Removing Spatial Boundaries in Immersive Mobile Communications

Jacob Young

a thesis submitted for the degree of

Doctor of Philosophy

at the University of Otago, Dunedin,

New Zealand.

30^{th} March 2020

Abstract

Despite a worldwide trend towards mobile computing, current telepresence experiences focus on stationary desktop computers, limiting how, when, and where researched solutions can be used. In this thesis I demonstrate that mobile phones are a capable platform for future research, showing the effectiveness of the communications possible through their inherent portability and ubiquity.

I first describe a framework upon which future systems can be built, which allows two distant users to explore one of several panoramic representations of the local environment by reorienting their device. User experiments demonstrate this framework's ability to induce a sense of presence within the space and between users, and show that capturing this environment live provides no significant benefits over constructing it incrementally.

This discovery enables a second application that allows users to explore a three-dimensional representation of their environment. Each user's position is shown as an avatar, with live facial capture to facilitate natural communication. Either may also see the full environment by occupying the same virtual space. This application is also evaluated and shown to provide efficient communications to its users, providing a novel untethered experience not possible on stationary hardware despite the inherent lack of computational ability available on mobile devices.

Acknowledgements

I would like to thank my supervisors, Tobias Langlotz, Steven Mills, and Holger Regenbrecht, in guiding me through my PhD and ensuring I didn't get carried away with making cool stuff instead of doing some real science, and Alexander Plopski for his valuable feedback regarding this thesis.

I also thank the other members of the HCI lab for their valuable ideas and ability to act as rubber ducks, particularly to Matthew Cook for his knowledge in mathematics and computer graphics, and to Rosa Lutz, Jonathan Sutton, and Oliver Reid for their help in conducting user studies with only craft beer as an incentive.

Many thanks also to Heather Cooper and Gail Mercer for organising my life for me, particularly to Heather as she enters retirement and leaves the department to descend into anarchy.

Finally, I give particular thanks to my fiancée Katie Tong, without whose moral, financial, and transportational support this research would not be possible.

Contents

List of Figures

Chapter 1

Introduction

People desire to be connected. This simple fact has guided millennia of technological progress, from the establishment of the first public postal service in the 6^{th} century BCE (Xenophon, 2005) to the invention of telegraphy in 1832 (History.com Editors, 2009b), the telephone in 1876 (History.com Editors, 2009a), and publicly available videoconferencing in 1936 (VSee, 2011). Each aimed to improve not only the convenience but the naturalness of conversation available; mail could take weeks to arrive, whereas telegrams could be received within a day. Telephones transmitted words instantaneously and allowed more personal speech rather than text, which videoconferencing made more personal by allowing the subtle cues provided by facial and body language.

Despite this, each still suffered from the same issue: their use was a deliberate process, often requiring premeditation in advance of when one actually wanted to use it. To send a letter one must go to the post office, to make a call one must be near a telephone, and to make a videocall one must be near a computer with a webcam and access to the internet. This inconvenience didn't occur only to the sender either; mail can take days to weeks to arrive, so no guarantee is made on when it will be seen, and telephones were only viable if the recipient also happened to be near the right phone at the time of the call.

This all changed with the introduction of the mobile phone. No longer was remote communication confined to one location, as calls could be made wherever was convenient for the initiator. The location of the receiver was also no longer an issue as they too were likely to be carrying their phone with them, making unanswered calls a rare occurrence. This immediacy fundamentally changed how communication takes place by removing all premeditation, allowing a "smallness in conversation" not possible before (Arminen and Weilenmann, 2009). With calls taking no effort to make or receive,

they could be made for any purpose, no matter how trivial.

Though more convenient, the communications media available to the average consumer have largely remained identical to those in 1936. No new means of interaction have been adopted by the wider population, meaning all current forms of communication were developed before the mobile phone and thus fail to take advantage of the ubiquity and convenience it introduced. Enterprise applications have gone in the complete opposite direction, setting aside entire rooms containing tens to hundreds of thousands of dollars worth of equipment to achieve what a mobile phone has been capable of for years. Videoconferencing thus remains the golden standard of what the average person can currently experience, and its manifestation on mobile phones is identical to that of desktop computers despite the difference in how these devices are used and what they are capable of.

1.1 The Emergence of Telepresence

This isn't to say that further methods of communication have yet to be researched. While current technologies make it obvious that speakers are in separate places and talking through some intermediate medium, the field of *telepresence* has emerged which attempts to obscure this fact and make it feel as if two speakers really are *present* in the same location. This would not only benefit social scenarios, but also collaborative tasks such as remote assembly (Fussell et al., 2000) and even crime scene investigation (Poelman et al., 2012). Many ways to achieve this have been researched, from simulating eye (Anjos et al., 2019) or hand contact (Wang and Quek, 2010) to physically manifesting the remote partner's actions through mechanical manipulation of the remote physical environment (Leithinger et al., 2014).

The most common approach is to allow users to fully explore their partner's space rather than have their view locked in the direction of the camera. This can be in the form of simple *egocentric* views from their partner's perspective, allowing rotational independence so that they can obtain novel views from one fixed viewpoint (Gauglitz et al., 2012; Kratz et al., 2014), or providing full freedom to move about a three-dimensional representation of that shared space as they wish (Fanello et al., 2016; Park et al., 2019) in an *exocentric* view. Giving users this freedom to explore remote environments has been shown to significantly increase their sense of presence within them (Jo and Hwang, 2013), and telepresence applications with this functionality tend to be preferred to traditional systems where their view would be locked to the direction

Figure 1.1: A Cisco Telepresence System[1], currently seen as the state-of-the-art in enterprise telepresence solutions. While one could fool themselves that their conversational partner really is on the other side of the desk, this requires an entire room to be set aside for a very specific type of conversation that provides no meaningful interaction methods between users.

of the camera (Gauglitz et al., 2012, 2014). This independence also proves beneficial in collaborative scenarios as it can lead to more efficient communication, resulting in tasks being completed faster (Fussell et al., 2000, 2004) with collaborators being more confident in the end result (Kasahara and Rekimoto, 2014).

Another possible step forward for remote communications is to allow gestures to be used to enhance conversation in a natural manner. In current videoconferencing systems these are technically supported as they can be shown in the regular camera stream, however this completely detaches them from their wider context and can make them difficult to understand as their perceived locations are not consistent between the two viewers. Proposed solutions include overlaying video of the user's hands on captures of the environment (Fussell et al., 2004; Kim et al., 2014) or using some other surrogate shown through a display of some kind (Kasahara and Rekimoto, 2014; Sodhi et al., 2013), though the former is often preferred (Fussell et al., 2004; Kim et al., 2014). Either approach can significantly reduce the time it takes to complete shared tasks (Fussell et al., 2004) by increasing the coordination between interlocutors (Fussell et al., 2000), and can ensure subtle body language conveying the user's current emotional state or understanding of shared instructions are not missed (Flor, 1998).

Despite their benefits, these advancements have yet to see any kind of use outside of the laboratories they are developed in or the large-scale enterprises that can afford them, whereas previous technological leaps have seen widespread adoption within years

[1]https://www.cisco.com/c/en/us/products/collaboration-endpoints/immersive-telePresence/index.html

of their invention. This supposed lack of interest may be due to the platforms these solutions tend to be developed for; many researchers seem to follow the mantra of "bigger is better", and thus use the most powerful and expensive desktop computers on hand in order to cram as many features as they can into their systems with the highest fidelity possible. This has led to many convincing experiences, however in their pursuit existing developments are discarded before they can be refined and thus never become feasible in an affordable way.

Focusing on these desktop systems also completely ignores a worldwide trend toward mobile computing. According to a 2019 report by Ofcom (2018), 94% of UK citizens aged 16 and over own their own mobile phone, with this number increasing each year. In comparison, only 24% own a desktop computer and 60% a laptop, both of which are seeing steep declines each year. Those that do own a desktop tend to value it less than their mobile phone, even in scenarios where graphical fidelity would usually be important; analytics platform Newzoo (2020) reports that the mobile games market is twice as large as the PC one and is increasing at four times the rate each year. This could be due to the increased usage opportunities presented by these mobile devices, however even at home smartphones are used for internet access more than desktops and laptops combined, three quarters of which is communication with a distant person in some manner (Ofcom, 2018).

Development of immersive communications for stationary systems thus seems difficult to justify and is a possible explanation for the supposed lack of interest from the wider public who are unwilling to regress to non-mobile communication. A shifted focus from desktop to mobile development would immediately bring many benefits to future communications solutions, allowing people to connect with each other wherever they may be without worry of where their communication partner is and what they're doing. Most importantly, any proposed solutions would be immediately available to most of the population using hardware they already own and prefer to use for such purposes.

1.2 Research Goals

With this thesis thus I explore how an immersive telepresence experience could be achieved using mobile devices, making any solutions immediately available to most consumers without requiring any further financial investment on their part. Such a solution would both be more accessible and more appealing, which could lead to

widespread adoption of new communications media and thus an improvement in how we communicate in our day-to-day lives. In doing so I must discover how communication in co-located scenarios is naturally performed, which techniques interlocutors use that contribute to a sense of naturalness and efficiency in communication, and which of these techniques can be replicated in telepresence systems to bring these benefits to remote scenarios.

Given the modest computational capabilities of mobile phones, these techniques must also be evaluated in their computational efficiency as any which are too demanding would remain out of reach of users while they wait for improvements in mobile hardware. Any implementation of these features would also serve as a contribution to the literature as the mobile phone has so far been ignored as a self-contained telepresence platform.

Finally, I also wish to explore how the portability and ubiquity of mobile devices can create a unique telepresence experience not possible on stationary hardware. It is my hope that this will encourage a refocused effort on mobile telepresence, allowing the findings of other researchers to no longer be restricted to laboratory or enterprise use and instead immediately benefit the public, eventually causing a paradigm shift in remote communications similar to the others seen through the introduction of previous technologies.

1.3 Contributions of this Thesis

In this thesis I explore how these experiences can be realised on mobile devices and the unique benefits this shift in focus could bring to remote communication. I begin by exploring theories of how this communication takes place and can thus become more effective, how the various techniques interlocutors use manifest in remote scenarios, and how the effectiveness and quality of communication can be evaluated and thus improved in a quantitative way. Using this knowledge, I then identify several requirements a remote communications system must meet in order to provide a rich, effective, and natural communications medium between two distant parties. Several existing telepresence systems are then evaluated to determine whether these requirements have been met, and if not, where they are so far lacking.

Once these requirements have been identified, I detail the implementation and evaluate the effectiveness of two applications I have developed for mobile devices that attempt to meet them. Each attempts to take advantage of the portability, ubiquity and

untethered nature of these devices to allow unique experiences not previously available, while also accounting for their decreased computational ability when compared to the systems usually used for such development.

The first, shown in Figure 1.2, allows distant users to freely view a 360° panorama of their communication partner's environment by rotating their device as if viewing it through their phone's integrated camera. Such viewpoint independence has been shown to increase a user's sense of presence within this space (Jo and Hwang, 2013), however with the limited computational capabilities of mobile devices there is a limit to how detailed and up-to-date this presentation of the environment can be. Five separate methods of constructing it were developed, each providing increased independence within the shared space either through increasing how much wider context could be seen outside of the transmitted video or how much of this could be viewed live rather than as static images. Users could also communicate through gestures, which were captured, transmitted to their partner and rendered unmediated in the environment, and their current view direction was shared to aid in context-dependent statements.

Each of the five environmental representations were compared through user testing to determine if a "sweet spot" existed where the extra computation required to increase the remote users's view independence was not worth the subsequent increase in presence; this is important as the easier an environment is to compute, the more devices it will be available on and thus the larger its potential audience. It was found that incrementally reconstructing a static view of the environment to supplement the live video stream is just as effective at inducing a sense of presence within that space as transmitting a full live panorama, showing that an immersive experience can be had without requiring an external 360° camera. This result, and the implementation details of the system used to discover it, were published in the IEEE Transactions on Visualization and Computer Graphics (Young et al., 2019).

This finding allowed the creation of a second application, shown in Figure 1.3, which brought this immersive experience into three dimensions where full live reconstructions would not be feasible without external hardware. A user can instead incrementally construct a static 3D representation of their surroundings by walking through it with their mobile phone, with the captured data transmitted live to their remote peer who can also freely move through this reconstruction by simply walking through their own space. Each can see the position of the other as a 3D avatar, over which live facial capture is displayed to allow face-to-face communication within this shared space. Since mobile phones are not yet capable of constructing these 3D environments at a

Figure 1.2: The mobile telepresence framework I implemented which is described in chapter 3. Two users meet in a shared panoramic space, which they can obtain independent views within by reorienting their device. Each may also communicate through gestures, which are captured by their mobile phone's integrated camera and spatially rendered in the environment.

realistic fidelity, users can also transition to a higher-resolution 2D view captured by a 360° camera by occupying the same location within the shared space. This is the first application of its kind to be developed for mobile devices, and thus for the first time the area that can be independently explored by a remote users is completely arbitrary in its size, scope, and location.

I also explore additional interaction methods that could be used in future in mobile teleconferencing systems using this system's hardware configuration. These were developed in full and thus their implementation is detailed, however the performance of mobile devices makes them currently infeasible for real-time use. The first of these is three-dimensional gesture tracking, which uses a depth sensor embedded in a mobile phone to capture a user's gestures and spatially render them within the environment. The second is gaze-based rendering, which allows the phone's display to be used as a "smart window" by using the angle between it and their face to adjust their viewing angle within the virtual environment. The third and final is full-body tracking, which tracks user's bodies within their device's 360° video capture and displays them within the environment rather than an artificial avatar.

This second system was also tested by novice users in a live remote scenario, comparing it to simple 360° videoconferencing to determine if a similar sense of presence could be induced in participants despite the reduction in visual quality. As this is the first system of its kind, a large focus was placed on its social aspects to determine whether this new interaction method is something that could see widespread adoption in the future. Participants much preferred this new system, seeing it as some

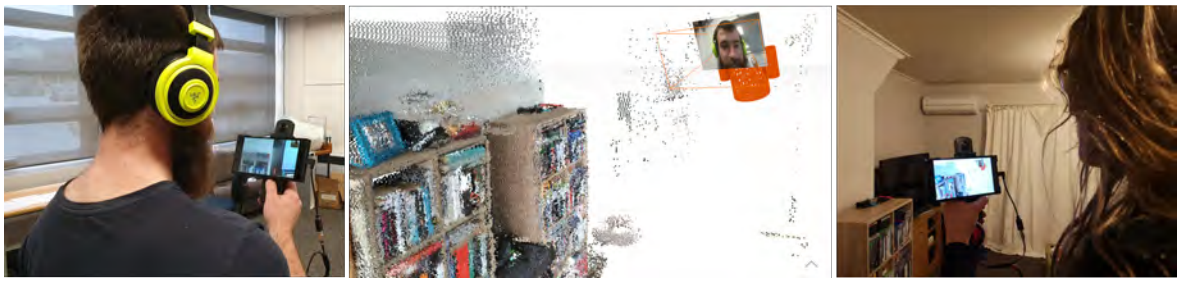


Figure 1.3: The second application developed for this work that brought full 3D interaction to mobile devices. The shared environment is now fully three-dimensional and incrementally reconstructed from the mobile phone's inbuilt sensors. Users can independently explore this virtual shared space by simply walking around their real one, with their current position and facial capture shown via a virtual avatar.

completely new experience as opposed to 360° videoconferencing which they saw as "too much like Skype", implying they do really see this as a new paradigm in remote communications and proving the feasibility of mobile phones as a platform for future telepresence research. These findings, and the implementation details of this system, were published in the Proceedings of the ACM on Interactive, Mobile, Wearable, and Ubiquitous Technologies (Young et al., 2020).

My contributions in this thesis are thus as follows:

- A framework for telepresence on purely mobile devices, described in detail so that it may be used as a basis for future research, and proving mobile phones capable of providing advanced communications experiences in real time.
- Proof of this framework's ability to induce a heightened sense of presence between speakers and within the environment, and that incrementally constructed static environments are sufficient to do so, lowering the computational requirements of future systems.
- A completely novel experience that brings 3D telepresence to mobile devices, which for the first time allows free exploration of arbitrarily large environments.
- Evidence that despite the inherent limitations in mobile processing, this 3D environment can be constructed and explored in real time, providing an enjoyable and more social experience for users while also proving mobile phones a capable and complete platform for future telepresence research.

It is my hope that these contributions will lead researchers to refocus their efforts towards mobile systems as their platform of choice for immersive telepresence sys-

tems. In doing so, not only would their research be immediately available to the wider population, but in a convenient and preferred form factor that allows the enhanced communication to take place whenever and wherever the user wishes. A general preference toward mobile hardware would make consumers more susceptible to adopting these new methods of communication, and the ubiquitous, portable, and untethered nature of these devices would enable brand new experiences never before seen.

Chapter 2

Related Work

To create an immersive collaborative experience we must first define how such collaboration occurs. This has been a commonly researched subject for many years, with many researchers discovering and defining what makes conversation efficient and natural for its participants. I begin by exploring the theories of how communication between people takes place through *conversational grounding* and a mutual sense of *presence*, then discuss how these processes can be enhanced in collaborative systems to make conversation as natural as it would be in real life. With this knowledge, I identify a list of requirements a telepresence system must meet in order to provide an efficient communications media to its users. Finally, I discuss previous systems that have attempted to fulfil these requirements and determine whether they have successfully done so.

2.1 Theories of Collaboration

Conversations rely on mutual knowledge between their participants. To discuss an object, everyone must agree which object is being talked about. To discuss a person, the same must be true. To discuss a task, all involved must know what is required and how it is to be done. Any discussion otherwise becomes meaningless without extensive descriptions of all objects involved. To avoid this, *common ground* between interlocutors must be established through a process called *conversational grounding* (Clark and Brennan, 2008), which occurs naturally in any conversation without conscious effort.

Grounding often manifests itself as a series of "turns" taken by speakers, each correcting the incorrect assumptions of the other as they slowly converge on some mutual understanding of the topic. This process begins with the presentation of an idea; this can be a statement, an instruction, or a question, each of which has a common goal:

to be acknowledged and understood by the recipient, after which common ground can be assumed to be established. Without this acknowledgement, the speaker will assume that they were not heard and will repeat their initial presentation until either they are sure it has been understood or they get frustrated and give up, with the latter case obviously being undesirable in any scenario.

This acknowledgement can come in many forms, and unless all are supported by the communication medium, conversation will be constantly interrupted as speakers assume they are not understood. The simplest is verbal confirmation, which is easily supported by any medium that transmits the user's voice. Others are more complicated; in a study of how people interact in side-by-side collaborative scenarios, Flor (1998) found that acknowledgement can often be conveyed with body language such as a nod, or more subtle movements such as a change in posture, and that speakers would often look to their collaborator to see these cues after making a presentation. This would necessitate that users can see each other during these tasks, however this acknowledgement can also manifest as the recipient performing a suggested action instead, also necessitating a view of the task space in collaborative scenarios (Kuzuoka et al., 2000).

Previous work has thus shown that collaborators work most efficiently when they are physically co-located within the task space (Fussell et al., 2000, 2004) as both of these views are easily available, thus to aid in remote collaboration it would be advantageous to study what resources co-located collaborators use. Flor (1998) proposes that collaboration is achieved through the *pushing* and *pulling* of information across media in the collaborative task space, which differ in how the exchange is initiated and have different requirements that must be met before they can occur. Pushing of information is when one person forces information onto the other; this requires that the information to be pushed can be freely shared with the conversational partner and that no conscious effort is required on their part to receive it. Pulling occurs when external information is obtained through a deliberate act, for example by looking at the other person's screen; to initiate a pulling exchange, the remote user must have some way of freely viewing and receiving information from their partner without that partner's involvement in the exchange. As seen in Figure 2.1, side-by-side communication is thus highly efficient in this regard due to the numerous methods of initiating these information exchanges it provides. To ensure an optimal environment for remote collaboration one must then emulate these co-present conditions as closely as possible to ensure that these channels of communication and subsequent opportunities for grounding are not lost.

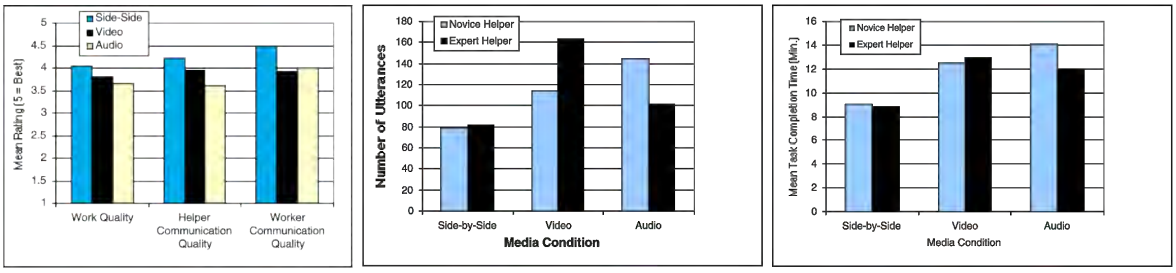


Figure 2.1: The quality of communication (left), number of utterances spoken during a collaborative task (centre), and time taken to complete that task (right) compared between side-by-side communication, video and audio teleconferencing, and audio-only teleconferencing, as reported by Fussell et al. (2000).

