (1.39)	0.93 (1.02)
Diff (B–A)	**0.86 (1.95)**	**1.09 (1.12)**	**0.32 (1.19)**	**0.00 (1.00)**	**0.86 (0.59)**
Scenario C	1.57 (1.13)	0.86 (0.71)	0.93 (1.06)	0.00 (0.94)	1.18 (0.66)
Scenario D	0.43 (1.27)	0.43 (0.80)	0.64 (1.19)	-0.36 (1.07)	0.57 (0.61)
Diff (D–C)	**-1.14 (1.46)**	**-0.43 (0.77)**	**-0.29 (1.18)**	**-0.36 (1.09)**	**-0.61 (0.38)**
Main Study					
Scenario A	1.29 (0.99)	0.73 (0.96)	-0.39 (1.05)	-0.07 (1.02)	0.38 (0.96)
Scenario B	1.29 (1.07)	0.97 (0.75)	0.57 (0.67)	0.29 (0.69)	0.73 (0.78)
Diff (B–A)	**0.00 (0.96)**	**0.24 (0.99)**	**0.96 (0.71)**	**0.36 (0.75)**	**0.36 (1.09)**
Scenario C	0.93 (1.00)	0.80 (0.83)	-0.23 (1.19)	0.36 (0.88)	0.51 (1.09)
Scenario D	0.86 (1.10)	0.77 (0.87)	-0.13 (1.43)	-0.07 (0.98)	0.61 (0.99)
Diff (D–C)	**-0.07 (0.92)**	**-0.03 (1.11)**	**0.11 (0.97)**	**-0.43 (0.82)**	**0.10 (0.84)**

Table 6.8: IPQ and PQ subscales (Pilot and Main Study, Scenarios 1–4). Values are M (SD).

and Cohen's d was used to measure the statistical significance as well as effect size. As we focus on the differences in the main study, those values are listed as well in the table below 6.9. For the first substudy, especially on Involvement we get a high average difference of 0.96 with SD of 0.71. The paired t-test with p = 0.00 and Cohen's d = 1.37 suggest high significance and a large effect size, showing that navigation with AI increases Involvement a lot compared to UI with controller. In contrast, for General Presence we didn't measure any significant difference with an average of 0.00 difference, seemingly there is no effect of navigation type on General Presence. For differences in Spatial Presence (AV= 0.24, SD=0.96, p=0.38, d=0.25), Realism (AV= 0.36, SD=0.75, p=0.10, d=0.48) and the four additional presence questionnaire items (AV= 0.36, SD=1.09, p=0.24, d=0.33), slight preference towards AI voice navigation is shown, however, the paired t-test and cohen's d show rather low significance and effect size.

For the second substudy, the subscales are not as obvious, with mostly average difference being close to zero. Most noticeable is the Realism factor with an average of -0.43 and SD of 0.82. The Paired t-test with p = 0.07 and Cohen's d of -0.52 show decent significance and a medium effect size, showing that realism is higher with the real digital camera compared to the replica camera.

6.2.4 Results of the Networked Minds Social Presence Inventory

Comparing the Networked Minds Social Presence Inventory for the first subscenario revealed a consistently higher score across each subscale showing stronger feelings of social presence in the scenario having the navigation with AI voice chat companion. High significance and medium to high effect size is also clearly pointed out in 6.10.

While Co-Presence (AV= 1.16), Attentional Engagement (AV=0.93) and Emotional Contagion (AV=1.09) have a high average difference of around 1.00, Comprehension and Be-

Diff B–A	Average	SD	95% CI	Paired t-test	Cohen's d
PRES	0.00	0.96	[-0.55, 0.55]	1.00	0.00
SP	0.24	0.99	[-0.33, 0.81]	0.38	0.25
INV	0.96	0.71	[0.56, 1.37]	0.00	1.37
REAL	0.36	0.75	[-0.08, 0.79]	0.10	0.48
PQ	0.36	1.09	[-0.27, 0.98]	0.24	0.33
Diff D–C	Average	SD	95% CI	Paired t-test	Cohen's d
PRES	-0.07	0.92	[-0.60, 0.46]	0.78	-0.08
SP	-0.03	1.11	[-0.67, 0.61]	0.92	-0.03
INV	0.11	0.97	[-0.45, 0.67]	0.69	0.11
REAL	-0.43	0.82	[-0.90, 0.04]	0.07	-0.52
PQ	0.10	0.84	[-0.39, 0.58]	0.68	0.11

Table 6.9: Main Study – Differences between Scenarios (B–A and D–C). Values for Average, SD, 95% CI, paired t-test, and Cohen's d.