Flor (1998) suggests that each collaborator subconsciously creates four internal representations of the task space: one of the task (such as which file the user should edit), one of the compositional structure of the task subject (such as how this file is structured), one of this subject's behaviour (such as how the file behaves when it's edited), and one of modifications applied to that subject (such as which edits have already been made). Clark and Brennan's theories of conversational grounding (Clark and Brennan, 2008) thus apply not only to social contexts but to collaborative ones as well, as grounding between collaborators aids in ensuring these representations become mutually consistent between collaborators. The internal model each has of the problem affects their proposed solutions, and thus differing representations between collaborators result in time wasted on differing solutions. This grounding is achieved not only through conversation, but through any interactive media present within the task space that allows the aforementioned pushing and pulling exchanges to be initiated.

This theory of grounding as applied to collaboration is supported by Dourish and Bellotti (1992), who believe that collaboration requires two kinds of awareness. *Character awareness* is high-level knowledge of the general task the collaborative partner is performing, such as knowing which file they're editing. *Content awareness* is more fine-grained awareness of their exact actions, such as knowing exactly what they are typing in that file. Dourish and Bellotti argue that effective collaboration requires seamless shifting between these two forms, and that this shifting is aided through the use of collaborative tools that provide feedback known to all parties; in other words, this shifting requires that sufficient common ground has been established between collaborators, that the representation of the task space stays consistent between them, and that information about the task can be freely pushed and pulled as required.

To test these theories, Flor (1998) observed how two programmers interacted when

modifying an existing piece of computer code. Each was seated next to the other and given a terminal that was easily within reach of their partner, allowing them to type on the other keyboard if so desired. The participants often explained their actions to the other to create a shared understanding of them. Once an issue arose that caused the code to behave in an unexpected manner, one participant (P1) turned to watch the other (P2) attempt to fix it, showing a seamless transition from character awareness to content awareness achieved by pulling information via a simple turn of the head. P2 explains his rationale behind the fix he was attempting, creating common ground in their internal representations of task subject modification, and P2 watched P1's screen whilst doing so to adapt his instructions to her current representation of the task space. Once P1 resolved the issue, she looked at P2's face to gauge his reaction to the news, providing a convenient visual hint as to how his model of the task has changed. Once they resumed work on the rest of the task, the two participants were often observed glancing at each other's screens to ensure they were not working on the same file to reduce conflicts and redundancy, often without the knowledge of the other so as to not distract them from their current action.

Therein lies the problem with remote collaboration in its current form. If grounding requires pushing and pulling across various collaborative media, then what good is a traditional system that provides only two such media, namely voice and unexplorable video? To make matters worse, one of these is essentially crippled; the dependence of video on the direction of the local camera renders it useless for pulling information, so it can only push whatever data the local user decides to point it at. This disallows such simple interactions as glancing from the workspace to the collaborative partner's face as it now introduces the additional step of verbally requesting that the camera be moved. The local user now dictates all attention, so pulling becomes impossible as the remote user's understanding of the task and thus their internal representation of it is completely dependent on the local user's possibly incorrect assumptions.

2.2 Presence

The degree of presence a user achieves within a virtual environment determines how they will perceive and interact with it. Once sufficient presence is achieved, a user's mental model of the virtual environment shifts from one on a screen to one within it, and they begin thinking not in terms of how their actions affect the system but how these actions affect the virtual objects within (Schubert et al., 2001). This allows

for complex interactions such as triggering fear through virtual stimuli (Regenbrecht et al., 1998) or providing sufficient ownership of virtual limbs for repair of neurological damage to the real ones to occur (Regenbrecht et al., 2011).

It's important to note here the distinction between presence and immersion (Schubert et al., 2001). Immersion is a property of the hardware or software used, and refers to its ability to provide realistic or convincing stimuli to the user. For example, a display with a high resolution would provide more immersion than a low-resolution one, and a head-mounted display would provide more still due to the ability it gives to naturally manipulate the virtual camera through head movements. Presence, on the other hand, is a psychological phenomenon rather than a technological one, and can be achieved with or without immersive hardware; for example, a low-resolution video game may provide more presence than a non-interactive virtual reality demo.

Several forms of presence have been identified, though for the purpose of this work we will focus on three: spatial presence, social presence, and co-presence. Each is distinct but intertwined, and the degree to which one is obtained can affect the perception of others.

Spatial presence is the most straightforward form to understand and the least controversial to define. It is the degree to which a user feels physically located within a real or virtual environment; the sense of really "being there" rather than viewing it through some surrogate (Biocca et al., 2003). The sense of spatial presence within one's own environment would be extremely high and serve as the upper limit that is achievable, whereas one would feel no spatial presence at all within a distant environment they have no awareness or understanding of.

Social presence is much more difficult to rigidly define due to the subjectivity inherent in any social context. Some have described it as the sense of "being there with another" (Müller et al., 2016; Tait and Billinghurst, 2015), making it akin to spatial presence with a person rather than an environment. Lombard and Ditton (1997) take a more technological approach and further develop the link between these two forms of presence by describing it as the degree to which conversation feels unmediated, implying that a higher degree of spatial presence within a space shared by two people would also increase the social presence felt between them. Technologies that provide this sense are often described as "warm, personal, sensitive, and sociable" (Hauber et al., 2005), and can provide more natural speaking environment as conversation is not dictated by available media.

Co-presence is more abstract and subjective than the other forms and is thus more

difficult to rigidly define. Müller et al. (2016) describe it as the sense of "being together in one place", implying it to be some combination of spatial and social presence. We shall instead use the definition by Campos-Castillo and Hitlin (2013), who counter this with the example of two people talking on the phone; despite the physical distance between the callers, they could feel as if they are more "present" with each other than they are with any strangers that they are spatially proximate to but have only a passing awareness of. Campos-Castillo and Hitlin account for this by instead defining co-presence as *mutual entrainment* between parties, where "entrainment" is a synchronisation of mutual attention, emotion, and behaviour. The emphasis here is on mutuality; these requirements must be met and felt by both parties, and each must feel that the feeling is reciprocated for true co-presence to be achieved. Spatial co-location can thus help in achieving co-presence between two parties as social and emotional conversation cues are more easily shared (Flor, 1998), but is not required in this definition and can thus be achieved in remote communications.

Achieving presence within a shared environment or with a communication partner can thus aid in conversation by reducing the number of turns required to establish common ground. Spatial presence gives common environmental context to both parties; if they both feel a high degree of presence within the same shared environment, then both will be receiving the same spatial and environmental cues and do not need to establish these separately. If a remote peer further has fully independent control within that space, then the environment itself can be used as a means to pull information such as done by the programmers in Flor's experiments (Flor, 1998). Social and co-presence similarly ease the grounding process by establishing mutual emotional understanding between parties, making it less likely that important conversational cues will be missed where grounding would need to be re-established. This sense of "grounding" can be quite subjective and have different meanings in different contexts, making it rather difficult to unambiguously measure. Presence, however, is well established as a tool to measure conversational efficiency through several industry-standard questionnaires (Schubert et al., 2001; Bailenson et al., 2005; Hauber et al., 2006; Biocca et al., 2003), and so will prove as a useful proxy through which we can measure the richness of conversation a system can provide.

2.3 Requirements for Efficient Communication

Much research has gone into what a telepresence system would need to provide these various forms of presence and subsequent opportunities for grounding in and between its users. Most findings relate to either how the shared environment is presented to its inhabitants or the methods of interaction possible between them. Here I outline the suggestions previous researchers have made regarding how these collaborative systems should be made in order to increase the sense of presence they induce and thus the potential for highly grounded and therefore efficient communication between their users.

2.3.1 Presentation of the Environment

Luff et al. (2003) argue that many existing telepresence solutions focus too much on face-to-face communication and ignore the user's wider environmental context. They believe that social interaction is accomplished largely through objects in the environment, and that users often make assumptions about the remote peer's knowledge of their local environment that turn out to be false. This can lead to frustration as instructions may have to be repeated or reworded in greater detail, requiring common ground to be re-established before communication can continue. For this reason users must have an intuitive and efficient means of interacting with their environment, and the relationship between users and objects must be constant and consistent to ensure their assumptions remain correct. This view is shared by Fussell et al. (2004), who found that views of the task space and each user's actions and direction of attention within it are more effective for conversational grounding than views of the user's face, leading to overall more efficient collaboration.

That's not to say that face-to-face communication doesn't have its benefits, and indeed the prior suggestion for environment-focused views may be due to the industry-focused nature of most experiments where efficiency is the only metric that matters. Even in such scenarios, Flor (1998) found that when performing a collaborative task side-by-side on separate terminals, participants would often look at their partner's face after voicing a solution, possibly to gauge their reaction to and agreement with the proposed solution. Kuzuoka et al. (2000) further found that in side-by-side communication, instructors would often glance at the worker to ensure that they were following and comprehending their instructions, and the worker would in turn express this comprehension through their own gestures and body arrangement, which can result in time

wasted if this comprehension needs to be explicitly confirmed (Clark and Krych, 2004; Taylor et al., 2009). This implies that even in task-oriented situations, emotional cues can still play an important part in facilitating efficient communication by confirming comprehension of instructions and thus should not be ignored.

2.3.2 Free Exploration of the Task Space

With a focus on views of the environment, one must consider how this is to be shown to each user. Fixing the remote user's field of view to the direction of the local camera has been shown to be detrimental to collaborative task performance (Fussell et al., 2000) as reorientation within the scene must be done through verbal instruction to the local user. Even then, users often find that a complete scan of the room is required before a mental model of it can be created, and afterwards are still confused about which way the camera was oriented when this scan was performed (Pece et al., 2013).

Allowing the remote user to navigate the environment freely through manipulation of a physical or virtual camera means they can perform this reorientation themself, allowing them to focus on communication rather than the medium through which they are performing it. Remote users were often observed using this freedom to focus on areas of the task space not seen by the local user (Kasahara and Rekimoto, 2014; Sodhi et al., 2013), allowing them to direct the focus of attention or describe features within the environment without intervention from the local user (Tang et al., 2017), avoid issues caused by incorrect assumptions on behalf of the local user (Flor, 1998), and use more of the environment in their discussions than would otherwise be possible (Taylor et al., 2009).

This view independence has been shown to significantly reduce the time taken to complete shared collaborative tasks with similar or greater accuracy (Jo and Hwang, 2013; Pece et al., 2013) and is preferred by users to traditional videoconferencing software (Gauglitz et al., 2012, 2014; Jo and Hwang, 2013). Participants also felt more confident that tasks had been performed correctly as they had more situational awareness within the task space (Kasahara and Rekimoto, 2014; Kratz and Ferreira, 2016) and felt more spatially present within it (Müller et al., 2016) than they would had their viewpoint been fixed. Care must be taken when creating such a space to ensure it remains consistent between peers; if sufficient fidelity, accuracy, or temporal consistency are not achieved across the connection, these benefits can often be negated (Kraut et al., 2002).

2.3.3 Indication of Gaze Direction

With the ability to freely explore the environment, the issue of coordinating viewpoints between users arises, particularly as remote partners often assume that anything visible to them will also be presented to the local user in the same manner (Luff et al., 2003). Tang et al. (2017) found that the inability of the local user to see the remote user's current gaze direction led to frustration as the remote user often assumed this information would be known when describing features in the environment. This led to complex verbal negotiation between participants to reorient themselves whenever they strayed too far from one another, and consequently they would keep their viewpoints close to avoid confusion, removing the point of having independent viewpoints at all.

This was also observed by Kuzuoka et al. (2000), who found that local users would often face the same direction as a remotely controlled robot to ensure their fields of view were similar, creating a strange reversal of roles where the remote user dictated the local user's attention rather than the inverse enforced by current teleconferencing applications. However, this reversed dictation meant that the local user could often predict which object would be referred to by the remote user before they even spoke as they could see which objects they were turning to face; this suggests that awareness of the communication partner's field of view provides an additional resource to conversation that allows for more complex interaction and provides more context for conversational grounding.

2.3.4 Gestural Interaction

Even with independent views within the environment, conversation can be a passive experience without some meaningful way for communicating parties to interact. Incorporating hand gestures into conversation can increase coordination between users (Fussell et al., 2000), significantly decreasing the time taken to perform remote collaborative tasks (Fussell et al., 2004; Gauglitz et al., 2012, 2014; Taylor et al., 2009) with no loss in accuracy (Kirk and Stanton Fraser, 2006). This is due to the context they provide to deictic references such as "this one here" or "it's over there", allowing for objects to be quickly identified without lengthy verbal description (Fussell et al., 2000; Taylor et al., 2009) and in some cases causing verbal communication to become inefficient and be avoided altogether (Bauer et al., 1999). Users can often become frustrated when these gestures aren't shared (Taylor et al., 2009) as they will still attempt to use them, even knowing they won't be seen (Tang et al., 2017), resulting in

systems that accommodate gestural interaction being preferred by users (Fussell et al., 2004; Gauglitz et al., 2014; Kim et al., 2014). Conventional video windows such as used in Skype prove insufficient for this task as users often find it difficult to associate referential gestures to the referred object due to the spatial disconnect between the video and the environment (Luff et al., 2003), meaning gestures must be specifically accommodated.

Several forms of gestures exist, and each serve to convey different types of information so all must be allowed in order for this communication channel to be utilised to its fullest. These can be divided into two main types: *Pointing gestures* are ones with specific directionality that are used to succinctly refer to people or objects through deictic references, which can often completely replace complex, lengthy sentences without any loss of information (Bauer et al., 1999) and can result in higher quality work that takes less time to complete (Fussell et al., 2000). *Representational gestures* are more complex hand movements that are used to represent the form or path of an object and are often used to show a remote peer how to move objects in order to complete a task. Though pointing makes up 70% of all gestures used during collaboration, these fail to provide any significant benefit alone if representational gestures are not also supported (Fussell et al., 2004).

Kirk et al. (2005) further divide these representational gestures into several categories not reproducible through abstract gesture surrogates. These include the obvious ones such as pointing at an object of interest, covering one that should be ignored, and mimicking movements such as rotation to show how the chosen object should be manipulated, but also less obvious ones such as users wiggling their fingers at the beginning of the task to determine the relationship between their own movements and those of their virtual gestures, wavering their hand over several potential objects while deciding which to choose, and resting their hands on a surface to signify they have finished giving the current instruction. Each of these plays an important part in conversation as they relay the helper's intentions, allowing the worker to pre-empt which action to perform before any instruction is even given.

The most profound gesture type is what the authors call the *inhabited hand*, which is when the remote helper's hands occupy the same space as the worker's and assumes a position the worker must match to solve the current task. Despite usually being impossible in real-world scenarios, users were often observed performing this gesture unprompted, implying that it came naturally to them as the most intuitive way to convey the current instruction. With such movements being difficult to predict, inter-

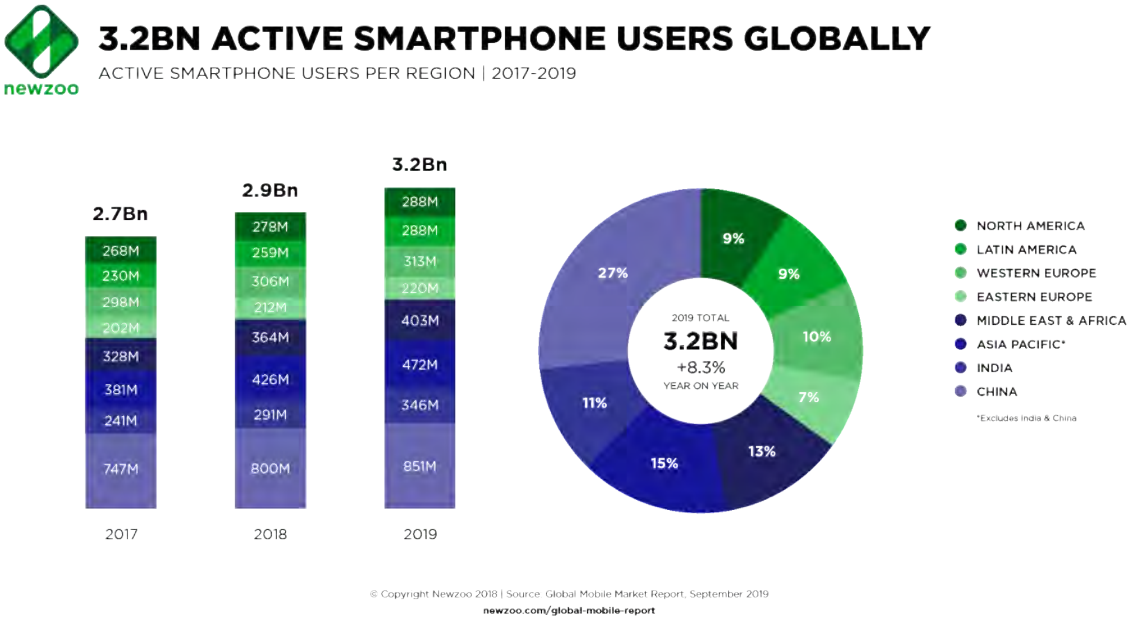


Figure 2.2: The number of active smartphone users globally per region according to Newzoo (2020).

action through a virtual surrogate becomes less viable as any unsupported movements will be missed and result in grounding needing to be re-established (Tang et al., 2017). Kim et al. (2014) and Fussell et al. (2004) thus found unmediated video of the users' hands to be more effective and preferred by users to intermediate surrogates.

2.3.5 Advantages of Mobile Systems

When discussing collaborative systems many authors place emphasis on and perform experiments through powerful and expensive desktop systems (Fanello et al., 2016; Pece et al., 2013; Stotko et al., 2019) or proprietary hardware (Fanello et al., 2016; Kasahara and Rekimoto, 2014; Kasahara et al., 2014; Kratz et al., 2014, 2015), neither of which are feasible solutions for the average user. These also significantly decrease the flexibility and portability of the proposed solution as communication may only take place either in one small, fixed location, disallowing communication in remote or outdoor areas, or require enough forethought to bring the required equipment, meaning the technology could never be used for spontaneous conversation.

Many of these issues could be resolved by a shift in focus from stationary to mobile systems. According to analytics platform Newzoo (2020), as of 2019 there were 3.2 billion active smartphone users globally, with this number increasing by 8.3% each

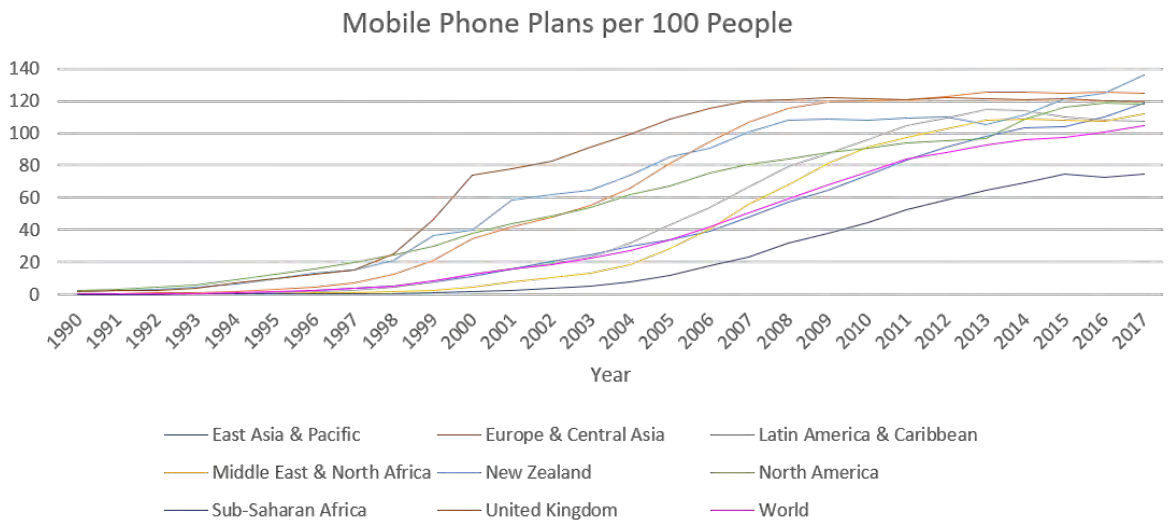


Figure 2.3: The number of mobile phone subscriptions per 100 people in each major region (Ritchie and Roser, 2019).

year (see Figure 2.2). With a global population of 7.75 billion as of the time of writing (Worldometer, 2020), this means that a teleconferencing application developed purely for mobile phones would be immediately available to 41% of the world through hardware they already own. In developed nations this number could be even higher; according to Ritchie and Roser (2019), most regions now have more active mobile phone subscriptions than people (Figure 2.3), implying that almost all will have access to a mobile phone in some way.