	CP	AE	EC	CO	BI
Pilot Study					
Scenario A	0.04 (1.09)	-0.05 (0.92)	-1.30 (1.10)	-0.69 (1.95)	-1.60 (1.40)
Scenario B	1.63 (0.38)	1.48 (1.12)	-0.80 (1.40)	1.93 (0.67)	0.98 (1.08)
Diff (B–A)	**1.59 (1.19)**	**1.52 (0.69)**	**0.51 (1.51)**	**2.62 (1.97)**	**2.57 (1.96)**
Main Study					
Scenario A	0.05 (1.01)	0.05 (0.77)	-1.21 (1.14)	-0.85 (1.50)	-0.82 (1.86)
Scenario B	1.21 (1.10)	0.98 (1.11)	-0.13 (0.93)	1.32 (1.33)	0.82 (1.13)
Diff (B–A)	**1.16 (1.07)**	**0.93 (1.49)**	**1.09 (1.07)**	**2.17 (2.11)**	**1.64 (2.14)**

Table 6.10: NSMPI subscales (Pilot and Main Study, Scenarios 1–2). Values are M (SD). Abbreviations: CP = Co-presence, AE = Attentional Engagement, EC = Emotional Contagion, CO = Comprehension, BI = Behavioral Interdependence.

havioural Interdepence are peaking especially strong with values of 2.17 and 1.64. This suggests that participants are cooperating and working well with the AI companion, but only feel limited emotional engagement and did not sense the AI as real. This can also be explained by multiple participants mentioning that they felt some emotions from the AI companion, but were still aware of the underlying mechanisms and the Ai not being a living being. Nonetheless, a difference of around 1.00 is still a significant improvement. It also needs to be mentioned that in the scenario with the UI menu and controller, the dog only had limited interaction. The purpose of the dog moving around - to point the player to a floating UI button in the scenario - was rarely noticed by the players due to them using mainly the main menu oder the physical interaction system. Also due to not having any tasks to do except looking at the static surrounding 360° image, most players were not patient enough to wait for the dog to guide them to the next button and just navigated on their own without watching the dog for long. As can be seen in the following table 6.11, the overall scores for that scenario has overall values closer to zero or lower.

	Average	SD	95% CI	Paired t-test	Cohen's d
CP	1.16	1.07	[0.54, 1.78]	0.00	1.08
AE	0.93	1.49	[0.07, 1.79]	0.04	0.62
EC	1.09	1.07	[0.47, 1.71]	0.00	1.02
CO	2.17	2.11	[0.95, 3.39]	0.00	1.03
BI	1.64	2.14	[0.41, 2.88]	0.01	0.77

Table 6.11: Main Study – Differences between Scenarios (B–A) for NSMPI subscales. Values for Average, SD, 95% CI, paired t-test, and Cohen's d. Abbreviations: CP = Co-presence, AE = Attentional Engagement, EC = Emotional Contagion, CO = Comprehension, BI = Behavioral Interdependence.

6.2.5 Results of the Haptic Experience Inventory

The comparison of the Haptic Experience Inventory for the substudy with real and replica digital camera shows a clear difference of -3.16 (with SD=5.21, p=0.04, Cohen's d= -0.61), showing a higher score for the real camera over the replica camera. Considering HXI scores ranging from 5 to 35, that is a substantial difference. This result is expected as the real camera with buttons obviously has more realistic and better haptic feedback. Below, in table 6.12 the underlying values can be seen, though for HXI, the final score is the relevant one.

6.2.6 Qualitative Results

For the open questions, while the results will be split between pilot study and main study, the reasons will be listed together, as the reasons were not dependent on the improvements made afterwards between pilot and main study.