Not only are mobile phones widespread, but they are often the device of choice for internet-based or communicative activities. According to the same study of UK citizens by Ofcom (2018), most people rank their smartphone as the most important device for accessing the internet, leading to 41.2% of respondents using their phone as their main internet device in any location, and 37.4% using their smartphone for this purpose at home. This is despite a desktop computer likely being available, which combined with laptops only saw 22.5% of respondents using it as their primary device at home (see Figure 2.4). By far the most popular activity to do while connected is to remotely communicate with others; 54% and 52% of internet traffic from females and males respectively are generated by dedicated communication apps, with a further 28% and 21% generated by social media applications, leading to an average of 77.5% of internet traffic facilitating remote communication in some way (Figure 2.5). If any of the immersive telepresence applications proposed by researchers were developed with mobile use in mind, they would thus be more likely to be accepted by consumers than

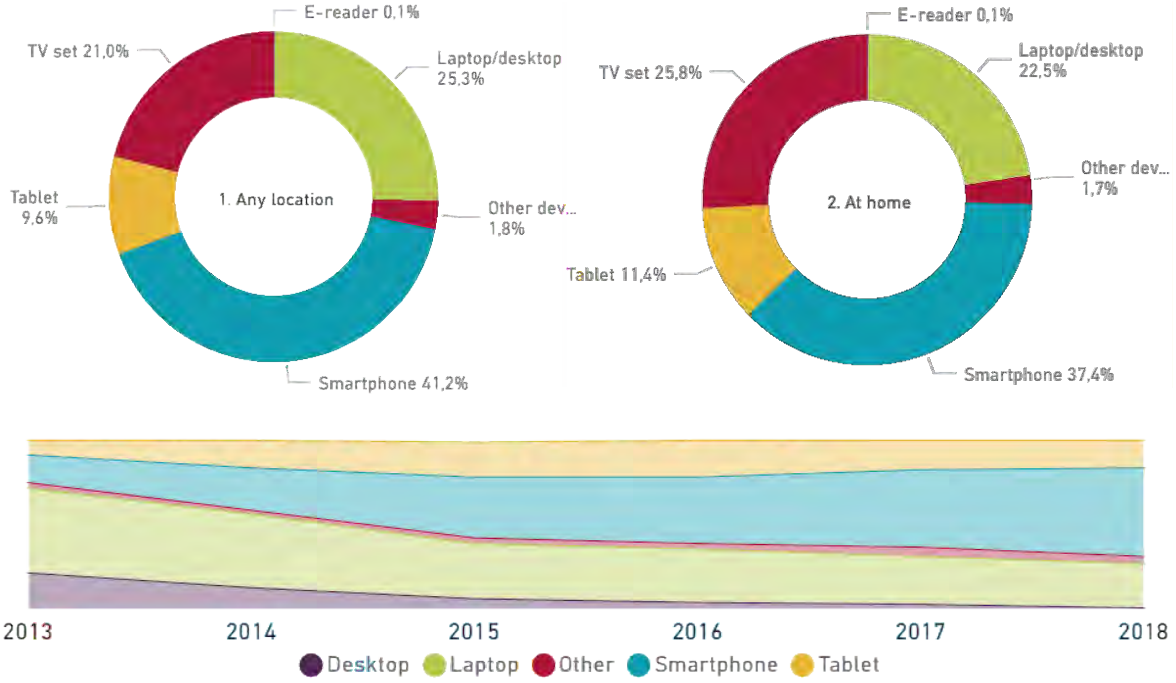


Figure 2.4: The results of a 2018 survey asking UK citizens which devices they use to access internet services at any location (top left) or at home (top right), with the proportion of people who chose each device as the most important for doing so over the last several years (bottom) (Ofcom, 2018)

if the current focus on desktop systems continued.

A shift towards mobile systems would also mean a shift away from wires. Most proposed telepresence systems use Head-Mounted Displays (HMDs) tethered to desktop systems, which aren't only restricted by the latter's reliance on mains power but also the length of the physical connection between the two devices. This is only exacerbated when environmental scanning is required, as many systems tend to use Kinect sensors[1] which are extremely limited in the area they can reconstruct due to this same issue. Mobile phones can conversely be used wherever the user wishes: they could take a remote partner through a museum, through their garden, or even show them the view from the top of a mountain. What's more, the caller would no longer have to consider where the recipient of the call is as conversation could be initiated from almost anywhere with the knowledge that the remote peer is also in a position to receive it. The only limitation to where such a call could take place is network availability, though almost every country now has access to 4G networks in some way (WorldTimeZone, 2019), with many seeing penetration of over 80% (Opensignal, 2019).

[1] https://www.xbox.com/en-US/xbox-one/accessories/kinect

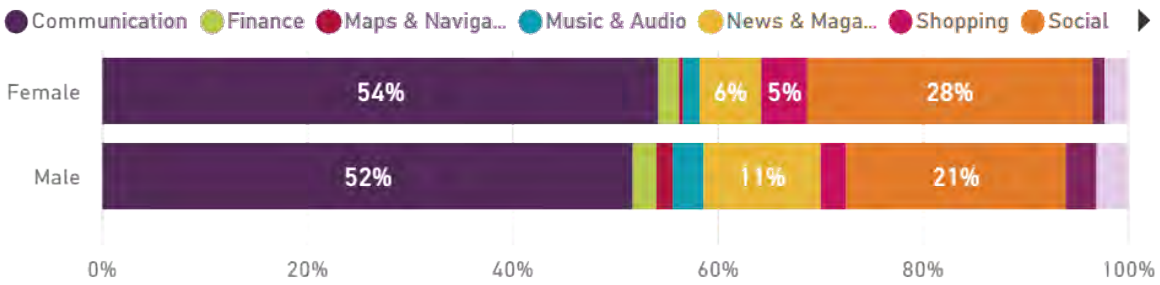


Figure 2.5: The types of apps used to access the internet according to a 2018 survey of UK citizens (Ofcom, 2018)

Arminen and Weilenmann (2009) found that this accessibility can fundamentally change how we communicate. Since calls can now be established whenever is convenient with no premeditation on the part of the receiver, "dead" moments such as during a commute can now be utilised for conversation. This is in fact what people tend to do, as 43% of adults send texts while commuting, 32% check social media, and 25% check their emails, with these numbers even higher amongst 18-43 year olds (Ofcom, 2018). These dead moments allow for seemingly insignificant interactions that would otherwise be impossible, such as letting a significant other know of your location and asking whether to pick anything up on the way home. These interactions arise almost absent-mindedly, with many calls originating from a simple desire to say hello, or even to see where the other person is, an exchange which would be entirely redundant with a stationary system. Arminen and Weilenmann argue that this "smallness" of conversation can only occur through mobile media and requires a high degree of social grounding between parties.

This mobility of course comes with its downsides; while conversation can become dynamic as the two parties move through their respective spaces, this also means that the grounding between them can be disrupted as their assumptions of the other's location is constantly disproven. This is most often seen at the beginning of conversation as callers ask "where are you?", and while this often means an extra step is required before grounding can be established, it can also provide an additional avenue of conversation.

2.3.6 Identifying Overall System Requirements

With these theories and suggestions taken into consideration, I thus identify the following requirements a telepresence system must meet in order to provide efficient collaboration and natural communication between its users:

1. The environment presented to users should remain visually and temporally consistent between them (Kraut et al., 2002; Taylor et al., 2009). Views of both the task space (Biber et al., 2005; Fussell et al., 2000) and users' faces or bodies (Flor, 1998) should be supported and freely switched between through no conscious effort on the user's part (Kuzuoka et al., 2000). The representations of task objects should make the relationships between them explicit and remain consistent throughout the interaction (Flor, 1998).

2. Both users should be able to freely explore this environment at their own discretion and completely independently from one another (Flor, 1998; Fussell et al., 2000; Kuzuoka, 1992; Kuzuoka et al., 2000; Taylor et al., 2009).

3. During this free exploration, the current position, field of view, and focus of attention of each user should be freely visible to their partner (Fussell et al., 2000; Kuzuoka et al., 2000; Tang et al., 2017; Taylor et al., 2009).

4. Hand and body gestures should be supported and freely available to both users (Biber et al., 2005; Fussell et al., 2000; Kuzuoka, 1992; Kuzuoka et al., 2000; Tang et al., 2017; Taylor et al., 2009). Both pointing gestures and representational gestures should be supported (Fussell et al., 2004; Sakong and Nam, 2006), preferably by mimicking the behaviour of the user's real hand as closely as possible (Kirk and Stanton Fraser, 2006). The interface to performing these gestures should distract from conversation as little as possible (Kirk et al., 2005) and require no additional effort on the part of the observer to be seen (Kirk and Stanton Fraser, 2006; Kuzuoka et al., 2000) and interpreted (Kirk et al., 2005).

5. Based on the research by Arminen and Weilenmann (2009) and the global trend toward mobile computing (Ofcom, 2018; Ritchie and Roser, 2019; Newzoo, 2020), the system must be fully realised on a mobile phone for maximum ubiquity, while shedding all reliance on tethered connections to power, internet, or other data lines to maximise the portability and potential use cases of the proposed solution.

2.4 Evaluation of Existing Systems

Many attempts have been made to incorporate one or more of these requirements into a telepresence system. However, most require specialised hardware or tether one or

both users to a desktop system, and despite this often fail to achieve interactive frame rates. Here we evaluate each of these attempts to determine where their faults lie.

2.4.1 View Independence in Two-Dimensional Environments

Due to the benefits it brings to remote collaboration, many systems place an emphasis on providing independent views for their users within some shared environment. Panoramas prove a popular representation for this due to the ease of which they can be produced and rendered, allowing rotational independence in large areas where translation would have no visible effect.

Chili (Jo and Hwang, 2013) allows partial view independence between users by projecting their camera images into a virtual spherical environment based on the orientation of their device. Each user can then freely obtain novel views within this environment by reorienting their device, allowing additional spatial context to be provided to their camera streams. Users can then draw over this environment using world-stabilised annotations to allow for basic pointing or representational gestures, and each can switch their view to their remote partner's front-facing camera to allow for face-to-face communication. Jo and Hwang found that this provided more spatial presence within the remote environment than traditional videoconferencing systems, with participants remarking that they felt they were there with their conversational partner. Task completion time was also found to be lower than when using a traditional system, and participants found the viewpoint control mechanism to be intuitive and preferable to verbally asking for reorientation of the camera, especially for the remote user. However, spatial presence is limited by the fact that the environment is only visible in the local user's field of view, with only an empty spherical grid visible elsewhere. Limiting interaction to drawn annotations also limits the complexity of possible gestures, which could be detrimental in complex tasks or environments where representational gestures are required.

Müller et al. (2016) improve upon Chili with *PanoVC*, which provides static context in areas previously viewed by the local user. Each frame from their camera is recorded into the environment, incrementally creating a panorama of their surroundings that can be independently viewed by the remote user by reorienting their device. Drawn annotations are supported for basic gesturing, and each user's field of view is shown to the other to ease coordination of viewpoints. They found that the addition of this static context increases the remote user's sense of spatial presence in the environment and the sense of social presence between users. However, the lack of meaningful interaction

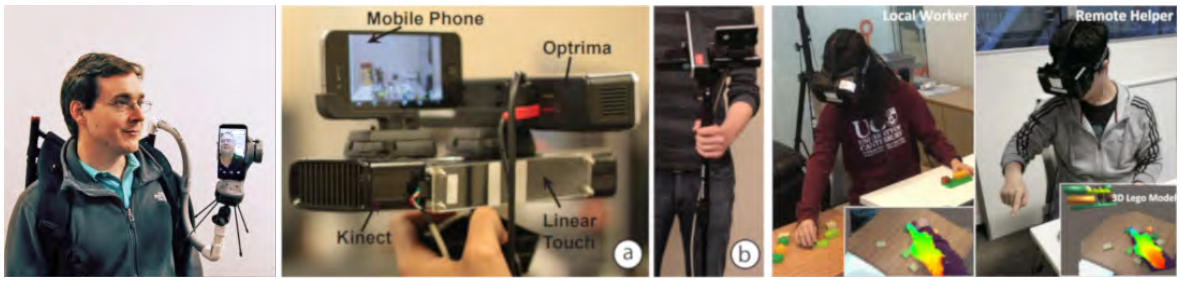


Figure 2.6: Previous systems that utilised inside-out tracking or reconstruction to provide users a shared environment in which to interact. The images are taken from their respective papers. (Left): Polly (Kratz et al., 2014). The local user attaches their mobile phone to a shoulder-mounted gimbal, allowing the remote user to control their view direction by physically rotating the device. (Centre): BeThere, which combines several sensors to create a 3D reconstruction of the local user's immediate view which interlocutors can interact within Sodhi et al. (2013). (Right): The work by Gao et al. (2016) which uses a Leap Motion controller mounted to an Oculus Rift to allow nominal view independence within a small 3D space.

between users resulted in the local user having no sense that the remote user was spatially present with them, and no significant increase in co-presence was achieved, which was attributed by the authors to a lack of visual representation of each user within the environment. The use of a cylindrical panorama also means that directly above and below the local user cannot be mapped into the environment, making it unsuitable for use in HMDs where views of this area are easily attainable. Despite these drawbacks, users expressed their preference for this type of communication over traditional videoconferencing.

PanoInserts (Pece et al., 2013) differs from PanoVC by providing static context of the whole environment from the time conferencing begins rather than relying on incremental construction. Before the video call is made, an environment with tracking markers placed around it is constructed and shared as a 360° cube map. One or more local users can then connect to the system using their mobile phones, each of which captures live video that is overlaid on the cube map with correct spatial orientation through a combination of marker-based tracking and feature-based stitching. A remote user can then view both the static context of the panorama and the live focus of the video streams, though the entire environment must be viewed all at once rather than through more intuitive first-person views. In a study involving placement of objects on a table surrounding the user, they found that PanoInserts resulted in a much lower rate of placement error than when using a traditional videoconferencing system. This is because participants often felt confused about the layout of the room as movement

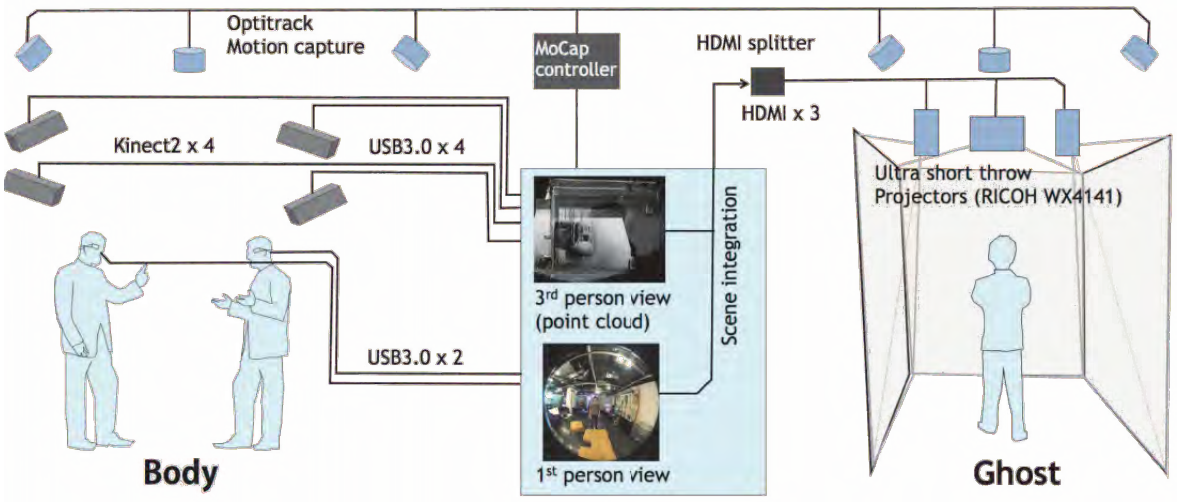


Figure 2.7: An illustration of Jackin Space (Komiyama et al., 2017), which serves as a typical example of systems which use outside-in camera placement to reconstruct the shared environment.

required verbal instruction to a local helper, and they thus often didn't know which way to turn to view different parts of it, indicating difficulty in self-localisation within the space. Despite this, users rated the traditional system as the most usable, possibly because PanoInserts is only capable of viewing the entire cubemap which may be unintuitive for inexperienced users. The authors suggested this could be remedied by allowing participants to view the space through an HMD so that they can obtain novel viewpoints by rotation of their head. PanoInserts was reported to only operate at 10 frames per second with two users despite their reliance on powerful desktop computers, a problem which would deteriorate further with the introduction of additional local clients. Interaction is also extremely limited, with only voice being exchanged between users.

Additional hardware is unfortunately unavoidable if the wider context to be independently explored is to consist of live video rather than static images. *Polly* (Kratz et al., 2014) utilises a specialised gimbal mounted on the local user's shoulder to allow for novel live views within the shared environment by giving the remote user physical control over the rotation of the local camera. Unfortunately the placement of the device on the shoulder makes gestures performed by the remote user difficult to translate to their correct position in the environment as both cannot be focused on at once, and it was shown to take significantly longer for the local user to determine the remote user's direction of attention than if the two users were physically co-present (Kratz et al., 2015). It was suggested that showing the remote user's current field of view to the

local user through an HMD could mitigate this, possibly combined with some other indicator to make it easy for this information to be known. Despite these drawbacks, Polly was still well received by participants: the local user felt that their focus on collaborative tasks was higher as mounting their phone on their shoulder meant both hands were free, and wearing the device caused them no social discomfort. Unfortunately this hardware is likely to be expensive and impractical for the average consumer, and the lack of interaction methods between users limits its use as a collaborative tool.

Tang et al. (2017) created a more consumer-friendly system that allows full rotational view independence within a live panoramic environment. The local user captures their surroundings with a 360° camera worn on a monopod attached to a backpack. Its video stream is sent to the remote user, who can obtain novel views within it by swiping or reorienting a tablet device. The remote user can also see the local user's head in the 360° video and thus infer their gaze direction, however the local user has no indication of where the remote user is looking. This often led to confusion between users, requiring complex verbal exchanges to coordinate their viewpoints when the remote user wished to dictate the direction of attention. Remote users expressed frustration at this lack of shared field of view awareness, and despite knowing of its absence still attempted to use it as a conversational resource, often reorienting their tablet to show the local user which direction to turn. This also introduced ambiguity to instructions; something as simple as "go back" could have several different meanings as it doesn't make clear whether to rotate, translate or undo a previous action. Consequently, remote users often stayed facing in the "forward" direction, nullifying the benefits of having independent views at all. The lack of meaningful interaction methods between users provided similar frustration; remote users would often point to landmarks or point in the direction the local user should face, accompanying the move with instructions such as "turn in this direction", but the lack of shared information made these deictic references meaningless.

2.4.2 View Independence in Three Dimensions

As promising as these systems are with the benefits they bring to communication, all share a common disadvantage: only rotational independence is possible, and the remote user is restricted to the position of the local one and thus cannot truly explore the environment without their intervention. Achieving this extra dimension of freedom also requires an extra dimension the environmental reconstruction, which researchers have thus far attempted in several ways.

The first is to place some surrogate within the environment and allow the remote user full control over its position. This is usually a robot equipped with cameras and other sensors, through which it is hoped the remote user will develop a sense of autonomy within the local environment. GestureMan was an example of this (Kuzuoka et al., 2000), which also attached a laser pointer to the robot to give the operator a limited ability to perform pointing gestures to aid in collaboration. Kuzuoka et al. found that though the remote user moved through the environment more than expected, this was still less than they would in a side-by-side scenario due to the delay between their intended movements and the robot's response. Three reasons were found for this movement: to guide the remote worker to specific locations, to acquire views of the current task object, and to observe the worker as they performed instructions, all of which require free viewing of both the conversational partner and the task space in order to work as intended. The laser pointer was used extensively to give instructions, many of which would be difficult or impossible without it, however this would not be sufficient to convey equally important representational gestures (Fussell et al., 2004).

The other method is to create a three-dimensional reconstruction of the space, usually before communication starts, and allow users the ability to navigate this virtual space themselves. This can be done in two ways: the first is with *outside-in* camera placement, which has one or more cameras facing towards the area to be captured. This ensures comprehensive coverage of this area that minimises the effects of occlusion, however the size of this area is limited by the placement of the cameras, and the setup required for this placement often excludes spontaneous interaction. The second is through *inside-out* capture, which uses a smaller number of cameras, often only one, facing outward from the user to the environment. This lends itself more to mobile use as the camera can be carried by the user and even attached to their display, however the placement of this camera must be more carefully considered as occlusion cannot be mitigated.

Gao et al. (2016) created a semi-mobile inside-out system by mounting depth sensors to the front of each user's desktop HMD and streaming the captured depth map and orientation to their peer. Contrary to prior research (Gauglitz et al., 2012, 2014; Jo and Hwang, 2013; Tang et al., 2017), no benefit was found to having this independent viewing method, likely due to the extremely limited space afforded by the HMD's tracking solution.

In their next iteration of the system, Gao et al. (2017) aggregate the maps captured by the HMD-mounted depth sensor into one coherent model before communication be-

gins. The remote user can thus visit previously reconstructed areas at their leisure, though again this exploration and the size of the reconstructed area is extremely restricted due to the limited range of the HMD. It was nonetheless found that objects are easier to find and identify when searched for in three dimensions, and acknowledgement of discoveries are simple to confirm as user's gaze directions are shared to their partner, though the static context afforded by the a priori reconstruction means that temporal changes cannot be seen.