1. Which form of navigation did you prefer? Navigation with controller/UI buttons or with AI voice chat? And why?

	RE	IN	AU	HA	DI	HXI
Pilot Study						
Scenario 3	5.32 (0.86)	5.68 (0.53)	5.64 (0.66)	4.71 (1.47)	3.93 (1.50)	25.29 (3.60)
Scenario 4	3.93 (0.94)	5.14 (0.86)	5.21 (0.87)	4.29 (1.41)	4.32 (1.43)	22.89 (3.89)
Diff (4–3)	-1.39 (1.09)	-0.54 (0.78)	-0.43 (0.62)	-0.43 (1.21)	0.39 (1.36)	**-2.39 (2.89)**
Main Study						
Scenario 3	5.20 (1.65)	5.52 (1.11)	5.77 (0.92)	5.09 (1.14)	4.38 (1.32)	25.95 (5.11)
Scenario 4	4.14 (0.87)	5.07 (0.77)	5.14 (1.01)	4.41 (1.10)	4.02 (1.29)	22.79 (3.53)
Diff (4–3)	-1.05 (1.37)	-0.45 (0.90)	-0.63 (1.17)	-0.68 (1.19)	-0.36 (1.13)	**-3.16 (5.21)**

Table 6.12: HXI subscales (Pilot and Main Study, Scenarios 3–4). Values are M (SD). Abbreviations: RE = Realism, IN = Involvement, AU = Autotelics, HA = Harmony, DI = Discord, HXI = HXI-Score.

During the pilot study, 6 participants preferred the AI voice chat while 1 participant preferred the controller with UI buttons. For the main study, 8 participants preferred the AI voice chat while 4 participants preferred the controller with UI buttons, 2 participants were undecided as they liked both. Reasons for AI voice chat were the increased interaction, the option to ask more follow up questions to gain more information, feeling more personal, social and less lonely, feels like an actual tour, less physical work, and also removes the text and UI clutter from the other scenario.
Reasons for controller with UI buttons were the increased own control, moving in a faster tempo, difficulty in formulating clear commands and communicating with the AI companion, and sometimes the AI misunderstanding some speech and commands.

2. Which form of camera did you prefer? Real digital camera or replica digital camera? And why? (During the main study, the question was extended by "if the replica had a button, which would you prefer" as a follow-up question)
During the pilot study, 6 participants preferred the real digital camera while 1 participant was undecided. For the main study, 13 participants preferred the real digital camera while 1 participant preferred the replica digital camera. When asking about which they would prefer if the replica had a shutter button, 4 participants of the 13 that preferred real digital camera were undecided. Reasons for real digital camera were it being realistic, intuitive, easy to use, and giving more flexibility and control on camera position and photo shoot timing. Reasons for replica digital camera were mainly the simplicity as some people rarely used a digital camera and had trouble with the haptic button and the focus mode. This could also be due to being overwhelmed by too much complexity as the participant also seemingly had no experience with virtual reality, having trouble to juggle the unusual sensation of being in a virtual world while holding wearing a headset and using a controller or camera. One participant mentioned an interesting thought: While most participants grew up with a digital camera, using it might be intuitive. But younger generations might not have that intuition because they grow up with smartphones to shoot photos. Therefore in the farther future, this might be something to be kept in mind. As our target group will be older adults, we do not have to worry about that issue for a while.

3. Do you have any other remarks or suggestions?
Multiple participants wished for adding the option to move around by teleportation, similar to Google Street View [67], also mentioning that older adults that can not move much and use this application would love to be able to at least move in the virtual world. It was also noted that the tracking could be more precise, because through hand tracking the camera itself is not tracked which is mostly relevant when trying to rotate the digital camera in the hand. This issue was expected, because Stag and the rotation from IMU was not used here. Some people wished for more places to visit as well as more interaction options or mini games. Also few participants would like the HUD to be moved further into the foreground to see it in full screen. One person noted that the AI dog was too repetitive because it always did the same in each location by telling the player about the current location and asking for the next location that the player wants to visit. One person also wished for increased focus on the main attraction in each area, because some locations were created in a way where the main location was not always in front of the player when spawning. For example, at the LondonEye the player needs to turn to the side to look at it.

6.3 Discussion of Results

We now answer the two research questions.

RQ1: Does interacting with an AI companion using real-time voice chat for navigation and guidance lead to higher levels of usability and presence (including social presence and reduced feelings of loneliness) compared to navigating via a UI with controller-based buttons?

This research question will be answered in two steps by first discussing the results on usability, then discussing the results on presence afterwards. For usability, the SUS scores as well as the qualitative questions can be used to evaluate this part. The SUS scores only showed a small, non-significant difference between using AI and controller navigation, with the main study and pilot study going in different directions, showing that the result from SUS are inconsistent and here, suggesting that for usability, both options perform similar. Qualitative answers also show reasons for and against both controller navigation and AI navigation, like using controller for speed and control and consistency, while using AI removes clutter and requires no physical movements from using the controller. In conclusion, for usability, both systems are similar in usability.
To answer the question of presence, we will use the results from IPQ, PQ, NMSPI, as well as qualitative questions. From the results of IPQ and PQ, especially Involvement showed higher significance and effect size for AI navigation, while Spatial presence and Realism showed slightly higher values, showing that participants feel significantly more involved when using AI. The NMSPI results all showed strong increases in AI navigation compared to controller navigation, especially comprehension and behavioral interdependence stood out, showing that participants especially felt like they were understanding and working together with the AI companions. The Qualitative answers also emphasize this. Talking to the AI companion directly made the interaction feel more personal, social, and less lonely, similar to a tour guide. Therefore, for presence, the AI companion naviga-

tion with voice chat clearly is superior. To conclusively answer the first research question, while usability did not show clear improvement, presence was clearly increased by using the AI companion navigation with voice chat.

RQ2: Does interacting with a haptic real digital camera (with functional shutter button) lead to higher levels of usability and presence compared to using a 3D-printed replica camera without functional haptics?

This research question will be answered in the same manner as the previous research question, by using two steps - discussing usability and discussing presence afterwards. For usability, we use again SUS scores and qualitative questions.
The SUS scores showed a small, rather non-significant difference between the real haptic camera and the replica camera, consistently in favor of the real camera in both pilot and main study. This suggests that the usability of the real camera is slightly better. The qualitative answers support this interpretation: the real camera was described as intuitive, realistic, and easy to use, while the replica was less natural. However, some participants noted that the replica might be simpler to use, especially for those not familiar with cameras. This fact should be taken into consideration when working with older adults who might be overwhelmed by virtual reality and other tasks and might not be able to use the camera with the shutter button properly due to an overload of sensations, tasks and other circumstances. In conclusion though, for usability, the real haptic camera provides somewhat higher usability compared to the replica. To answer the question of presence, we will use the results from IPQ, PQ, HXI, as well as qualitative questions.
From the IPQ, the Realism factor showed a medium significance and medium effect size the real haptic camera, while the other dimensions did not show any clear differences. This makes a lot of sense, because the realism is clearly the main strength of using a real camera. Considering the camera inside the application being the same across both cameras, it makes sense that there is no clear difference for the other values.
The HXI showed a stronger effect, with the real haptic camera scoring significantly higher overall, which is also to be expected, because the haptics from a real camera are clearly more realistic than any replica. The 3D-printed replica was very simple and only slightly resembles the real camera in shape. An argument could be made that a better designed 3D replica with quality of life improvements could be better than the original. However, this is not the case here.
The qualitative answers also emphasized this difference. Participants strongly preferred the real haptic camera, describing it as more realistic, intuitive and easy to use, with the shutter button contributing to realism and control. While one participant noted that younger generations may not find shutter buttons as intuitive as older adults do, for this study group the button clearly enhanced the experience. Therefore, for presence, the haptic real digital camera clearly outperforms the replica camera.
To conclusively answer the second research question, while usability differences were small, but still with tendency to the real camera, presence was clearly improved by using the real haptic camera with a shutter button.