Stotko et al. proposed a more mobile solution with SLAMCast (Stotko et al., 2019), which uses a handheld Kinect sensor to greatly increase the size of the reconstructed area. This solution is still extremely limited as the Kinect is tethered by a power cable, disallowing true large-scale environments. The authors proposed using a mobile phone as an alternative capture device, though this option was not explored further. Four powerful desktop computers were used to combine the captured data, each with discrete Graphics Processing Units (GPUs) too expensive for most consumers, and a desktop HMD was used as the viewing device which meant the full environment could not be explored without artificial means of movement.

The main issue with these systems is that no way of capturing collaborators' body language is accommodated as users tend to be behind the capture device. Outside-in systems solve this by capturing not only the environment but everyone inside it, allowing both to be seen simultaneously at the expense of severely limiting the size of the explorable area.

One such system is Holoportation (Fanello et al., 2016), which creates a full volumetric scan of the remote user by surrounding them with custom RGBD camera arrays. This scan is shared in real time with the local user, who can see them virtually placed in their environment via a Hololens optical see-through display. Each of these arrays consists of three cameras and a structured light emitter, the data from which is sent to four separate desktop computers. Each of these contains two discrete enthusiast-grade GPUs and an equally high-grade CPU; while the resulting reconstruction is of a very high quality, such a setup is obviously not feasible for the vast majority of the population, even without accounting for the high cost of the Hololens itself.

Park et al. (2019) propose a more affordable alternative that instead surrounds communicating parties with Kinect sensors. These are much more affordable and are commercially available, increasing consumer availability at the expense of producing lower resolution scans than Holoportation's custom camera arrays. These scans are presented in voxelised grids, isolating users from their surroundings and placing them

into a shared virtual environment displayed in an Oculus Rift. Each user may freely walk through this space with views possible from any angle due to the 360° coverage provided by the outside-in sensor placement. Despite being more affordable, this setup suffers from the limited tracking space provided by the Oculus and the fixed nature of the Kinects, severely constraining the explorable area.

2.4.3 View Independence with Mixed Dimensionality

With the advantages of inside-out and outside-in capture it seems feasible that the two could be combined in some manner to offset their disadvantages. Komiyama et al. created such a system with JackIn Space (Komiyama et al., 2017), which as seen in Figure 2.7 has the usual array of depth sensors to reconstruct a small area in 3D, but if the remote user wishes to see outside of this confined space they can transition to an egocentric view of the space as captured by a 360° camera mounted to the local user's head. Areas not covered by the depth sensors can thus be seen, however not explored as the local user is also restricted to this area due to their HMD's limited tracking space. Users appreciated the ability to perform this transition between egocentric and exocentric viewing positions, however the means to do so is a deliberate computer-mediated process and so was not as well regarded.

Teo et al. (2019) propose hand gestures as a more natural way to perform this transition. Users can freely explore a static 3D environment created beforehand through photogrammetry, and the user can transition to an egocentric view from the local user's head-mounted 360° camera by performing a "double thumbs up" gesture to see live capture with full coverage of the area. The ability to switch between these two viewing modes was seen as useful and was thus greatly preferred by users to either viewing position by itself, though the transition was too abrupt and thus caused simulator sickness in many participants. Relying on gestures to trigger this change may also be an unwise solution, especially one which could feasibly be used in conversation and trigger accidental transitions.

2.4.4 Collaborative Interaction

Other systems have placed a greater emphasis on interaction between users than how the environment is presented. Most make use of a virtual hand or other abstract gesturing tool, though previous work has shown that unmediated video of the user's hands are preferred (Kirk and Stanton Fraser, 2006).

BeThere (Sodhi et al., 2013) allows six Degrees of Freedom (DoF) movement within a remote environment through use of a specialised inside-out depth sensing array. A Kinect sensor, mounted to a tripod, is used to create a three-dimensional point cloud of the shared environment. This is sent to the remote user, who can then navigate by reorienting their own tripod which has its position tracked by a similar Kinect setup. Users view this environment through a mobile tablet, which is placed on top of the tripod to align the view of its camera with that of the Kinect. An additional depth sensor is mounted to the side of the tripod which captures any hand gestures the remote user may wish to make. This captured gesture is shown as a virtual model within the space so that gestures can be shown with the correct three-dimensional context, and providing this surrogate ensures that the occluded side of the hand is still visible. Users found this system useful, noting the ability to look around objects and obtain views of the task space independently of the local user. The short range of the Kinect sensor proved the limiting factor in this regard as it only allowed for limited translational independence within the space, and as the environment was not incrementally created the explorable area remained extremely small. Users found the hand model intuitive to use as it matched their own hand movements closely, and it permitted more complex representational gestures than an annotation-based system would allow. Despite its usefulness, the setup required for this system would be impractical for most users due to its cost. The system also proved too heavy to hold for extended periods of time, necessitating the tripod and limiting its mobility, which further highlighted the short range of the Kinect. Additionally, the number of devices required necessitated offloading computation to a desktop computer, so despite claims of this being a mobile system this could not be further from the case.

JackIn (Kasahara and Rekimoto, 2014) provides a similar experience to BeThere but with more portability. The local user wears a transparent HMD with an in-built camera. Its video feed is sent to the remote user's device, which performs Simultaneous Localisation and Mapping (SLAM) on the incoming image to incrementally build a panoramic model of the shared environment. The remote user can then independently browse this environment from a first-person perspective by performing specific gestures towards their device's display. These gestures are captured by a Leap Motion controller[2], which also allows the user to highlight objects of interest with a cursor by simply pointing at them with a finger. Allowing the user independent views within the environment was shown to aid in their spatial understanding of it, and users often

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

Figure 2.8: Several ways in which gestures could be incorporated into remote communications. (Left): From Kasahara and Rekimoto (2014). A primitive cursor or other annotation is used to facilitate pointing and basic representational gestures. (Centre): Raw capture of the user's hand is overlaid on the environment to provide the full range of gestures. (Right): From Sodhi et al. (2013). The user's hand capture is extrapolated to create a surrogate model through which gestures can be performed in 3D.

felt more confident that they were performing collaborative tasks correctly as a result. They were also often observed focusing on different areas of the task space than the local user, allowing them to dictate the direction of attention and work independently from the local user. The incorporation of gestures also led to more efficient collaboration due to the introduction of deictic references, though the lack of support for representational gestures limits the ways in which these can be used. Issues were also caused by the Leap controller as there was a discrepancy in real and virtual pointing locations due to incorrect calibration, and even if the remote setup were made portable the use of this sensor would make it unsuitable for outdoor scenarios due to its reliance on infrared light.

Gauglitz et al. (Gauglitz et al., 2012, 2014) developed a slightly more robust system that provides more freedom for the local user. They use a mobile tablet's camera to capture their environment and perform SLAM tracking to determine their orientation within it. These frames and their associated orientation are then sent to the remote user's desktop device which constructs a 3D model of the space. They can then independently view this environment using their mouse, and can place world-stabilised annotations within it to communicate with the local user. If these annotations are outside of their current field of view, an arrow is displayed at the edge of their device's display to show its relative location. These annotations were found to greatly improve task performance, allowing for more tasks to be completed in the time allotted to each participant. Users greatly preferred this system over a traditional video conferencing system and rated the annotations as "extremely helpful", so much so that verbal communication became inefficient in comparison. The use of a mobile device for the

local user is also beneficial as it allows them to freely explore a possibly remote task space, however limiting the remote user to desktop use potentially speaks against a spontaneous use case as they would have to be expecting the incoming call.

Other systems provide interaction through unmediated video of the user's hands, though these also have their limitations. *HandsInAir* (Huang and Alem, 2013) shows this unmediated video to the local user through a transparent HMD, which has a camera attached to the front of it so that the remote user can see the shared task space. This allows for gestures to be seen directly in the environment, though the remote user's view is dependent on the direction of the local user's camera, so unexpected head movements may result in these gestures being shown with the incorrect spatial context as they are not world-stabilised.

2.5 Summary of the Literature

Conversation takes place through a process of grounding, where information is pushed and pulled across various communications media until common knowledge between interlocutors can be assumed. The amount of information that can be pushed and pulled is limited by the degree of presence felt during communication, whether that be the spatial presence within the shared environment or the social and co-presence felt between conversational partners. To provide an efficient communications platform, a system must therefore aim to maximise the amount of presence it can induce in order to maximise the efficiency and naturalness of the conversation that can take place.

Two ways of doing this exist which have proven effective in the past. The first is to allow distant users to not only see their partner's environment through their fixed video stream, but to allow them to virtually step in to that environment and freely explore it themselves. This then increases their sense of spatial presence within that shared space, and information can freely be pulled from it without intervention from the user physically present there.

The second is to support gestures as a way to enhance verbal statements. These can provide subtle cues that can aid in establishing common ground, and can also serve as instructional aids during collaborative tasks where completion time must be minimised. Both pointing and representational gestures must be supported, and unmediated video is preferred to gesture surrogates as it allows both these types of gestures to be shared effortlessly.

Communication of any kind is preferred when it can be done through mobile means,

allowing it to fill in otherwise "dead" moments in a person's day. This opens up the kinds of conversations that can be had, providing opportunities for socialisation that would otherwise be impossible using traditional desktop computers. More people now own a mobile phone than a desktop or laptop computer combined, so a shifted research focus to mobile computing would instantly allow more people to benefit from that research through a platform they prefer to use for communication, making them more likely to adopt new proposed technologies.

To the best of my knowledge no systems exist that provide independent views and gestural interaction within a consistent, shared, and mobile environment. Many provide free exploration of the environment, though often limit the explorable area by confining users to small tracking spaces, require expensive proprietary hardware, fail to provide rich interactions within the shared space, or all of the above. Gestures have been explored, though either fail when the camera is allowed to freely move or are performed through unnatural surrogates that limit the range of gestures that can be conveyed. Finally, almost all of these systems focus on desktop development, severely limiting who can use them, where they can be used, and how they can be used, potentially contributing to a lack of interest from the wider public.

Chapter 3

A Foundation for Mobile Telepresence

Due to the lack of existing mobile solutions that satisfy the requirements for rich collaboration identified in the previous chapter, in this chapter I present a framework for mobile telepresence that allows users a higher degree of spatial presence within a shared real-world panoramic environment than can be provided using traditional videoconferencing systems.

Users of this application may obtain independent views within this space by simply reorienting their device, granting the remote user full rotational independence from the direction of the local user's camera. This independence provides more opportunities for the remote user to pull information from this space without intervention from the local user, making conversation more efficient Flor (1998) and increasing the sense of presence within that space (Jo and Hwang, 2013; Müller et al., 2016). To allow for viewpoints to still be easily coordinated the Field of View (FoV) of each user is shown, and either user may further immerse themselves within the space by viewing the application through a mobile HMD such as the Google Daydream.

I also aim to increase the sense of co-presence between users of this system by showing their unmediated hand gestures correctly aligned within the environment, allowing for natural conversational cues such as pointing and representational gestures. The use of ubiquitous mobile hardware ensures that this system may be used by anyone regard-

less of physical location or lack of premeditation, and its networking implementation allows use from anywhere with sufficient internet or mobile network coverage.

As the computational capabilities of mobile phones are limited, five different approaches to creating this shared environment are explored that between them sample the full continuum of view independence possible in such a panoramic space. The implementation of each is detailed along with an extensive evaluation of their performance on a modern mobile device, allowing them to be used in future research. The results of a preliminary user study are also reported in the following chapter, which confirms this system's ability to induce the desired sense of spatial presence within the environment and co-presence between users while identifying the degree of view independence required to do so. These contributions will guide future development of mobile telepresence systems and prove them to be a valuable and capable platform for future telepresence research.

The contents of this and the following chapter were published in the IEEE Transactions on Visualization and Computer Graphics[1] and presented at IEEE Virtual Reality 2019[2] where it was nominated for Best Journal Paper (Young et al., 2019).

3.1 Implementing a Mobile Framework

This framework was developed for the Android operating system and is compatible with any supported device, though mobile phones were the focus of testing to fully capitalise on their ubiquity. A first-generation Google Pixel was used for all testing and development, which was released in 2016 and so ensures that any relatively modern phone will be capable of running the application.

The *local user* captures a panoramic representation of their surroundings using either their mobile phone's inbuilt camera or an external 360° one. This reconstructed environment is shared with the *remote user*, who can obtain independent views within it by simply reorienting their device. The two users may then communicate through voice, gestures and shared FoV awareness, allowing for natural, intuitive conversation that closely mimics that used in side-by-side scenarios. An overview of the hardware used and the information exchanged between clients is shown in Figure 3.1.

The system supports several methods to construct the shared environment, called the *modes of interaction*. These are presented in order of increasing view independence:

[1] https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?reload=true&punumber=2945
[2] http://ieeevr.org/2019/

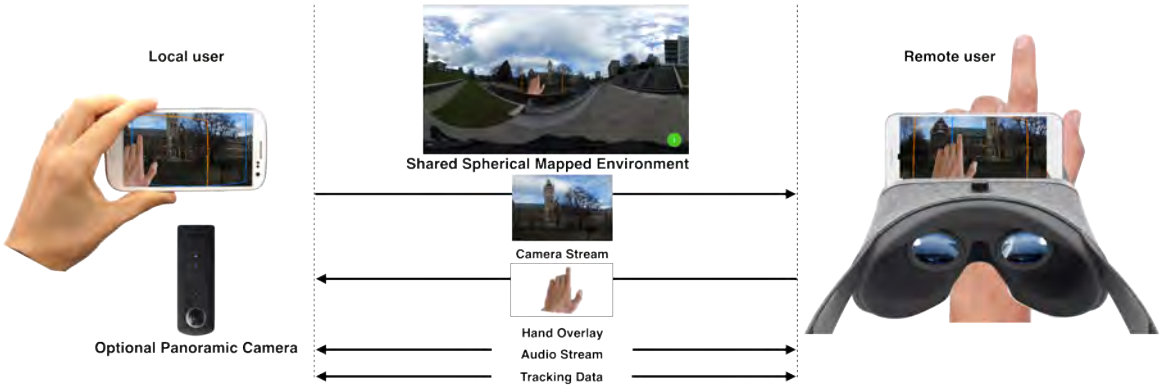


Figure 3.1: An overview of the presented mobile telepresence framework. The local user streams their device's orientation information and live video of their surroundings to a remote user, who uses this data to reconstruct a panoramic representation of it in one of several ways. Users may communicate through hand gestures, which are captured and spatially mapped into the environment to preserve their context, or through voice, allowing for conversation as if the two parties were side-by-side. Either user may also immerse themselves further within the space by viewing it through a mobile HMD.

- *Live Video Calling*: The local user's live camera feed is shown directly to the remote user, who cannot manipulate their viewpoint and thus is restricted to only viewing the current video stream. This gives a similar experience to traditional video-calling applications such as Skype[3], but the remote user's hands are overlaid on the video feed to allow for gestural interaction. No view independence at all is provided due to the users' views being completely coupled. Live Video Calling was included as a baseline to compare the more complex modes against, though could also prove useful in situations where the local user wishes to completely dictate the direction of attention.

- *Live Spatial Video Calling*: Both users are placed at the centre of a virtual sphere which they perceive as a monochrome grid. The local user's camera images are projected to the inside surface of this sphere based on the orientation of their device, and so the direction each frame is projected matches the direction in which it was captured in the real world, thus preserving spatial relationships between objects in the environment and giving this rotational context to their camera stream. The remote user can then control their viewing direction and subsequently where their hand gestures are shown by reorienting their device, decoupling their view from the direction of the local user's camera. Partial view

[3]https://www.skype.com/en/

independence is thus provided, however only the spherical grid will be visible outside of the local user's current FoV. The remote user's view is therefore still mostly dictated to them, but as the local user can see their FoV too this indicator can be used for spatial queries and partial dictation of attention. This mode takes inspiration from Chili (Jo and Hwang, 2013) but provides additional interactivity between users.

- *Incremental Panoramic Calling*: Operates similarly to Live Spatial Video Calling, but each time an image from the local user's camera is projected onto the inside surface of the sphere it is permanently recorded there. Over time this creates a static panorama of the environment, increasing view independence by allowing the remote user to view previously visited areas at their leisure while also providing live focus within the local user's FoV. However, the local user must have already viewed an area before it becomes visible, allowing them to dictate where up-to-date views can be seen and thus where independent viewing can occur.

- *Panoramic Calling with Live Inserts*: A full panorama of the environment is captured and shared before communication begins, then projected onto the full surface of the sphere once a call is initiated. This allows the remote user almost full rotational view independence as they can view any area without relying on the local user visiting it first. Since panorama construction is performed offline slower, more accurate techniques may be used than in the previous modes, however this increase in quality introduces a heavy penalty to spontaneity of use by introducing a required degree of premeditation and temporal changes will never be reflected. The local user's camera stream is still projected into this static environment, providing a live focus area, though only where they dictate it should be.

- *Live Panoramic Video Calling*: Rather than reconstructing the environment from frames captured by the mobile phone's internal camera, an external panoramic camera is used to capture the entire environment in real time. Full rotational view independence between users is thus achieved as the remote user can see live video wherever they may look, however this comes at the cost of ubiquity, cost of entry, and spontaneity of use as it requires additional hardware.

An example of each mode in use is shown in Figure 3.3. Each allows the environment to be viewed either through the mobile phone's screen or more immersively through a

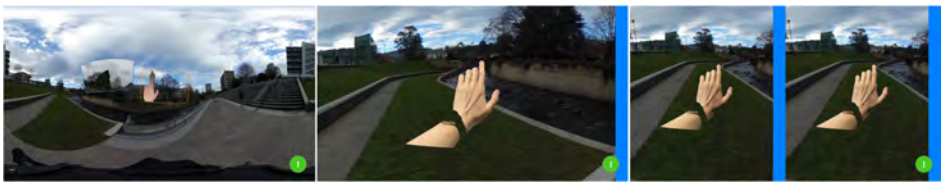


Figure 3.2: The possible ways to view the shared environment. (Left): An equirectangular projection of the entire panoramic space. (Centre): A first-person view of the environment based on the orientation of the user's device. (Right): A pseudo-stereoscopic projection for viewing in a mobile HMD.

mobile HMD. Each produces a first-person view based on users' orientations, however an equirectangular projection of the full environment sphere may also be viewed, as shown in the left of Figure 3.2, if either user wishes to see the entire environment at once. For ease of reference all modes other than Live Video Calling will be referred to as the *spatial modes* throughout the rest of this work.

To allow for coordination of viewpoints when users' views are decoupled, the current gaze direction of each is displayed as a coloured outline projected onto the environment around each user's FoV. If users are looking in different directions and this outline is not visible, one edge of the screen is instead coloured to indicate which direction the user would need to turn in order to see their partner's FoV indicator. This allows for context to be preserved in spatially sensitive utterances such as "look over here" or "what's this?" without any deliberate action required.

Gestures have been shown to be an integral part of everyday conversation (Fussell et al., 2004; Gauglitz et al., 2012), and so the framework allows users to incorporate their hands into conversation in an intuitive way that does not interfere with the construction or viewing of the environment. For the local user their hands will be visible in their existing camera stream, however for the remote user their hands are first segmented from the background before projecting their camera images into the environment. The remote user can thus perform gestures such as pointing without having their view obstructed by their own local environment. Verbal communication is also supported through inbuilt VoIP capabilities.

3.2 Technical Foundation

All modes of interaction require that images be captured and transmitted to the remote peer along with the device's latest orientation information. To reflect this and allow run-time selection of which mode to use, all modes share the same low-level subsystem

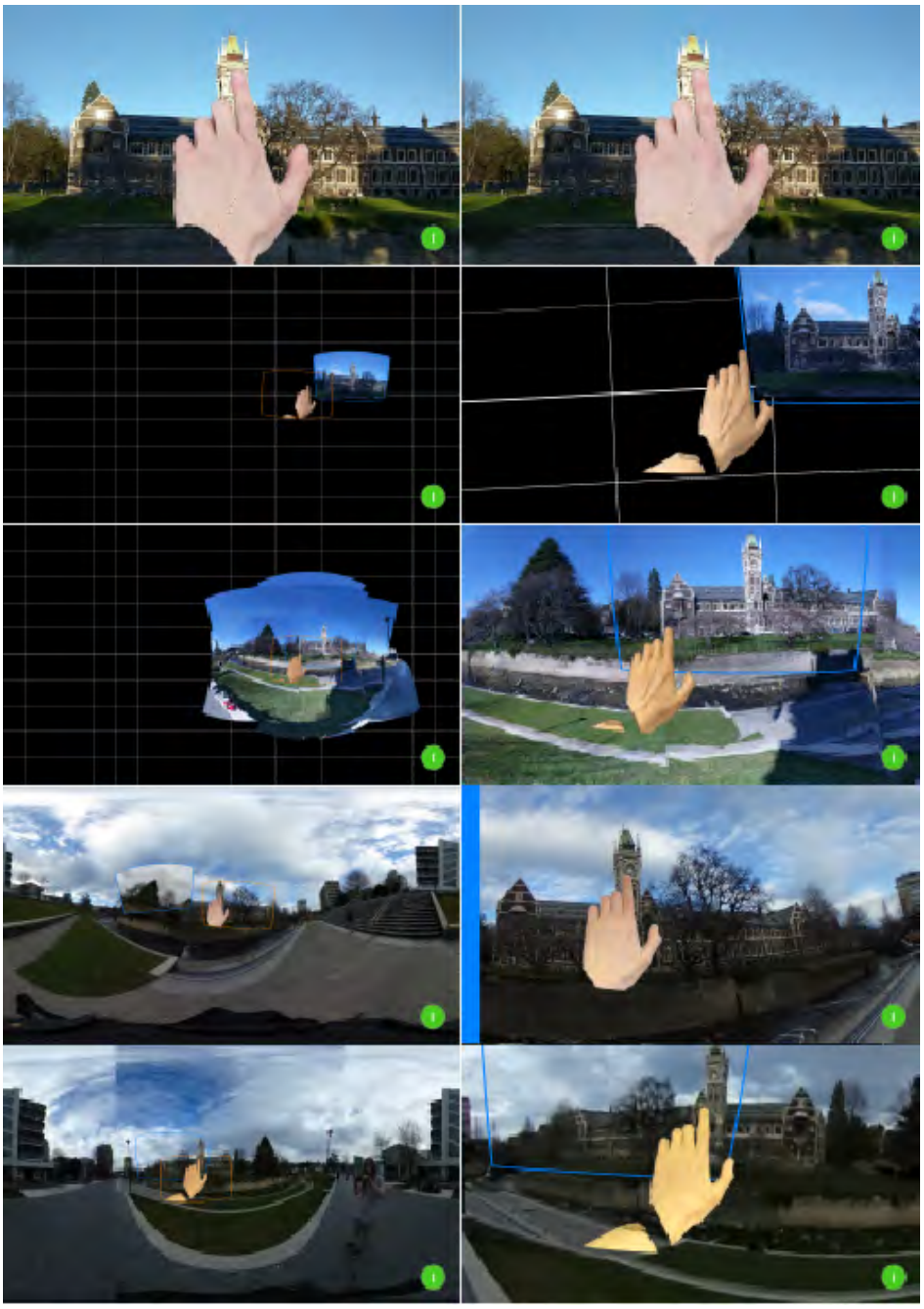


Figure 3.3: Each mode of interaction in use, with the full panoramic environment on the left and the unprojected first-person view on the right. In order of increasing view independence from top to bottom, these are: Live Video Calling, Live Spatial Video Calling, Incremental Panoramic Calling, Panoramic Calling with Live Inserts, Live Panoramic Video Calling. The top row shows a view of the entire panorama constructed using that mode, whereas the bottom row shows a unique view of this environment for the remote user based on the orientation of their device.

which will be described in the following sections. The interactions between these various subsystems are illustrated in Figure 3.4. To maintain real-time performance all rendering and processing of images is done in C++ whenever possible, requiring them to first be passed through the Java-Native Interface (JNI) as they can only be captured via the Java API.

3.2.1 Camera Access

Access to the mobile device's camera is provided by Google's Camera2 API[4]. This grants access to the raw camera stream as well as parameters such as exposure, focus, and white balance at run-time, which are essential for building environments with constant illumination and achieving accurate segmentation of the users' hands in any lighting conditions.

Images are captured at a resolution of 1280×720 in all modes. Higher resolutions are possible but would require greater network bandwidth to achieve acceptable end-to-end latency. Each frame is recorded to two separate buffers; one is CPU-controlled and is for the networking module to send to the remote peer, and the other is a GPU-controlled OpenGL texture that will be displayed locally during rendering. Frames are captured in NV12 format and so must be converted to I420 for the networking module and RGBA for rendering; these conversions are performed using OpenCV[5].

3.2.2 Orientation Estimation

To provide independent views for each user, some way of tracking their device's orientation must be implemented. Existing solutions such as ARCore[6] or ARKit[7] were considered but dismissed as their limited availability at the time would have compromised the convenience of the application and provide an unnecessary barrier to its use.

Two orientation methods have thus been developed that can be used interchangeably at run-time. Both produce a three-dimensional rotation matrix, which is passed to the renderer so that a correct view of the environment may be calculated. This matrix is also encoded within each camera frame before being sent to the remote peer so that

[4]https://developer.android.com/reference/android/hardware/camera2/package-summary
[5]https://opencv.org
[6]https://developers.google.com.ar/
[7]https://developer.apple.com/arkit/

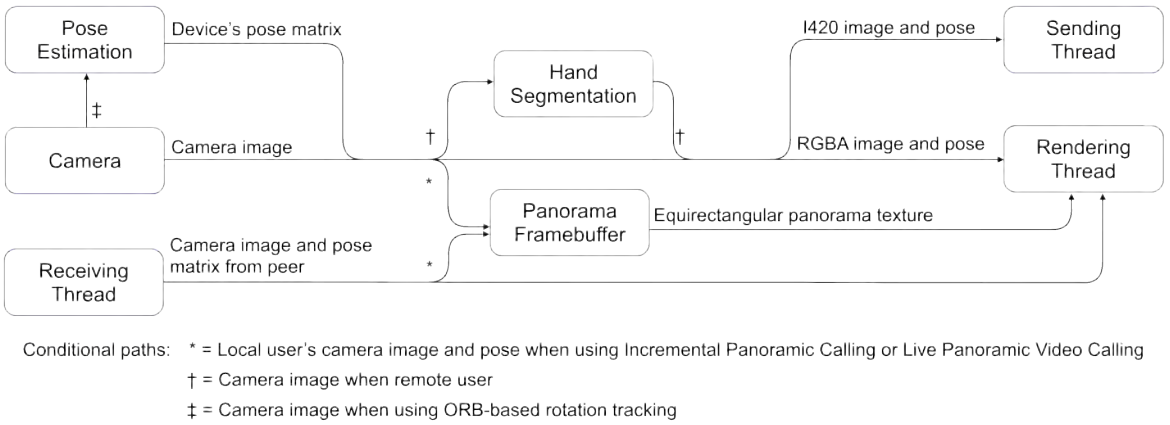


Figure 3.4: The low-level subsystems shared by each mode and the interactions between them. Paths marked $*$, † or ‡ are only conditionally followed.

it may be correctly positioned within the environment. Movement is restricted to rotation as the panoramic environment is position-invariant; translational movement would require depth-sensing hardware or real-time dense SLAM to ensure the environment reacts appropriately to the user's actions and so is omitted for now.

The first and fastest method is to simply use a fusion of the device's inbuilt sensor values after processing them with a Kalman filter; for this I use an implementation by Pacha (2013). This approach can estimate the user's orientation in real-time but is susceptible to drift inherent in these sensors. For most modes this drift is acceptable as the orientation calculated does not have to perfectly align with reality, however in Incremental Panoramic Calling and Panoramic Calling with Live Inserts these discrepancies can cause visible seams in the constructed panorama or overlaid live video; for this reason a more complex method has been implemented that uses image feature detection to calculate the user's orientation with higher precision at the cost of real-time calculation.

The absolute orientation R_{a-1} of the device during the first frame of tracking is assumed to be that calculated by the sensor fusion approach. The set of feature points in the first camera image f_{t-1} is also found using the ORB method (Rublee et al., 2011), which combines FAST keypoint detection and BRIEF descriptors. For each subsequent image its feature set f_t is also calculated using this same method. All matches between it and the one before (f_{t-1}) are found using a brute force approach, comparing each point in f_t and f_{t-1} and assigning a similarity score to each pair based on the immediate neighbourhood of their keypoints. For each point the pair with the highest similarity score is kept and added to the set m if this score is above 70% of the

highest possible and is at least double that of the next best matching pair.

After finding m, a homography-based estimator is used to find the relative rotation R_r between the two images which is refined through bundle adjustment. This estimated rotation is checked against the rotation R_s obtained by the sensor fusion estimator by calculating their difference

$$D = R_r R_s^{-1}. \tag{3.1}$$

The frame is discarded if a discrepancy of more than 5° is found due to the difficulty of recovering from erroneous estimates. Once a suitable relative rotation is found between the two frames the new frame's absolute rotation R_a is calculated such that

$$R_a = R_{a-1} R_r \tag{3.2}$$

which is returned as the device's orientation matrix R.

3.2.3 Networking and Synchronization

To facilitate the connection between clients I use Google's open-source implementation of WebRTC[8], an API that provides efficient matchmaking and streaming of audio and video. A central server is used for initial matching of clients but subsequent communication is entirely peer-to-peer. It is assumed for now that neither client will be behind a strict NAT or firewall and thus no support for STUN or TURN servers, which aid in bypassing these restrictions, was implemented for this framework.

When a connection is established between two users, three communication channels are opened for each: one for OPUS-encoded audio, one for VP8-encoded video, and one for raw data streams, which is used for sending the user's orientation matrix each frame. Since no means of synchronising video and data packets is provided by WebRTC, an identification tag is embedded into two 8×8 pixel regions in the top left corner of each image as in Figure 3.5; multiple pixels are required to ensure this value can be reliably read after video compression. The pixels removed to accommodate this tag are stored in the relevant data packet along with a matching identification tag and that image's orientation matrix. When a client receives an frame over the video channel it waits for the accompanying data packet to arrive and reconstructs the image before passing it to the rendering module.

[8] https://webrtc.org

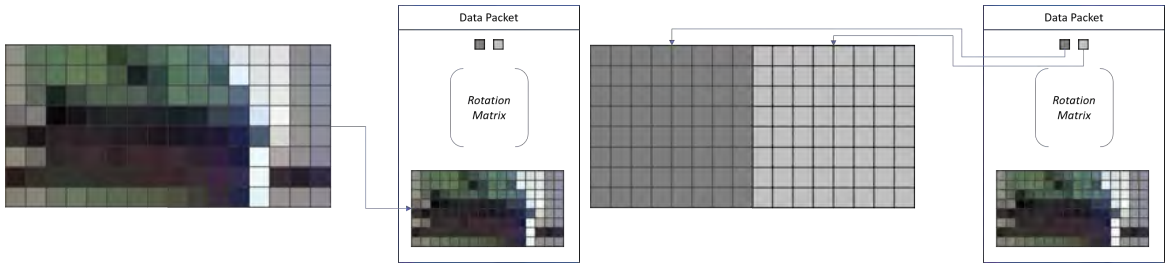


Figure 3.5: How the data packet's identification is embedded in its corresponding image. (Left): Two 8x8 regions of pixels in the top-left corner of the image are removed and stored in the orientation packet. (Right): These pixels are replaced with the packet's two-byte id tag. When received, this process allows these disparate packets to be synchronised and combined.

3.3 Constructing the Shared Environment

With low-level functionality implemented we can now focus on how to construct the shared environment. As the interactions afforded by each mode of interaction are largely the same and differ only in the degree of independence they grant users, each mode is built upon the same foundation which I describe here.

3.3.1 Representation of the Shared Environment

Though the spatial modes differ in how their environment is created, all provide some spherical panoramic space in which users can interact. For ease of implementation and to better facilitate switching between modes at run-time, all modes share a common representation of how they store this: as an equirectangularly-projected panorama that is updated differently depending on the mode in use.

Due to the technical limitations of mobile devices this panorama is limited to strictly two dimensions; more complex representations such as a point cloud or mesh are possible but would inhibit performance with minimal benefit due to the inability for translational movement. A common approach to constructing such a panorama is to model it as a cylinder (Agarwala et al., 2005; Müller et al., 2016), however this does not lend itself to viewing in an HMD as the vertical poles are not preserved. A spherical model is thus used instead which allows these areas to be visible.

An equirectangular projection of the environment sphere is stored in a texture-backed OpenGL framebuffer. The maximum size of any side of such a buffer is 4096 texels and so this is used as the panorama's width, and the equirectangular model requires a 2:1 aspect ratio, giving a final resolution of 4096×2048 for the entire panorama.

Figure 3.6: An illustration of how the environment is stored in memory. (Left): An equirectangular projection of a 360° environment stored as a flat 2D panorama. (Right): This panorama once projected to the surface of a sphere. Each user is virtually placed within the centre of this, and rotating their device allows them to gain independent views of its inner surface.

Higher resolutions are possible but would require storing the panorama across multiple framebuffers, greatly increasing performance overhead while only providing minimal gains in quality. All operations using this framebuffer are performed using fragment shaders accessed via C++ to maximise performance. This panorama texture, and its subsequent visualisation as a user-encompassing sphere, are shown in Figure 3.6.

Since no translational movement of the device is recorded it is assumed that all rotations are performed around the camera's optical centre. This requires unnatural movement from users as they must pivot themselves around the device rather than the usual inverse, and thus it is likely that discrepancies between perceived and actual rotations will occur. This discrepancy will be negligible for sufficiently distant objects (Diverdi et al., 2008), so this sphere is conceptualised as being infinitely large such that all points on the panorama are at an infinite distance to the camera, making the environment effectively position-insensitive.

3.3.2 View Unprojection

With the environment stored, some method of viewing it must now be implemented. In Live Video Calling this is simple as the local user's camera images can be rendered directly to the screen, however the equirectangular projection used for the spatial modes means the environment must first be unprojected from its equirectangular form before being viewable in any meaningful way. This unprojection must take the orientation of the user's device into consideration in order to allow them to independently control

their view within the space.

This unprojection process is performed within the rendering shader. A virtual camera is created with an 82° FoV, an intrinsic matrix K which matches that of their phone's integrated camera, and an orientation matrix R retrieved from the either estimator. For each fragment in the screen buffer we cast a unit vector $\mathbf{m}' = (m_x, m_y, m_z)$ from the centre of the environment sphere $\mathbf{M}$ such that

$$\mathbf{m}' = K^T R^T \mathbf{M} \tag{3.3}$$

which we normalise to obtain the coordinate $\mathbf{m} = \frac{\mathbf{m}'}{m'_z}$ of the texel in the panorama buffer to display at the current fragment.

For a more immersive experience, either user may alternatively choose to view the environment through a mobile HMD. The lack of depth data makes stereoscopic viewing impractical, so a vertical slice containing the centre of the unprojected image is simply shown to each eye. Most objects are too distant for this to be noticeable, especially when outdoors, and a lack of depth information has previously been shown to have no detrimental effect on performance of collaborative tasks (Kratz and Ferreira, 2016).

Alternatively, if the user wishes to see the entire scene they may instead view the full equirectangular panorama directly, which is achieved by simply copying the contents of the panorama buffer to the screen buffer. This gives an unnatural and unintuitive view of the space, but may be useful if the user wishes to quickly find their communication partner or a specific object without having to manually search for them.

3.3.3 Projecting Images into Panorama Space

For each user's latest camera image to be visible during unprojection they must first be projected into the environment based on their associated orientation matrix.

To perform this projection, for each fragment in the screen buffer we first transform its corresponding texture coordinates $\mathbf{t} = (t_x, t_y)$, obtained via the unprojection algorithm, to the equirectangular sphere map coordinates (θ, ϕ) such that

$$(\theta, \phi) = \left(2\pi \left(t_x - \tfrac{1}{2}\right),\ \ \pi t_y\right) \tag{3.4}$$

where t_x, $t_y \in [0, 1]$ and $\theta \in [\text{-}\pi, \pi]$, $\phi \in \left[\frac{\text{-}\pi}{2}, \frac{\pi}{2}\right]$ are the azimuth and inclination on the environment sphere respectively. These coordinates are then projected to a unit vector

$\mathbf{u} = (u_x, u_y, u_z)$ on this sphere such that

$$\mathbf{u} = \begin{bmatrix} \sin(\phi)\sin(\theta) \\ \cos(\phi) \\ \sin(\phi)\cos(\theta) \end{bmatrix} \tag{3.5}$$

This sphere-space projection is then rotated based on the current frame's orientation matrix R, followed by a projection to camera space using the camera's intrinsic matrix K, giving the camera-space coordinate $\mathbf{u}'$ for each fragment such that

$$\mathbf{u}' = KR^T\mathbf{u} \tag{3.6}$$

The current fragment is not projected to if the calculated values of $\mathbf{u}'$ are less than 0 to prevent the image from being shown in both the forward and back projection. The calculated coordinates are then normalised across the camera's pixel coordinate space, giving the final pixel coordinates $\mathbf{v} = (v_x, v_y)$ to sample from such that

$$\mathbf{v} = \left(\frac{u'_x}{r_x u'_z}, \frac{u'_y}{r_y u'_z} \right) \tag{3.7}$$

where $(r_x, r_y) = (1280, 720)$ is the resolution of the camera. If both $v_x, v_y \in [0, 1)$ then the pixel at v in the input image is displayed in or recorded to the current fragment.

3.3.4 Field of View Awareness

To ensure that each user knows where the other is looking at all times, a coloured outline is drawn around their current FoV. During the projection step each fragment is coloured blue (for the local user's FoV) or orange (for the remote user's FoV) if there exists

$$v \in \mathbf{v},\ n \in \{0, 1\} : |v - n| \leq \epsilon$$

where $\mathbf{v}$ is the pixel coordinate calculated in Equation 3.7 at the end of the projection step, in essence drawing a box around everything being rendered to the remote peer's screen. In cases where this indicator would not be visible due to the users' views not overlapping, a fragment is also coloured if there exists

$$c_i \in \mathbf{c},\ v_i \in \mathbf{v},\ n \in \{0, 1\} : |c_i - n| \leq \epsilon,\ v_i \notin [0, 1)\,,\ \operatorname{sgn}(c_i) = \operatorname{sgn}(v_i)$$

for the fragment's screen-space coordinate $\mathbf{c}$, providing an indicator as to which direction the user needs to turn in order to see their partner's gaze indicator. Both of these indicators can be seen in Figure 3.7.

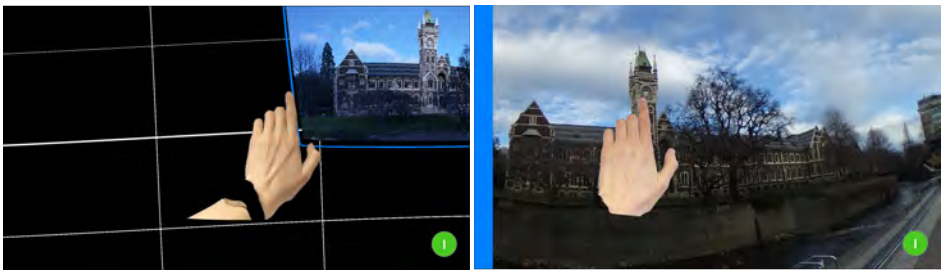


Figure 3.7: The indicator used to show each user's current field of view. (Left): When the two users' views overlap, a coloured box is drawn around the partner's FoV to show where they are currently looking. (Right): When the two users' views do not overlap, one edge of the screen will be coloured to show which direction the user needs to turn in order to see their partner's view direction.

3.3.5 Hand Segmentation

For intuitive hand-based gesturing to be possible, both users must be able to gesture and have these be visible and presented with the same spatial context in which they were originally made. Representational gestures must be supported (Fussell et al., 2004), as discussed in subsection 2.3.4, and the means with which this is done must support HMD use which excludes interactions such as touching the screen. This is simple in most cases for the local user as their hands will be visible in their pre-existing video stream, however more work is required to facilitate the same functionality for the remote user.

To ensure that their gestures are shown and that the targets of these are not obscured by other objects in their camera stream, the remote user's hands must be identified within each image, isolated from the background, and projected into the panorama within their current FoV using the method in subsection 3.3.3. This segmentation process must be fast enough to allow real-time interaction, and must be robust enough to be independent of lighting conditions and skin colour. Because of these requirements and the limitations of mobile processing, a simple colour-based approach is used.

Identification of hands within the remote user's camera image is performed on a per-pixel basis. A pixel is assumed to not belong to a hand if its YUV and corresponding RGB values satisfy the following conditions proposed by Al-Tairi et al. (2014):

$$u \in (80, 130)\,,\ v \in (136, 200)\,,\ r > 80,\ g > 30,\ b > 15,\ |r - g| > 15.$$

Compatibility with a range of skin colours is ensured by ignoring the Y channel because human skin tones generally share a similar hue and only vary in lightness (Yang and Waibel, 1996). Using this classification scheme, a binary segmentation mask is created

which marks foreground (hand) pixels as black and background (non-hand) pixels as white.

This gives a rough segmentation without much processing required, however any skin-coloured non-hand objects will still be included in the output. To mitigate this, one or more filtering techniques can be applied to refine the results. Each of these can be enabled or disabled at run-time to adapt to the current conditions, and are applied in the order presented:

- Normalised box, Gaussian, or Median filtering can be applied to the segmentation mask to remove small noisy regions. The size of the kernel used can be adjusted depending on the user's needs, with larger kernels removing larger objects at the expense of higher processing requirements.

- The mask can be eroded and/or dilated to quickly remove very small regions of noise and fill any holes present in the foreground.

- Pixels can be removed if their Euclidean distance to the nearest non-skin pixel is below some small, adjustable threshold. This removes small regions of noise at the expense of widening any holes in the foreground and eroding the edges of the user's hands.

- The GrabCut algorithm (Rother et al., 2004) can be applied, with the previously black mask values marked as probable foreground and white values as probable background. This provides quite accurate results but is too computationally expensive to perform in real time on a mobile device.