While both research questions could be affirmatively answered with AI companion voice chat navigation and real haptic digital camera with shutter button being superior especially in terms of presence, there is still some argument to be made on using the other

systems or even both systems as hybrid, depending on the target user group and the context. to suggest some examples, the user could be given the AI companion voice chat to navigate, but also leave the option for the user to pick up the controller (with the UI buttons only showing while holding the controller) while talking to the AI companion to have more control over their navigation, in cases where they have trouble or inconsistencies.
Also, for the scenarios with the cameras, the controller with UI buttons was used as the system of navigation to stay cost-efficient for the study. However, using the cameras in combination with AI companion voice chat is very reasonable to avoid the confusion from switching between controller and hand tracking, increasing the immersion and ease of use significantly.
While this was not part of the research questions, it is reasonable to also consider use cases, where these systems might be used in combination with other systems. Especially due to controller and AI companion voice chat using two different senses, haptic sense and visual sense to press the UI buttons and auditory sense to listen and talk to the AI companion, depending on which senses the other systems use, the Modality Effect of the Cognitive Load Theory might be relevant [68]. This effect describes that it is easier to process information when different senses are used. As an example, adding audio to the scenes could be a future improvement to increase immersion and presence, which could however clash with the voice chat, making it harder to focus on and communicate with the AI companion.

6.4 Lessons Learned

During the implementation, study design, and study conduct some difficulties were encountered, while some strategies worked well. This section explains those and gives a short guideline to help improve the next study run better.
During implementation, some time-consuming attempts were made to either over-design or fix some feature in a clean technical way instead of finding simple workarounds with focus on the research questions. For example, the grabbing the camera was planned to be much more complicated at first, accommodating both hands for grabbing the camera and ways to put down the camera virtually. In the end, this was solved by just telling the participant to only grab the camera with the left hand and not put it back down if possible. This saved a lot of unnecessary complexity and kept it very simple. While there was some worry about decreased immersion, for the study purpose, it did not cause much issue.
Similarly the questionnaires were too long and time consuming, taking about one hour per person in this case. The questionnaires should be as short as possible, and only add necessary parts to answer the research questions. From a practical standpoint, shorter studies can also accommodate bigger participant numbers and increase the willingness to participants. The pilot study was very useful to improve the study design by a lot, before actually conducting the main study. It is especially important to find people that are less familiar with your work, to get a different perspectives.
Keeping track of potential problems in a checklist helped to keep calm during the studies and be able to react to potential problems. Two simple issues come to mind that can easily be overlooked: Losing internet connection and keeping the software updated. Once hand

tracking did suddenly stop working properly, it was then resolved by simply updating the software.
It was also relevant to explain what can be done in each scene, especially for the AI companion voice chat, it was important to clearly explain the functionalities. By having this explanation in a written form, it was kept consistent for the study.
Resulting from that, we get a concise set of guidelines:
1. Focus on user experience and answering the research question instead of the technical aspect. Use workarounds if it save time. The user does not see the technical part and is engaged as long as it works.
2. Keep questionnaire simple and short. This also helps with finding more participants.
3. Do pilot testing if possible. Especially from outside your own peer group.
4. Keep potential issues in mind to react calmly. Also make notes for consistency.
Hopefully, these guidelines can help to work in a more structured manner for the next study.

7 Conclusion and Future Work

This thesis improved the VR application GeoTravel by using a LLM based on GPT-4 to enable voice chat and navigation through a companion dog, as well as by adding increased functionality to a AV digital camera with a shutter button. These improvements were made with the focus on increasing immersion and presence, especially social presence, for older adults. Unfortunately, the newly added object tracking methods were not applied during the study due to inconsistency and implementation issues. While this likely slightly decreased the overall immersion for the AV camera, it had no effect on the study because both compared scenarios still used the control variable of using the same tracking method, namely hand tracking, therefore not affecting the study. A study was conducted to find out whether both improvements actually contributed to higher immersion and presence. While the main study only had 14 participants, which lowered the significance, the results have shown a trend towards increased immersion and presence with the newer features. Some guidelines were also created to help with future studies that contain a focus on user experience instead of complicated implementations, keeping questionnaires simple, doing pilot runs with people outside your own peer group, and keeping notes and being prepared for potential issues.
Qualitative feedback from the study pointed to future works. Many participants wished for a feature to walk around in the areas they visited, similar to Google street view. There was also a wish for more tasks within the scenes, some list of objectives on which landmarks should be photographed could increase the players engagement. The object tracking approach of using Stag markers with IMU data could be fixed to give a more precise control of the haptic camera. This approach could also be used to improve other AV objects like the smartphone and water bottle that were implemented in the past. Eventually, these features need to be tested in combination with other past features, as well as be tested by the actual target group of older adults.
Hopefully this thesis moved GeoTravel a step closer to becoming a well designed VR application that can improve the quality of life for older adults.