Once the final mask has been created it is applied to the image via a bitwise AND operation. As WebRTC requires images to be transmitted in I420 format which provides no means of storing transparency, any pixels to be removed are marked black, all of which will be ignored during rendering. Despite a focus on real-time processing, accurate segmentation can still be achieved on uncontrolled backgrounds, as can be seen in Figure 3.8.

3.4 Modes of Interaction

With these processes implemented they must only be combined in order to realise the various modes of interaction. Each makes use of the previously described subsystems

Figure 3.8: An example of the real-time skin segmentation the application can achieve. (Left): The original image. (Centre): The same image after colour-based segmentation. (Right): The image after distance-based thresholding on the segmentation mask.

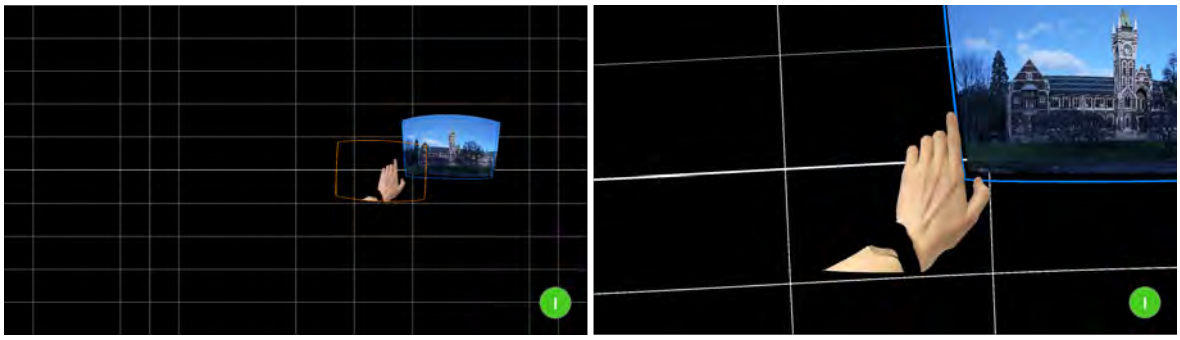
in different ways in order to achieve their desired reconstruction of the environment, though usually only differing in how new images are projected to the panorama.

3.4.1 Live Video Calling

Live Video Calling is the most straightforward mode to implement as it does not require orientation tracking nor projection to or unprojection from the panorama buffer. As frames arrive from the user's camera they are simply sent straight to their communication partner, and no data channel is required so the synchronisation process is skipped. In the case of the remote user, this image first passes through the hand segmentation module to remove the background. The two images are then rendered directly to the screen buffer, with each fragment sampling from the remote user's image if the relevant pixel is non-black and sampling from the local user's image otherwise, resulting in the remote user's gestures always being visible though difficult to correctly perform as the local user controls where they are shown.

3.4.2 Live Spatial Video Calling

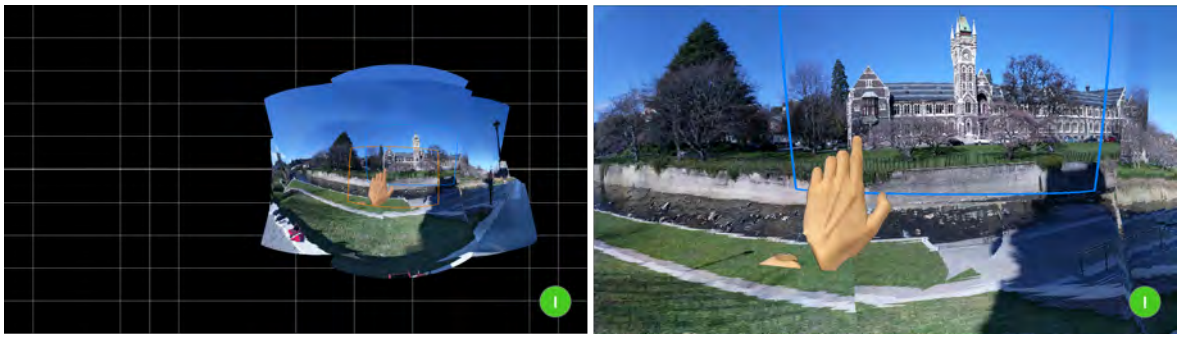
Live Spatial Video Calling is more complex as we now have to provide users unique views and spatially map their camera images based on their devices' orientations.



For each fragment in the render shader, the unprojection step outlined in subsection 3.3.2 is performed using the user's latest orientation matrix to determine which texel in the environment panorama to sample from. Both users' latest camera images are then projected into the environment, after segmentation in the case of the remote user, and if either image or their resulting FoV indicators intersect with the chosen texel then these are rendered to the current fragment. Preference is given to the remote user's indicator and image to ensure their gestures are always visible.

Users will only see their own FoV indicator when viewing the entire panorama as it is not rendered as part of the unprojection process. If neither image is visible in the relevant panorama texel then the spherical grid pattern will instead be shown in their place to give users a consistent sense of orientation in otherwise empty areas.

3.4.3 Incremental Panoramic Calling



While Live Spatial Video Calling can allow for independent views of the space, it does not provide any static context for areas outside of the local user's FoV and so the local user can still effectively dictate what their partner is able to look at. Incremental Panoramic Calling mitigates this by recording each of the local user's camera frames whenever they are projected into the environment; over time this will construct a panorama of the entire space, allowing the remote user to independently

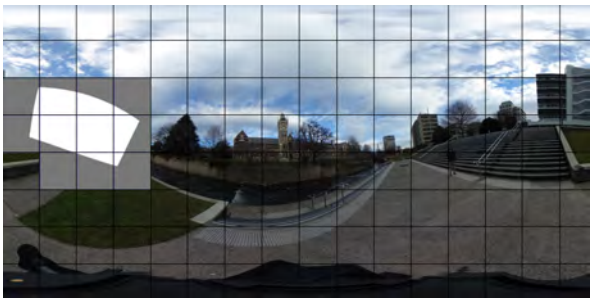


Figure 3.9: An illustration of how culling quads are overlaid on the panorama framebuffer to prevent the projection shader executing for every fragment. Here the white area shows the projection area and the grey squares show the quads that will have the shader executed on them. The coloured regions will not have the projection performed on them at all, drastically reducing computation time.

view previously seen areas at their leisure.

This requires an additional step before rendering. Whenever a camera frame arrives from the local user, an additional shader is executed that projects it into the environment but with the panorama framebuffer as the output target rather than the screen buffer, making the image visible to subsequent invocations of the rendering shader. Newer pixels always overwrite old ones to ensure that the latest information is available to both parties, and to prevent excessive processing overhead it is assumed that the estimated device orientation is accurate and thus no additional stitching is performed.

A side effect of this process is that the local user's gestures may now be permanently recorded in the environment, occluding the background and making subsequent gestures more difficult to interpret. To mitigate this, the hand segmentation algorithm described in subsection 3.3.5 is performed during this projection, but in reverse: a pixel is not projected into the panorama if it is determined to belong to a hand. The original image is unaffected by this process and thus any gestures will still be visible after rendering.

As the panorama framebuffer is used as the output for the projection shader, it would usually operate on all 4096×2048 fragments in this buffer even though only a small fraction of them will be altered, unnecessarily impacting performance. To prevent this, the panorama is overlaid with 144 culling quads, each of which will be used as the output for the shader only if it is determined that it will be affected by the projection. This process is illustrated in Figure 3.9.

For a camera with diagonal field of view F we can determine if a culling quad will

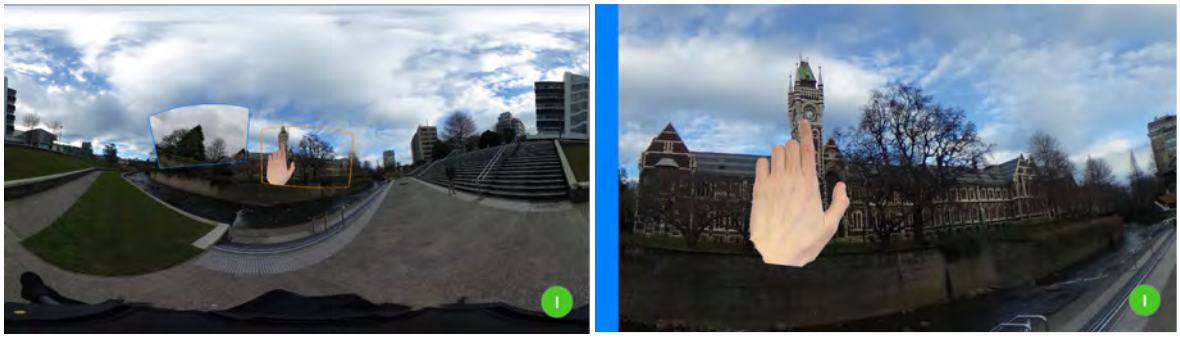
intersect with the area being projected to if any of its sphere space corners $\mathbf{c}$ satisfy

$$R\mathbf{z} \cdot \mathbf{c} \geq \cos\left(\frac{F}{2}\right) \tag{3.8}$$

where $\mathbf{z}$ is the unit vector along the z axis and R the orientation matrix for the frame being projected.

This mode operates similarly to PanoVC (Müller et al., 2016) but differs in a few key areas. The use of a cylindrical panorama better accommodates HMD use, increasing the potential for immersion within the space and ensuring everything directly above and below the local user is visible. Orientation matrices are also calculated locally and sent along with each camera image, meaning they don't have to be needlessly recalculated by the receiver and thus improving performance and environmental consistency. Integration of gestures also provides a representation of the remote user that PanoVC lacked, which Müller et al. identified as a limiting factor in the amount of co-presence the system could induce.

3.4.4 Panoramic Calling with Live Inserts



The device's inbuilt camera is capable of building convincing reconstructions of the shared environment, however their quality is sensitive to inaccurate orientation estimations and temporal changes in the environment. Sensors are susceptible to drift over time, and translational movement can cause confusion in vision-based stitching, so artefacts will likely be present within the constructed panorama in less-than-ideal conditions. Panoramic Calling with Live Inserts attempts to avoid these artefacts by allowing the panorama to be constructed before communication starts, enabling external hardware or slower, more accurate stitching to be used. We assume that this panorama will already be equirectangularly projected and so it is simply copied into the environment buffer when the application starts; other than this initial step, the implementation is identical to Live Spatial Video Calling.

Figure 3.10: A pre-captured panorama after being split and unprojected into multiple segments to give undistorted views of each area within the space. These segments are stored along with the orientation with which they were obtained and their set of feature descriptors so that new frames can be positioned relative to them.

The existing orientation estimation methods return rotations relative to the user's initial view direction. This is acceptable for the other modes as there is no absolute world-space orientation images need to adhere to, however in Panoramic Calling with Live Inserts this is no longer the case. To ensure that objects in the live video align with their locations within the static panorama it must be determined where these objects are; for this an additional orientation estimation method has been implemented that calculates the relative rotation between each frame and pre-defined segments of the panorama.

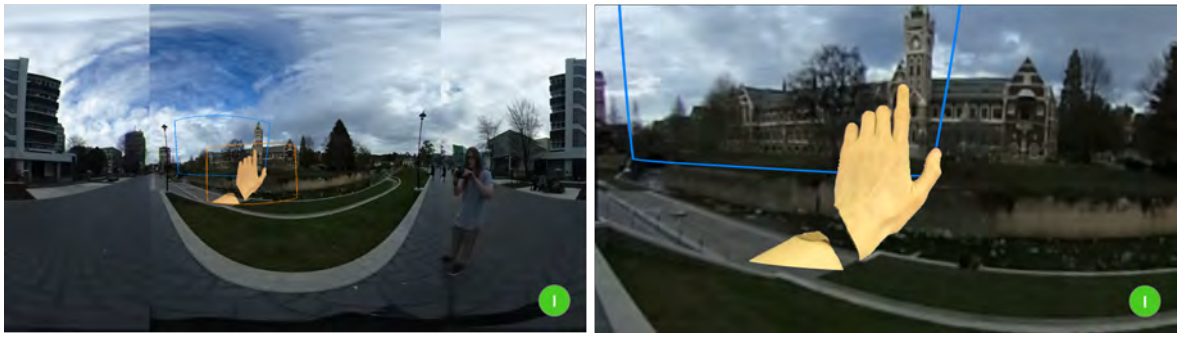
Once the panorama is loaded from disk, an undistorted image is unprojected from it at set angles which, as seen in Figure 3.10, cover the entire space between them with some overlap. These are stored along with the orientation at which that segment was obtained, resulting in a two-dimensional array of images such as seen in Figure 3.10. For each segment we then find and store its set of feature descriptors using the ORB method (Rublee et al., 2011).

For each frame that arrives from the camera we then detect its set of features, again using the ORB method. As the pitch of an object in the pre-captured panorama is likely to match its relative pitch in the real world, we find a likely row of images that the new image could match based on the pitch estimated by the sensor-fusion approach. Feature matching is then performed against all segments in this row, and after finding the segment with the most matches the relative rotation between it and the new image are calculated, which is multiplied by the absolute orientation matrix of the chosen segment to find the absolute orientation of the device.

This matching procedure is predictably slow, taking 453ms on average; it is thus only done every 500 frames, with the calculated orientation used as an offset for sensor-based tracking to maintain real-time performance.

This modes provides a similar experience to PanoInserts (Pece et al., 2013), but despite the computational limitations of mobile phones manages this at a much higher frame rate. The system also provides much more natural interaction through unmediated gestures and FoV indicators, providing representation of the remote user that PanoInserts lacks.

3.4.5 Live Panoramic Video Calling



The static nature of Panoramic Calling with Live Inserts means that temporal changes within the environment will not be made known to the remote user. Live Panoramic Video Calling utilises an external 360° camera to capture and share the full environment in real time. In this framework I use the Ricoh Theta S[9], an inexpensive and portable camera designed for use with mobile phones, which can be seen in Figure 3.11.

Interfacing with the Theta is usually performed using its own API through HTTP requests, however this requires connecting the mobile phone to a network broadcast by the camera that only accepts one client, making external connections with the communication partner impossible. The phone is instead connected to the camera via USB using UVCCamera[10], an open-source library that allows Android devices to interface with cameras over USB.

Frames are streamed from the camera using the dual fisheye format in Figure 3.12. For this to be useful, it must first be converted to an equirectangular projection so that it may be textured on the environment sphere. As with Incremental Panoramic

[9]https://theta360.com/en/about/theta/s.html
[10]https://github.com/saki4510t/UVCCamera

Figure 3.11: The Ricoh Theta S, a camera used for 360° image capture in Live Panoramic Video Calling. (Left): A close-up of the Theta being worn around the user's neck. (Right): The Theta connected to a mobile phone via a USB cable.

Calling, this projection is performed in an additional fragment shader before rendering with the panorama buffer as the output target.

For each fragment in the panorama framebuffer I calculate its latitude φ and longitude λ on the environment sphere

$$(\varphi, \lambda) = \left(\frac{\text{-}\pi}{2} + \frac{\pi t_y}{r_y}, \text{-}\pi + \frac{2\pi t_x}{r_x} \right) \tag{3.9}$$

where (t_x, t_y) is the coordinate of the current fragment and $(r_x, r_y) = (4096, 2048)$ the resolution of the panorama framebuffer. These are projected to a unit vector $\mathbf{m}$ on the environment sphere, which is rotated so that it aligns with the coordinate space of the panorama:

$$\mathbf{m} = \begin{bmatrix} 0 & 0 & \text{-}1 \\ \text{-}1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \cos(\varphi)\cos(\lambda) \\ \sin(\varphi) \\ \text{-}\cos(\varphi)\sin(\lambda) \end{bmatrix} \tag{3.10}$$

From here the spherical equirectangular coordinates are calculated

$$\begin{aligned} \theta &= \begin{cases} 0 & \text{if } |\mathbf{m}_y| \geq 1, \\ \cos^{\text{-}1}(|\mathbf{m}_y|) & \text{if } |\mathbf{m}_y| < 1 \end{cases} \\ \phi &= \begin{cases} \tan^{\text{-}1}\left(\frac{\mathbf{m}_x}{\mathbf{m}_z}\right) & \text{if } \mathbf{m}_y < 0, \\ \tan^{\text{-}1}\left(\frac{\text{-}\mathbf{m}_x}{\mathbf{m}_z}\right) & \text{if } \mathbf{m}_y \geq 0 \end{cases} \end{aligned} \tag{3.11}$$

where θ is the azimuth and ϕ the inclination of the current fragment on the environment sphere, with corresponding texel coordinates (u, v) in the fisheye texture

$$(u, v) = (t_x, t_y) \cdot \begin{bmatrix} c_x + \frac{r\pi}{2}\cos(\phi)\left|\sin\left(\frac{\theta}{2}\right)\right| \\ c_y + \frac{r\pi}{2}\sin(\phi)\left|\sin\left(\frac{\theta}{2}\right)\right| \end{bmatrix}. \tag{3.12}$$

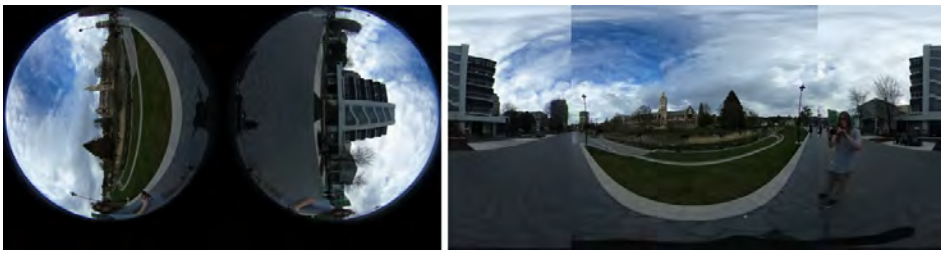


Figure 3.12: An example of the 360° images obtained by the Ricoh Theta S. (Left): The duel-fisheye format natively captured by the camera. (Right): The same image after being projected into the equirectangular panorama.

Here (c_x, c_y) is the centre point of one of the fisheye images seen in Figure 3.12 and r its radius. If $\mathbf{m}_y < 0$ the texel is taken from the left image, otherwise it is taken from the right. As each lens of the Ricoh covers slightly more than 180° and its intrinsics are not publicly available a perfect projection is difficult without computationally expensive stitching, causing visible seams where the two images meet.

3.5 Evaluation of Requirements

So does this framework fulfil the requirements set out in subsection 2.3.6, and thus maximise the presence induced in its users?

The first such requirement was that the environment should remain consistent between clients and facilitate views of both the task space (where applicable) and the users themselves. The panorama is guaranteed to remain consistent and symmetrical across the connection as both clients use the same orientation matrices for projecting images and WebRTC's use of TCP for transmitting packets ensures these matrices will not be lost. The representations of objects also makes the relationships between them explicit in all modes except Live Video Calling as the remote user can see their spatial context and thus infer their placement relative to each other. Unfortunately, in its current form this system facilitates no way of viewing the user. This is largely due to modern phones not allowing their forward- and backward-facing cameras to be used simultaneously due to limited bandwidth on the camera bus, and so such views would always require an external camera to be attached to the phone. While this would technically be possible with the external Ricoh Theta used for Live Panoramic Video Calling, this was purposefully not added to ensure functionality between the modes of interaction differed only in the view independence they offered.

This framework does however mostly fulfil the second requirement, which is that

users should be able to explore the environment completely independently from one another. While this obviously differs between modes of interaction, each other than Live Video Calling allows the remote user to obtain completely independent views by simply rotating their device, and each successive mode ensures more content will be available to view in areas outside of the local user's current field of view. Where this framework falters is in its ability to allow translational movement within the space; as the remote user is constrained to the local user's position at all times, there is no way to actually walk around the space other than request that the local user do it for them, breaking any static panoramas in the process.

The third requirement was that while exploring, users should be aware of where their partner is at all times through sharing of their current position and gaze direction. Positional indicators are not required in a two-dimensional environment as peers will always be co-located, and the current gaze direction is shown in this framework by outlining each user's current field of view. This is only an approximate representation; it could be that the user is looking at any number of objects within their FoV, especially when wearing an HMD where this view completely encompasses their vision, however more fine-grained visualisation would require eye tracking and thus more external hardware.

The fourth was that each user's gestures and body language should be shared to aid in natural conversation. In this framework unmediated video of each user's hands are generated and spatially rendered in the environment, ensuring both pointing and representational gestures can be used without any context required to interpret them being lost. However, other body language is not currently shared as there is no means to capture it. It's possible that each user could be tracked using the 360° camera in Live Panoramic Video Calling, which is a concept that will further be explored in section 5.3.

The final requirement was that all features are facilitated through purely mobile devices. This has been achieved; of all the modes of interaction only Live Panoramic Video Calling requires any additional hardware, and this is only for environmental capture. All other processing is done on the mobile phone, which itself is several years old and thus much less capable than many other consumer-grade devices. Achieving this means that all of the features outlined in this chapter are now available to the majority of the population on a device they already own.

Chapter 4

Evaluation of the Mobile Framework

With the foundations laid it must now be determined whether the interactions afforded by this framework are sufficient in inducing the desired sense of presence within and between its users, and whether the system is capable of doing this with acceptable performance. A full technical evaluation of the application in each of the modes of interaction is thus detailed, with each of these also evaluated by novice users to determine the degree of freedom a user must have within an environment in order to feel a sufficient sense of spatial presence within it. Though the application proves capable of providing real-time interaction, such a finding would allow the hardware resources allocated to environmental reconstruction to be used in additional features, providing more opportunities for interaction or higher visual fidelity without detrimental effect to the presence felt within the shared space.

4.1 Technical Evaluation

For this experience to provide a real sense of presence between its users it is necessary for it to operate in real time. This requires a high frame rate so that movement appears fluid and low latency so that actions performed by users provide immediate visual feedback from their actions as they would in the real world. The application has thus been benchmarked to determine whether a mobile device is capable of delivering such an experience with sufficient performance.

Two Google Pixels running Android 7.1.2 were used for testing. While accurate measurements are difficult to obtain, the camera-to-screen latency of this device was measured to be at worst 128ms, a Wi-Fi connection over a local-area network was similarly measured to introduce at most 128ms of latency, and the USB connection

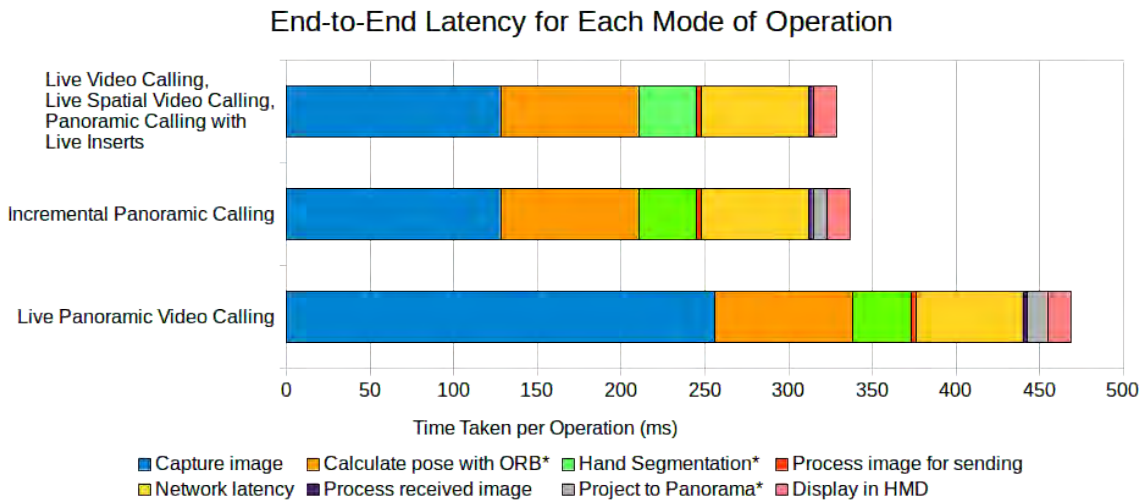


Figure 4.1: The average end-to-end latency of each mode, assuming that users are viewing a unique unprojected view of the environment. Sections marked * are not always required: ORB tracking is optional, projecting to the environment outside of the renderer is only required in Incremental Panoramic Calling and Live Panoramic Video Calling, and hand segmentation only occurs on the remote user's device, so these values can be reduced further in many cases. The standard deviations were negligible and are thus omitted.

used for the Ricoh Theta introduces 256ms of latency when capturing 360° video. These latencies are unavoidable and thus any similar system has a best-case end-to-end latency of 256ms.

The Pixel's camera operates at 30frames per second (fps), and the Ricoh Theta S is only capable of streaming images to third-party applications at a rate of 15fps, capping the rate at which the environment can be updated to these values in their respective modes. The Pixel's screen updates at 60Hz, making this the highest frame rate the application can achieve due to Android's enforced vertical synchronization.

The average end-to-end latency, the time elapsed between an image's capture and its display on the remote peer's screen, is shown in Figure 4.1, and the average frame rate and time to process each frame is shown in Figure 4.2. These values proved constant over several measures, resulting in negligible standard deviations that are thus omitted. The application achieves 60fps rendering in most configurations, which is the highest attainable on the Pixel. Incremental Panoramic Calling and Live Panoramic Video Calling have slightly lower performance due to the extra projection required to construct the panorama.

The local user can expect to see images from their own camera only 141ms after they are captured in most cases, which is only slightly higher than camera's inher-

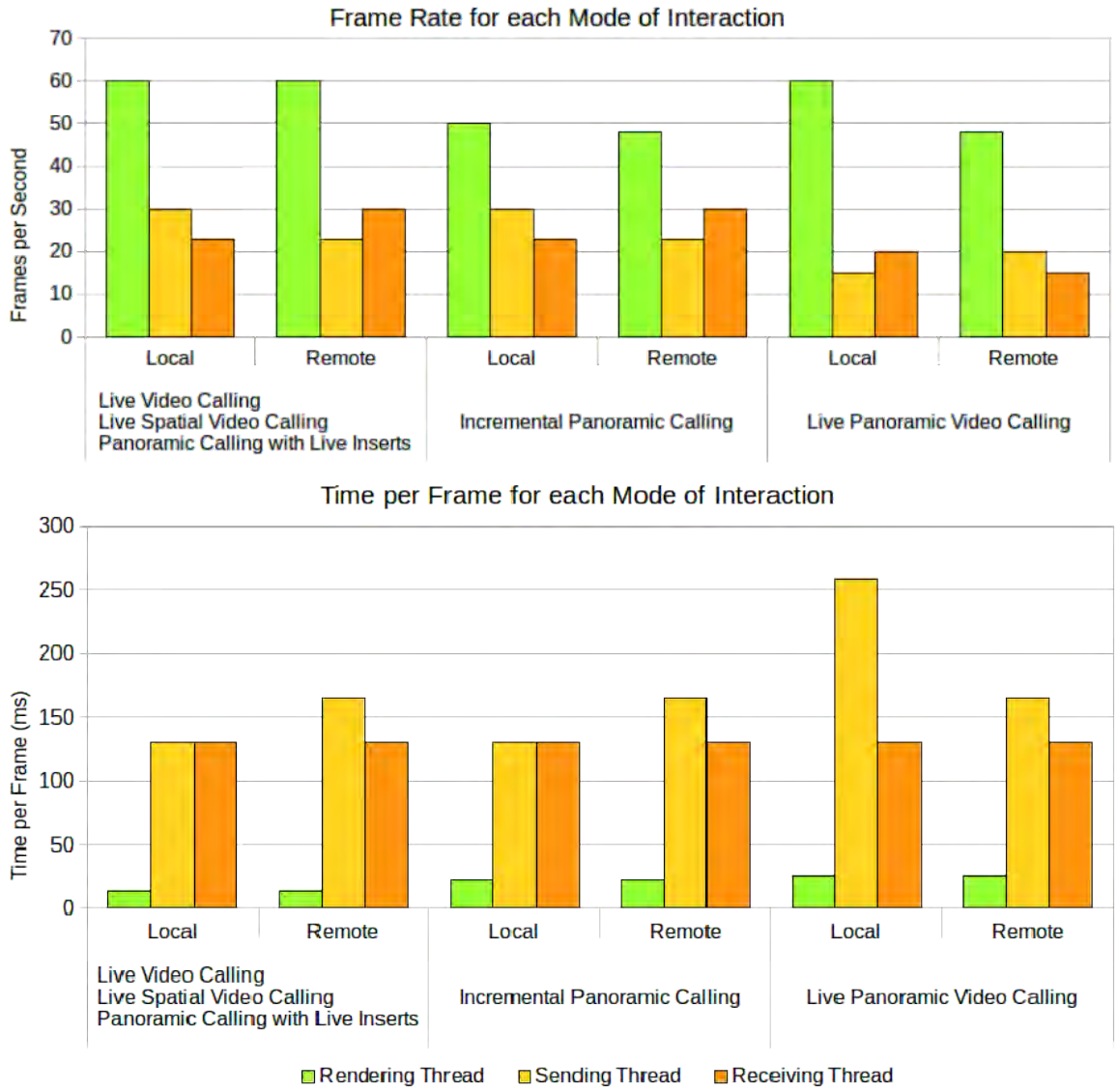


Figure 4.2: The average frame rate (top) and time to compute each frame (bottom) for each mode of interaction for the local and remote user. Each of the application's main threads execute asynchronously and affect output differently so are evaluated separately. These results assume that sensor-based tracking is used. Standard deviations were negligible and thus omitted.

ent latency. The remote user can expect to see these in approximately 302.34ms or 351.54ms depending on whether they were captured by the local user's integrated or external camera. The remote user can similarly expect to see their own camera image, including their segmented hands, in 175ms, and expect that the local user will see it in approximately 246.19ms. This latency is lower than that experienced in other systems that saw no related ill effects (Müller et al., 2016; Tang et al., 2017), and is almost as

low as possible given the limitations of the camera and network, implying that future hardware revisions will bring these numbers even lower.

The application is comprised of three main threads: the *rendering thread*, the *sending thread*, and the *receiving thread*. Here we evaluate the tasks each is responsible for, how long these tasks take to perform, and how these values affect the overall perceived performance of the application.

4.1.1 Rendering

The rendering thread is responsible for updating and displaying the environment sphere. It performs the following tasks each frame:

1. The latest frames are retrieved from both the camera (for local images) and the receiving thread (for remote images) as well as their associated orientation matrices. The time this takes is negligible (<1ms).

2. In Incremental Panoramic Calling or Live Panoramic Video Calling, the latest camera image is projected into the environment. This takes 8.42ms for images from the inbuilt camera and 12.00ms for panoramic images. The local user's hands may also be removed from the image during projection to avoid occlusions in the environment which requires a further 34ms of processing time.

3. The newly updated environment is rendered to the display, which also requires projecting both users' camera images into the environment. This takes 14.05ms to display the full panorama, 13.75 for a unique unprojected view based on the user's orientation, or 13.97ms for a pseudo-sterescopic HMD view.

The performance of this thread determines the perceived performance of the application and as such is the most important to optimise. A low frame rate would affect the update rate of both the display and orientation estimates, and high latency could result in higher risk of motion sickness when an HMD is used as head rotations would not provide immediate feedback.

Fortunately, this thread executes at 60fps in most cases. The remote user may see lower performance in both Incremental Panoramic Calling and Live Panoramic Video Calling; this is due to increased resource contention caused by hand segmentation and the additional projection.

4.1.2 Sending Frames

The sending thread is responsible for processing frames and passing them to the network module so that they may be successfully sent to the remote client. It is responsible for the following tasks each frame:

1. Retrieves the latest local image from the camera. This takes at most 128ms for the device camera and 256ms for the Ricoh Theta S.

2. On the remote user's device, this image is then passed through the hand segmentation module to isolate their gestures. This takes 34ms on average.

3. The latest orientation estimate is retrieved. When using sensor fusion the time this takes is negligible (<1ms). The vision-based approach is much slower, requiring 82.38ms on average and reducing the thread's frame rate to 14fps.

4. The image is then processed for sending. This involves creation of its accompanying data packet that will contain the orientation matrix, an identification tag, and the pixels removed when embedding that tag into the image. This takes approximately 2.57ms.

The performance this thread can achieve is limited by the camera used, resulting in best-case latency of 128ms or 256ms and a maximum frame rate of 30fps or 15fps when using the integrated or panoramic camera respectively. This results in far fewer frames being captured than the application is capable of rendering, so future hardware revisions will result in immediate benefits.

Poor performance in this thread would result in users seeing the environment updated only infrequently, which may result in important information being lost. High latency could also cause conversation to become difficult or awkward as both users would have to wait a perceivable amount of time to see the other's actions.

4.1.3 Receiving Frames

The receiving thread is responsible for processing frames as they arrive over the network so that they can be used by the renderer. It is responsible for the following each frame:

1. Retrieves the latest frame from the remote client, which takes at most 128ms due to network latency.

2. Reconstructs the received images with the pixels removed to accommodate the identification tag, then passes the reconstructed frame to the renderer. This takes 2.57ms on average.

As with the sending thread, the latency here is limited by the network delay and its frame rate capped to the capture rate of the remote client's camera. This thread has the least computation to perform and thus can easily keep up with demand.

4.2 User Evaluation

We know that allowing independent views between users can increase their sense of spatial presence within the environmment (Jo and Hwang, 2013), but it remains to be seen if this increase is subject to diminishing returns; that is, if some "sweet spot" exists where an increase in view independence results in only a negligible increase in the presence induced. Such a finding would be of huge importance to mobile telepresence; the limited computational capability of these devices means not all features can always be included in an application, so if construction of the environment is less computationally demanding then the resources saved can be used to create richer interactions within it.

To find this sweet spot, a study was conducted on novice users who were asked to test the system and evaluate the degree to which they felt present within a shared environment and with the remote person they shared it with. Before finding this aforementioned "sweet spot" we must first confirm that a correlation between view independence and presence exists, and that the link between the two seen in previous research is not just a binary increase. There are thus two hypotheses to confirm:

1. There is a correlation between the degree of view independence provided to users and their sense of spatial presence within the environment.

2. There is a correlation between the degree of view independence provided to users and their sense of co-presence with their communication partner.

Each condition presented the participant with a different mode of interaction as these differ in how much free view independence they provide. As a limited number of participants took part, the number of conditions evaluated was limited to ensure results were useful.

Participants viewed the environment through a Google Daydream HMD, both to allow themselves to fully immerse within the space without distraction, and to prove

the system suitable for HMD use. Live Video Calling was thus excluded from testing as it does not lend itself to comfortable HMD use due to users having no control over their view, and a comparison between it and other modes has been done before (Jo and Hwang, 2013). Panoramic Calling with Live Inserts was similarly excluded due to the low participant count as it differs the least from the other modes, essentially serving as a best-case scenario for Incremental Panoramic Callingin that the environment it presents is similarly static in nature; if it was found that the sweet spot existed somewhere between Incremental Panoramic Calling and Live Panoramic Video Calling, thus suggesting a difference between live and static environments rather than immediate and incrementally constructed ones, the intention was that further testing would be performed.

This experiment was conducted with the approval of the University of Otago Human Ethics Committee (Non-Health).

4.2.1 Study Design

19 participants took part, all between the ages of 18 and 65. Each identified as having no prior history of simulator sickness, and only seven had prior experience with virtual reality. The study followed a within-subjects design where the independent variable was the mode of interaction used. All modes were viewed through a Google Daydream HMD. The degree to which participants felt spatially present within the presented environment and co-present with their communication partner were evaluated using modified versions of questionnaires designed by Schubert et al. (2001) and Biocca et al. (2003) respectively; these consisted of statements about the user's experience, with the participant indicating the degree to which they agreed with each statement via a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). Statements were worded in a way such that a higher score indicates more presence was induced. Space was left at the end of the questionnaire for participants to write any free-form comments they may have had about their experiences with the system, and any simulator sickness experienced during the experiment was evaluated using a questionnaire by Kennedy et al. (1993) after all conditions had been completed. The complete questionnaire may be viewed in Appendix A.

Each condition consisted of an informal conversation with a remotely located study mediator, whose video and orientation data was recorded ahead of time to ensure that the environment remained consistent between participants and conditions; this was then streamed into the application so that participants could still make use of

the system's interactive elements. Pre-recording the environment in this way allowed unpredictable factors such as lighting to be controlled, preventing extreme conditions from interrupting the experiment or biasing its results. Due to the difficulty of placing a 360° camera such that the user is not visible, this environment was restricted to 180° in all conditions to ensure a consistent first-person view between them. Participants were informed that the video they were seeing was not live before the experiment began, and their own video and orientation data was still captured and broadcast live to the study mediator over Wi-Fi so that their gestures and view direction could be seen in relation to the environment. Both the participant and the mediator were situated within the same room so could hear each other speak without use of the application's VoIP features, though the mediator was not visible to the participant to prevent this from biasing results. The same environment and study mediator were used for all participants and conditions.

As interacting with a partially simulated communication partner may affect how co-present one may feel with them, the experiment was repeated with live video and orientation data to determine the system's suitability to a real-life scenario. For this experiment the study mediator was physically located within the same environment as used for the last one with all data transmitted in real time over Wi-Fi. Seven participants were recruited, each fitting into the same demographic as for the previous experiment.

4.2.2 Procedure

The ordering of conditions was randomised for each participant to reduce potential learning effects. For each condition the participant was instructed on how to operate that mode of interaction and given two minutes with the system in a pre-recorded environment to familiarise themselves with its use. This was done with pre-recorded video and orientation data and without a communication partner so that they would not be discouraged from experimentation.

After these two minutes had elapsed, participants were connected to the study mediator and virtually placed within an urban outdoor hill-top environment where they were asked to have a two and a half minute informal conversation with the mediator, who showed them various landmarks and encouraged them to discuss any others they could see. They were also asked if they knew of any other landmarks in the area to encourage active participation in the conversation. Once finished, participants were then asked to complete the presence questionnaires. This procedure was repeated for

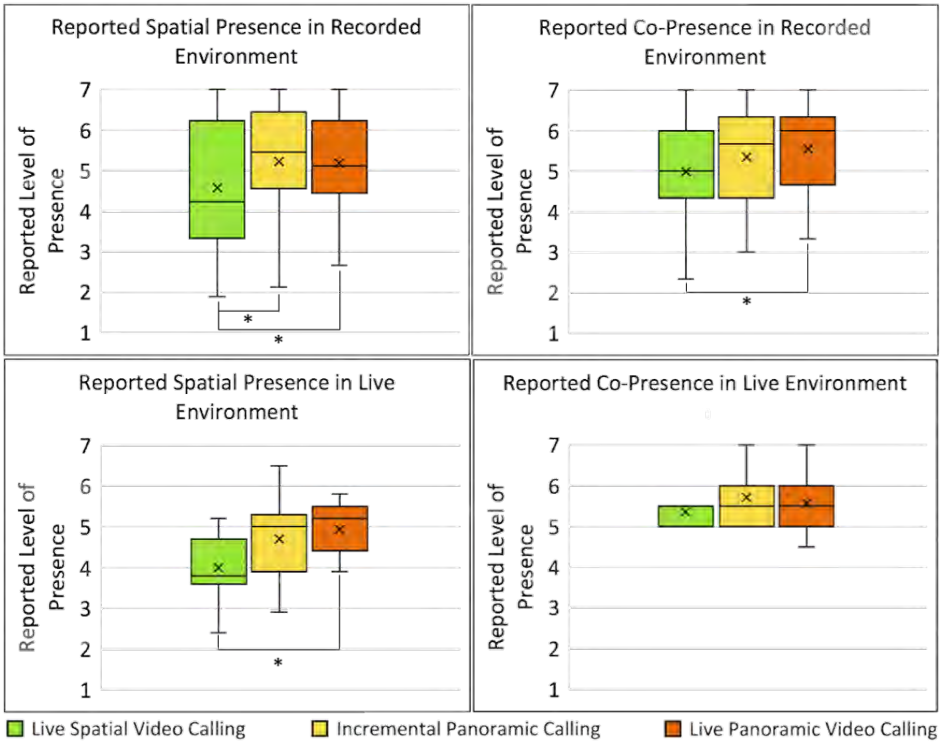


Figure 4.3: Participants' reported levels of spatial and co-presence within the pre-recorded (top) and live (bottom) environments. These were taken as the means of the questionnaire responses related to that form of presence.

the remaining conditions, after which the participant completed the simulator sickness questionnaire and was gifted a $20 NZD voucher.

Participants were informed that they could conclude the study at any point without disadvantage to themselves if they experienced any symptoms related to simulator sickness.

4.2.3 Results

The spatial and co-presence scores for each condition were determined by the mean of the relevant questionnaire responses and are shown in Figure 4.3. Live Spatial Video Calling (C1) scored the lowest of the modes tested for both forms of presence in the pre-recorded environment with mean scores of 4.56 ($\sigma = 1.48$) and 4.98 ($\sigma = 1.23$) for spatial and co-presence respectively. Incremental Panoramic Calling (C2) was higher rated, with mean scores of 5.22 ($\sigma = 1.43$) and 5.35 ($\sigma = 1.35$), and Live Panoramic Video Calling (C3) achieved similar results with mean scores of 5.19 ($\sigma = 1.18$) and 5.54 ($\sigma = 0.94$). Friedman tests showed a significant difference between conditions ($\alpha = 0.05$) in both spatial presence ($p = 0.019$) and co-presence ($p = 0.015$). Wilcoxon

signed-rank tests revealed that C1 induced significantly lower spatial presence than both C2 ($p = 0.012$) and C3 ($p = 0.019$), with no significant difference between C2 and C3 ($p = 0.917$). Co-presence was similarly distributed, with C3 scoring significantly higher than C1 ($p = 0.008$) but with no significant difference between C1 and C2 ($p = 0.343$) or between C2 and C3 ($p = 0.586$).

The live study provided similar results as C1 again was rated the lowest for both forms of presence with mean scores of 3.99 ($\sigma = 0.90$) and 5.39 ($\sigma = 0.24$) for spatial and co-presence respectively. C2 and C3 also scored similarly, with C2 achieving mean scores of 4.70 ($\sigma = 1.17$) and 5.71 ($\sigma = 0.70$) and C3 scoring 4.94 ($\sigma = 0.67$) and 5.57 ($\sigma = 0.79$) for spatial and co-presence. Friedman tests again showed a significant difference in induced spatial presence between conditions ($p = 0.030$), but this time no significant difference in co-presence was found ($p = 0.368$). Wilcoxon signed-rank tests found that participants felt significantly more spatially present within the environment in C3 than in C1 ($p = 0.035$), but this time no significant difference was found between C1 and C2 ($p = 0.051$) or between C2 and C3 ($p = 1.00$).