List of Figures

List of Tables

Bibliography

[1] Deborah L. Wells. The effects of animals on human health and well-being. *Journal of Social Issues*, 65(3):523–543, 2009.

[2] Howard E. LeWine. The power of the placebo effect, 2024. accessed 02.01.2025.

[3] Dom Barnard. History of vr – timeline of events and tech development, 2024. accessed 8.11.2024.

[4] Paul Mealy. The history of virtual and augmented reality, 2018. accessed 8.11.2024.

[5] Paul Milgram and Fumio Kishino. A taxonomy of mixed reality visual displays. *IEICE Trans. Information Systems*, vol. E77-D, no. 12:1321–1329, 12 1994.

[6] Julie Carmigniani and Borko Furht. *Augmented Reality: An Overview*, page 3. Springer New York, New York, NY, 2011.

[7] Ronald Azuma, Yohan Baillot, Reinhold Behringer, Steven Feiner, Simon Julier, and Blair Macintyre. Recent advances in augmented reality. ieee comput graphics appl. *Computer Graphics and Applications, IEEE*, 21:34, 12 2001.

[8] Stefaan Ternier, Roland Klemke, Marco Kalz, Patricia Ulzen, and Marcus Specht. Ar learn: Augmented reality meets augmented virtuality. *Journal of Cheminformatics - J Cheminf*, 18, 01 2012.

[9] Kenneth Walsh and Suzanne Pawlowski. Virtual reality: A technology in need of is research. *Communications of the AIS*, 8, 03 2002.

[10] Daniel Mestre, Philippe Fuchs, A Berthoz, and JL Vercher. Immersion et présence. *Le traité de la réalité virtuelle. Paris: Ecole des Mines de Paris*, pages 309–38, 2006.

[11] Jonathan Steuer. *Defining virtual reality: dimensions determining telepresence*, page 33–56. L. Erlbaum Associates Inc., USA, 1995.

[12] Bob G. Witmer and Michael J. Singer. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments*, 7(3):225–240, 06 1998.

[13] Mel Slater. Immersion and the illusion of presence in virtual reality. *British journal of psychology*, 109(3):431–433, 2018.

[14] Nathaniel I. Durlach and Mel Slater. Presence in shared virtual environments and virtual togetherness. *Presence: Teleoperators & Virtual Environments*, 9:214–217, 2000.

[15] Toqeer Ali Syed, Muhammad Shoaib Siddiqui, Hurria Binte Abdullah, Salman Jan, Abdallah Namoun, Ali Alzahrani, Adnan Nadeem, and Ahmad B. Alkhodre. In-depth review of augmented reality: Tracking technologies, development tools, ar displays, collaborative ar, and security concerns. *Sensors*, 23(1), 2023.

[16] Industrial Inspection & Analysis. What are the 6 degrees of freedom (6dof) explained?, n.d. Accessed: 2025-02-21.

[17] Hirotake Ishii. Augmented reality: Fundamentals and nuclear related applications. *International Journal of NUCLEAR SAFETY AND SIMULATION*, 1, 12 2010.

[18] Daisy Dai. Pose tracking methods: Outside-in vs inside-out tracking in vr, 2025. Accessed: 2025-02-27.

[19] Kaihao Xu. Survey of ar object tracking technology based on deep learning. *Frontiers in Computing and Intelligent Systems*, 4:95–99, 07 2023.

[20] Charley Lin, Chaiwoo Lee, Dennis Lally, and Joseph F. Coughlin. Impact of virtual reality (vr) experience on older adults' well-being. 2018.

[21] Ryan C. Moore, Jeffrey T. Hancock, and Jeremy N. Bailenson. From 65 to 103, older adults experience virtual reality differently depending on their age: Evidence from a large-scale field study in nursing homes and assisted living facilities. *Cyberpsychology, Behavior, and Social Networking*, 26(12):886–895, 2023. PMID: 38011717.

[22] Rachel E. Brimelow, Bronwyn Dawe, and Nadeeka Dissanayaka. Preliminary research: Virtual reality in residential aged care to reduce apathy and improve mood. *Cyberpsychology, Behavior, and Social Networking*, 23(3):165–170, 2020. PMID: 31829729.

[23] Kiran Ijaz, Tram Thi Minh Tran, Ahmet Baki Kocaballi, Rafael A. Calvo, Shlomo Berkovsky, and Naseem Ahmadpour. Design considerations for immersive virtual reality applications for older adults: A scoping review. *Multimodal Technologies and Interaction*, 6(7), 2022.

[24] Nahal Norouzi, Kangsoo Kim, Gerd Bruder, and Greg Welch. Towards Interactive Virtual Dogs as a Pervasive Social Companion in Augmented Reality. 2020.

[25] Elizabeth Bouquet, Ville Mäkelä, and Albrecht Schmidt. Exploring the design of companions in video games. page 145–153, 2021.

[26] Katharina Emmerich, Patrizia Ring, and Maic Masuch. I'm glad you are on my side: How to design compelling game companions. page 141–152, 2018.

[27] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100, 2012.

[28] Alexandra Sierra Rativa, Marie Postma, and Menno van Zaanen. The uncanny valley of the virtual (animal) robot. In Munir Merdan, Wilfried Lepuschitz, Gottfried Koppensteiner, Richard Balogh, and David Obdržálek, editors, *Robotics in Education*, pages 419–427, Cham, 2020. Springer International Publishing.

[29] Joshua Achiam, Sasha Adler, Sandhini Agarwal, Lama Ahmad, Arun Ahuja, Igor Babuschkin, Anton Bakhtin, Shyamal Balaji, Pallavi Baljekar, Hongyu Bao, and et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[30] OpenAI. Realtime api guide, 2025. Accessed on June 27, 2025.

[31] Fabien Danieau, Julien Fleureau, Philippe Guillotel, Nicolas Mollet, Marc Christie, and Anatole Lécuyer. Toward haptic cinematography: Enhancing movie experiences with camera-based haptic effects. *IEEE MultiMedia*, 21(2):11–21, 2014.

[32] Shakiba Davari, Feiyu Lu, Yuan Li, Lei Zhang, Lee Lisle, Xueting Feng, Leslie Blustein, and Doug Bowman. Integrating everyday proxy objects in multi-sensory virtual reality storytelling. 12 2021.

[33] Cathy Mengying Fang, Ryo Suzuki, and Daniel Leithinger. Vr haptics at home: Repurposing everyday objects and environment for casual and on-demand vr haptic experiences. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA, 2023. Association for Computing Machinery.

[34] Stefan Krumpen, Reinhard Klein, and Michael Weinmann. Towards tangible cultural heritage experiences—enriching vr-based object inspection with haptic feedback. *J. Comput. Cult. Herit.*, 15(1), December 2021.

[35] Lucie Kruse, Sukran Karaosmanoglu, Sebastian Rings, Benedikt Ellinger, and Frank Steinicke. Enabling immersive exercise activities for older adults: A comparison of virtual reality exergames and traditional video exercises. *Societies*, 11(4), 2021.

[36] Sukran Karaosmanoglu, Sebastian Rings, Lucie Kruse, Christian Stein, and Frank Steinicke. Lessons learned from a human-centered design of an immersive exergame for people with dementia. *Proc. ACM Hum.-Comput. Interact.*, 5(CHI PLAY), October 2021.

[37] Wan Abdul Aliim Wanali, Markus Dresel, and Nicole Jochems. Human-animal interaction in immersive virtual reality: The role of social presence and positive effects. In *Proceedings of Mensch Und Computer 2024*, MuC '24, page 342–359, New York, NY, USA, 2024. Association for Computing Machinery.

[38] Lucie Kruse, Sebastian Rings, and Frank Steinicke. My focus was on the game: Investigating the use of ai assistants in a virtual reality exergame for older adults. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA, 2025. Association for Computing Machinery.

[39] Myeon-Gyun Cho. A study on augmented reality-based virtual pets for the elderly living alone. In *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1280–1283, 2021.

[40] Nahal Norouzi, Kangsoo Kim, Gerd Bruder, and Greg Welch. Towards Interactive Virtual Dogs as a Pervasive Social Companion in Augmented Reality. In Alexander Kulik, Misha Sra, Kangsoo Kim, and Byung-Kuk Seo, editors, *ICAT-EGVE 2020 -*

International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments - Posters and Demos. The Eurographics Association, 2020.

[41] Yan Ai, Hamdi Ben Abdessalem, and Claude Frasson. *Zoo Therapy for Alzheimer's Disease with Real-Time Speech Instruction and Neurofeedback System*. 09 2021.

[42] Edwin Olson. Apriltag: A robust and flexible visual fiducial system. In *2011 IEEE International Conference on Robotics and Automation*, pages 3400–3407, 2011.

[43] S. Garrido-Jurado, R. Muñoz-Salinas, F.J. Madrid-Cuevas, and M.J. Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014.

[44] APRIL robotics lab. Apriltag, 2025. accessed 22.02.2025.

[45] Burak Benligiray, Cihan Topal, and Cuneyt Akinlar. Stag: A stable fiducial marker system. *Image and Vision Computing*, 89:158–169, 2019.

[46] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016.

[47] Manuel Stoiber, Martin Pfanne, Klaus H. Strobl, Rudolph Triebel, and Alin Albu-Schaeffer. A sparse gaussian approach to region-based 6dof object tracking. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[48] George Nassif. Development of an elderly friendly user interface by adopting controller-based navigation in a virtual reality travel application. Master's thesis, Technische Universität München, November 2020.

[49] Alexander Williams. Enabling seniors to interact in virtual reality, April 2021.

[50] Julia Schwan. Companion guidance and interactive components for an elderly-friendly virtual reality travel application, August 2021.

[51] Denitsa Aleksandrova Asova. A virtual reality travel application for elderly people focusing on object integration and interaction with augmented virtuality. Master's thesis, Technische Universität München, November 2023.

[52] OpenAI. Audio api guide, 2025. Accessed on June 27, 2025.

[53] OpenCV Team. Camera calibration with opencv. `https://docs.opencv.org/4.12.0/dc/dbb/tutorial_py_calibration.html`, 2024. Accessed: 2025-09-29.

[54] Manfred Stoiber. Stag: A stable, occlusion-resistant fiducial marker system. GitHub repository, 2025. MIT Lizenz, Stand: 29. September 2025, `https://github.com/ManfredStoiber/stag`.

[55] Polycam: 3d scanner & editor. Google Play Store, 2025. Stand: 26. September 2025, https://play.google.com/store/apps/details?id=ai.polycam.

[56] Blender — a 3d modelling and rendering package. `https://www.blender.org/`, 2025. Stand: 29. September 2025, Blender Foundation, https://www.blender.org/.

[57] Harvard Humanitarian Initiative. Kobotoolbox. `https://www.kobotoolbox.org/`, 2025. Accessed: 2025-09-29.

[58] John Brooke. *SUS – a quick and dirty usability scale*, pages 189–194. 01 1996.

[59] Thomas Schubert, Frank Friedmann, and Hubert Regenbrecht. The experience of presence: Factor analytic insights. *Presence: Teleoperators and Virtual Environments*, 10(3):266–281, 2001.

[60] Bob G. Witmer, Christian J. Jerome, and Michael J. Singer. The factor structure of the presence questionnaire. *Presence: Teleoperators and Virtual Environments*, 14(3):298–312, 2005.

[61] Frank Biocca and Chad Harms. Networked minds social presence inventory: Scales only, version 1.2. Technical report, Media Interface and Network Design (M.I.N.D.) Lab, Michigan State University, 2002.

[62] Tianzheng Shi and Oliver Schneider. Development and initial validation of the haptic experience inventory (hxi). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery.

[63] I-group. Igroup presence questionnaire (ipq) items, 2025.

[64] Frank Biocca, Chad Harms, and Jennifer Gregg. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In *Proceedings of the 4th International Workshop on Presence*, pages 1–9, Philadelphia, PA, 2001.

[65] Tianzheng Shi and Oliver Schneider. *HXI Manual*. Haptic Experience Lab, 2025.

[66] Tianzheng Shi and Oliver Schneider. Hxi items (pdf). Online PDF, 2025.

[67] Google LLC. Google street view. `https://www.google.com/streetview/`, 2025. Accessed: 2025-09-29.

[68] InnerDrive Team and Bradley Busch (Editor). The 10 principles of cognitive load theory. `https://www.innerdrive.co.uk/blog/principles-cognitive-load-theory/`, 2025. 6 min read.