Results of the simulator sickness questionnaire suggest that few symptoms were experienced by participants. Responses to each symptom were coded to allow for numerical analysis ("None" = 0, "Severe" = 3), resulting in an average response of 0.313 ($\sigma = 0.192$) across all symptoms. No participants felt it necessary to conclude the study due to these symptoms despite there being no disadvantage in doing so.

4.3 Discussion

Overall, the system was well received by participants, with those unfamiliar with VR having no difficulty adapting to the means of interacting with it. The frame rate and latency proved adequate, and the experience was smooth enough that little simulator sickness was felt. It is unsure whether this would hold true for longer exposure times (Kennedy et al., 2000), though as it stands this system could easily be adopted by the wider public without issue. Many participants saw the system's potential as the future of telecommunications, with comments such as "After I took [the HMD] off I had forgotten exactly where I was... really did feel like I was there" and that it "makes one actually feel they are really in the same place with the other person. Taking communication to another level!".

4.3.1 Spatial Presence

As per the first hypothesis, it was believed that there would be a positive correlation between the degree of view independence provided to users and the spatial presence they felt within the virtual environment. This was partially supported by the results of these experiments; both Incremental Panoramic Calling and Live Panoramic Video Calling induced significantly higher spatial presence than Live Spatial Video Calling, which suggests that allowing the remote user to view areas outside of the local user's field of view could be beneficial to an increased feeling of presence. Participant reactions further support this hypothesis, with comments such as "[C3] felt more immersive [than C1] as you can see the whole area", "Looking through a small window [in C1] made it harder to fully immerse", and "I felt more immersed with being able to see outside the blue square".

This difference in spatial presence was not seen between C2 and C3 in either experiment, which could suggest that the partial panorama used in Incremental Panoramic Calling is just as effective at inducing presence within the environment as the full panorama used in Live Panoramic Video Calling. This implies that this is where our desired sweet spot lies; wider context outside of the live video is required for users to feel present within the space, though it doesn't matter whether this is constructed incrementally throughout the call or updated in real time. Participant comments echoed this sentiment; one noted that "Due to the headset only displaying a small field of view, and not having any control on orientation, it did not feel as encapsulating as [C2 or C3]", one that it "didn't really feel I was in that environment maybe because I saw my surrounding as like an isolated dark place and all I had to focus on was the [FoV indicator]", and another that "I felt more immersed with being able to see outside the [mediator's FoV] even if it was just a still image".

The similarity in spatial presence induced between Incremental Panoramic Calling and Live Panoramic Video Calling could also be due to the limited resolution provided by the Ricoh Theta, which was mentioned by participants as a detriment with comments such as "The lower resolution [in C3] made me feel slightly dizzy", "I felt the resolution on this really impacted how immersed I felt", and "because the image was so grainy I felt like I was in a game rather than a real location". Improvements in panoramic camera resolution may mitigate this in future, however network limitations constrain the resolution of images that can be streamed in real time and so using all available pixels over a smaller area as in Incremental Panoramic Calling will always result in a higher resolution environment than using the same number over the full

panorama.

It was assumed that the Ricoh Theta's reduced frame rate would cause a similar detriment in induced presence, however no participants mentioned this or indicated that they noticed it at all. In fact, the performance of the system as a whole seemed satisfactory for users; only one reported any slowdown, which was caused by unscheduled background processing and not experienced in subsequent conditions.

Despite Live Spatial Video Calling inducing significantly less spatial presence than the other conditions, participants still felt that the limited view control was sufficient in making them feel as though they were within the presented environment. One noted that "even though the environment is less engaging in terms of most of your vision isn't occupied.., as I became more comfortable and used to the environment I could map it better in my head". This could suggest that even though no wider context is shown to the user as they look around, being able to follow a moving video feed was sufficient in making the spatial relationship between objects in the presented environment clear.

4.3.2 Co-Presence

The second hypothesis, that there would be a positive correlation between the degree of view independence provided to users and the co-presence they felt with their communication partner, was not supported by the experiments. A small correlation was found, with Live Panoramic Video Calling inducing significantly more co-presence than Live Spatial Video Calling in the pre-recorded environment, but not enough for a significant difference to present itself between the other conditions or for this to be replicated in the second experiment. It's not known why this difference wasn't repeated in the live study; perhaps video compression caused the quality of the 360° images to decrease even further to the point of being detrimental to the experience, or the low participant count meant that those who saw quality as an issue were more influential in the final results.

The ability to perform gestures to augment conversation was well received by participants. Deictic references were frequently used to point to objects of interest, and representational gestures were also often used for tasks such as tracing the path of a river. Participants felt that this contributed to their sense of presence within the environment, with one saying that "being able to see the movements of my own hands creates a much higher sense of engagement than if this feature was not included, and drives most of the sense of 'being there'". No participants noticed any artefacts from incorrect segmentation, even with their hands in front of an unprepared background.

Co-presence was rated highly across all conditions, which was particularly surprising in the pre-recorded environment as participants knew that the mediator wasn't really there and thus could not properly react to their actions. This may have been biased by the mediator's physical presence within the room during the experiment, however the similar scores between the two experiments suggests no such bias exists.

4.4 Summary

The last two chapters have provided a viable framework on which future telepresence experiences can be based, with the knowledge of how it can be implemented and results showing that it provides a natural and efficient communications tool to its users. All features were achieved using purely mobile hardware, meaning they are now available on a platform accessible by the majority of the population, and interactions occur in real time to deliver the smooth conversational experience users have come to expect from existing videoconferencing solutions.

The framework provides an immersive panoramic environment within which users can interact, constructed from the local user's surroundings based on video captured by their mobile phone. Interlocutors can obtain independent views within this environment by simply reorienting their device and know that their current gaze direction is shared with their partner so that deictic references can be made without breaking any common ground. These references can be made even clearer through the use of gestures, which are captured and spatially rendered in the environment to provide a natural and intuitive aid to conversation.

Five means of constructing this shared environment were implemented. Live Video Calling acts similarly to traditional videoconferencing aided by gestural interaction. Live Spatial Video Calling enhances conversation by spatially rendering the local user's video feed based on the orientation of their device, providing limited but helpful spatial context to the environment. Incremental Panoramic Calling further increases this context by recording each image as it is rendered, over time incrementally creating a static panorama of the local user's surroundings. Panoramic Calling with Live Inserts removes this construction step by performing it before conversation begins, allowing the entire environment to be visible throughout the entire call. Finally, Live Panoramic Video Calling removes the concept of reconstruction altogether, instead using an external 360° camera to capture and render the entire environment live in real time.

A subset of these modes of interaction were tested by novice users, proving the

framework's ability to induce a heightened sense of spatial presence within them. A comparison between them showed no effect of the way the environment is created on the co-presence felt between users, but that they felt equally spatially present within an incrementally-constructed static environment as they did in a live, fully-encompassing one.

This latter discovery is of particular importance as it significantly lowers the hardware requirements of any telepresence platform. 360° video requires an additional device consumers aren't likely to own, so without this the cost of entry for consumers is further lowered and the resources previously allocated to full environmental reconstruction can be used to enhance the possible interactions between users without detrimental effect on the remote user's perception of the shared space and thus the possible richness of communication.

Chapter 5

Telepresence in Three Dimensions

The main limitation of the framework presented in chapter 3 is its inability for users to stray from their partner's viewing position. Full rotational independence could easily be obtained, however the remote user is always bound to the local user's position and thus has no freedom to walk around the shared space and explore it as they wish. While participants enjoyed the prototype, they also found it uncanny that they were essentially inside of the study mediator.

The obvious solution is to provide six degrees of freedom within a shared three-dimensional environment. Not only would users be truly independent from one another, but even more static context could be provided to users which could potentially induce more spatial presence than possible with rotations in a two-dimensional space.

This is a concept that has been visited many times in the past and is characterised mainly by outside-in systems where the area to capture is surrounded by RGBD cameras. Unfortunately these 3D environments would be difficult to implement on a single hand-held device that could only obtain limited views of the environment, and so most previous efforts focus on desktop systems where the possible tracking space is severely limited (Fanello et al., 2016; Gao et al., 2017; Kasahara and Rekimoto, 2014; Komiyama et al., 2017). Others have seen the benefits mobile devices bring through spontaneity and arbitrarily-large tracking spaces (Sodhi et al., 2013; Gauglitz et al., 2012, 2014), though usually the phone is used as a simple display tethered to a desktop computer

and so these benefits are nullified. To the best of my knowledge no such systems had yet been developed that provide free movement through full-sized 3D environments without requiring expensive stationary or proprietary hardware.

In the previous chapter it was discovered that providing users static context outside of the usual live camera stream can significantly increase the sense of presence they feel within the shared environment. Furthermore, it is not important whether these views are complete and transmitted in real time or static and incrementally constructed during the course of conversation; both provide this same heightened sense of presence, and neither affects the co-presence felt between interlocutors in any significant way. Live and complete views often require additional hardware not yet integrated into mobile phones and are more computationally demanding, so this finding is advantageous as it means a single mobile device can still provide this heightened sense of presence to its users.

This also paves the way for full 3D remote environments to be possible on mobile hardware. If live and full coverage of such a space was required, this would be impossible using a single self-contained handheld device as multiple cameras would need to be placed around the area. Removing these restrictions means that the environment to explore can be incrementally constructed as the local user walks around it, and the static nature of these previously explored areas would serve no detriment in the presence they induce.

To this end I developed Mobileportation, a system that provides full 6DoF exploration of remote three-dimensional environments incrementally constructed on purely mobile hardware. The only devices required are a mobile phone with an integrated RGBD sensor, which is used to incrementally construct a 3D point cloud of the shared environment that can be explored exocentrically from outside of the local user's position, and an external 360° camera tethered to the phone via USB, which captures a complete higher-resolution live view of the user's immediate surroundings that can be explored egocentrically from inside the local user's position. This camera also tracks and captures views of each user's face that will be spatially rendered within the captured environment, allowing for natural face-to-face communication to occur without detracting from views of the space. This mobile configuration allows completely free exploration of an arbitrarily-large environment without regard to cables, tracking spaces, or other tethers, and provides an experience as close to true face-to-face communication as is currently available on consumer-grade hardware.

I show that despite the technical limitations of these mobile devices, this experience

can still be achieved in real time and with no noticeable latency over both WiFi and cellular networks, allowing for immersive communication wherever the user desires. I also show through user experimentation that this experience is much more fun and social than traditional 360° videoconferencing systems with no sacrifices in portability, affordability, spatial presence, or co-presence, while for the first time allowing truly arbitrarily-large environments to be captured and explored in real time without lengthy precomputation.

The description and evaluation of this system have been published in the Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT) (Young et al., 2020) and will be presented at UbiComp 2020[1].

5.1 System Overview

Mobileportation was developed as a novel approach to *nomadic* telepresence that combines the strengths of 2D and 3D environment reconstructions to enable a new communication metaphor not previously available on untethered mobile devices. This combination of environmental representations provides several benefits: while panoramas are simple to build and provide immediate coverage of the full environment, they are only valid from the camera's position and so inherently exclude translational exploration. Full 3D reconstructions are conceptually and computationally much more difficult to create and tend to have a lower fidelity than their panoramic counterparts, but allow for full 6DoF movement without introducing visible distortions. This system thus allows for novel interaction in object-focused scenarios such as in Figure 5.1 by capturing the target object in high detail and showing it in context with its surroundings, and also in environment-focused scenarios such as in Figure 5.5 where arbitrarily large spaces need to be reconstructed and explorable with six degrees of freedom.

To combine these two reconstruction methods, Mobileportation provides two ways of viewing the environment. No intentional switching between them is required; rather, the way the environment is presented depends on the distance between users, allowing them to focus on exploration rather than operating the application. These two modes and the transition between them are shown in Figure 5.2 and are as follows:

- *Exocentric View*: A 3D reconstruction of the local user's surroundings is incrementally captured as they walk around it using an RGBD sensor embedded

[1] http://ubicomp.org/ubicomp2020/index.html

Figure 5.1: An example of Mobileportation used in an object-focused scenario. (a): A user finds an object of interest that they want to share with a remotely located person. (b): The user creates a 3D reconstruction of this object by walking around it with their mobile device. As they do so, the resulting 3D data and the live video from the 360° camera are sent to the remote communication partner so that they can also view this object in real time. (c): The resulting 3D reconstruction of the object of interest. (d): This object as seen in from an egocentric viewing position, overlaid on the 360° video so that it can be seen within its wider environmental context.

within their mobile phone. Both users may walk freely through this space and obtain truly independent views by simply walking around their real one, with 6DoF tracking provided by this same sensor. This freedom of movement allows for immersive and independent exploration, however due to the incremental nature of the reconstruction only areas already visited by the local user will be visible to the remote one, and at a lower fidelity than video capture could provide. Each user's current position is shown as a 3D avatar and gaze indicator, with their face captured and overlaid on it to allow for face-to-face communication.

- *Egocentric View*: The live video from the 360° camera is shown from the local user's position to provide an immediate, high-detail view of the environment that users can explore in 3DoF by rotating their mobile phone. Any sufficiently close 3D data will still be visible, providing parallax to give the illusion of limited translational movement. As users are now co-located and will no longer be able to see each other's avatars, their face is instead shown in the top-right corner of their partner's display, though their gaze indicator is still rendered as usual.

To transition from an exocentric view to an egocentric one, a user must simply walk towards their partner's avatar. As the distance between them decreases, the 360° video will slowly fade in, and a "snap in" mechanism will smoothly guide the two users together. Similarly, when in this egocentric view, either user can simply walk away

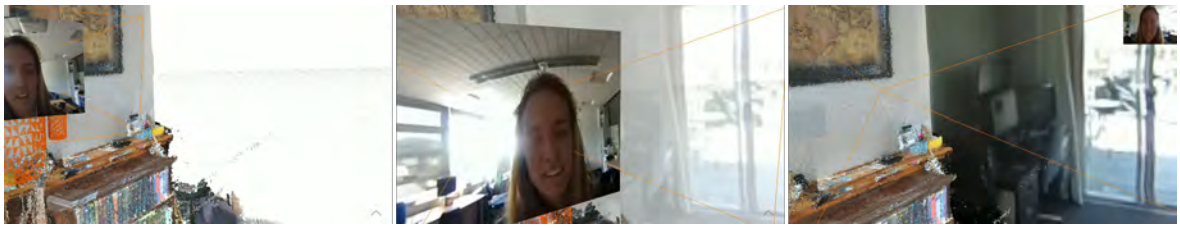


Figure 5.2: The application interface as seen by the user. (Left): Exocentric view. The two users are in separate locations within the virtual environment so only the 3D reconstruction is visible. The remote communication partner is shown as an avatar with their face capture overlaid. (Centre): As the users come closer together, the live video from the 360° camera gradually becomes visible. (Right): Egocentric view. The two users are now co-located and the 360° video capture has become fully opaque. 3D data is still visible for objects close enough for motion parallax to be noticeable.

from their partner to gradually transition back into the exocentric view. It is thus not possible for the local user to "carry" the remote one through the space, though as is the case in 2D environments (Jo and Hwang, 2013), requiring the remote user to follow their partner's movements through the space could provide additional spatial context and thus increase their spatial presence within it.

All that is required to create this 3D environment is a mobile phone with some means of depth capture. While this may seem rare, recent trends in mobile phone design make this combination far more likely in the near future. Many high-end devices such as the Samsung Galaxy S10+[2] or the Huawei P30 Pro[3] now have embedded Time of Flight (ToF) sensors, and even many that don't are capable of stereoscopic depth capture due to the inclusion of multiple rear-facing cameras. 360° video capture may also be more accessible in future, with devices such as the Essential Phone[4] or Motorola's moto z series[5] supporting modular 360° cameras, or the Samsung Galaxy series now integrating wide-angle 123° FoV lenses into the phone itself. If these trends continue, it's very possible that within the next few years, an experience such as Mobileportation will be available using a single self-contained device.

[2]https://www.samsung.com/us/mobile/galaxy-s10/
[3]https://consumer.huawei.com/en/phones/p30-pro/
[4]https://www.essential.com/
[5]https://www.motorola.com/us/products/moto-z-family

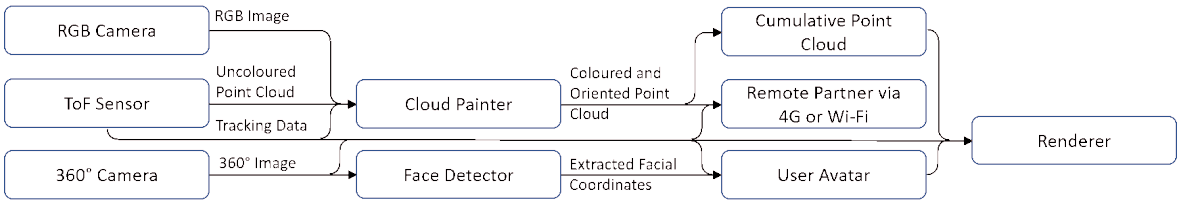


Figure 5.3: An overview of how data flows through the system's various modules to give the overall experience. The data here moves from left to right, so all is captured by the RGB, ToF or 360° camera and passed through several modules before being rendered or sent to the remote peer.

5.2 Implementation

Mobileportation was developed for the Android operating system. Almost all processing is done with C++ through the JNI, with only networking and the user interface handled Java-side. Depth capture is provided by Google's Tango API, which provides an interface to a Lenovo Phab 2 Pro's[6] integrated depth camera. This phone also performs all computation including rendering, tracking, networking, and scene reconstruction without help from external servers. Each frame, a point cloud is captured using this sensor and combined with the latest image from the phone's RGB camera to create coloured, oriented point clouds of the user's current view, which are then sent to the remote peer and stored in a cumulative point cloud stored on each user's GPU.

Panorama capture is again performed using a Ricoh Theta S connected via USB, this time attached to the mobile phone with a purpose-built 3D-printed mount as shown in Figure 5.4. The Lenovo Phab 2 Pro is unable to provide sufficient power for the Ricoh through its micro-USB port so an external power bank is required, though this is not necessary on modern USB-C compliant phones. Captures from this camera are transmitted across the network unprocessed; once received, images are passed through the face detection module, the result of which is used along with each user's tracking data to spatially render their avatar within the environment. The full 360° capture is also projected to the inside surface of a sphere surrounding the local user to facilitate egocentric viewing.

In the following sections I detail the specific implementation and algorithms of each of these processes, as well as key optimisations that were required for this experience to be possible in real time on the Lenovo's modest hardware. An overview of the interactions between these various modules is illustrated in Figure 5.3

[6]https://www.lenovo.com/us/en/smart-devices/-lenovo-smartphones/phab-series/Lenovo-Phab-2-Pro/p/WMD00000220

Figure 5.4: The hardware required for Mobileportation. The Lenovo Phab 2 Pro handles all tracking, rendering, reconstruction, and other computation, while the Ricoh Theta S provides 360° video capture and is connected to the phone via USB. The devices are combined using a purpose-built 3D-printed mount.

5.2.1 Depth Acquisition

The Tango API makes use of a mobile phone's integrated ToF camera to construct a depth map of the area in front of the sensor, then uses this data as a basis for SLAM to calculate the user's position within 3D space. With these two pieces of information we can then construct a point cloud consisting of objects directly in front of the user and find its absolute position within the space, allowing these captures to be combined over time to create full reconstructions of arbitrarily large environments. This cloud is not coloured, however, and so it must be combined with images from the phone's RGB camera to acquire a recognisable reconstruction.

At the beginning of each render frame the latest point cloud is requested from Tango. The Lenovo Phab 2 Pro's ToF sensor has a low capture rate and is only capable of producing these five times per second, so this process is skipped on intermediate frames to focus on other computation. This cloud is defined in screen-space, so each point's coordinate vector is multiplied by the user's co-temporal pose matrix to find its position in world space. To minimise computation this is performed within an OpenGL compute shader, which also projects each point into the latest RGB image to determine its colour. This projection takes the intrinsic and extrinsic parameters of the two cameras into account to eliminate distortion, though their aspect ratio differs so some points must be discarded as they lay outside the RGB camera's field of view. Each coloured point consists of its four-byte RGBA value and its 12-byte XYZ coordinate vector, resulting in 128 bits per point. The alpha value is not used, but is kept to maintain efficient memory alignment in 64-bit architectures.

This newly coloured point cloud is then passed to the renderer so that it can be

transferred to the GPU. To satisfy our real-time low-latency requirements, complex structures such as octrees cannot be used as insertion operations are too slow or memory usage is too high. Each cloud is instead simply appended to the end of an array stored in a GPU-owned vertex buffer object; this reduces the time spent storing new points each frame, but prevents checking for duplicates and thus much memory and rendering time can be wasted when previously-visited areas are captured again or subsequent frames overlap.

To combat this, each cloud is also appended to a CPU buffer controlled by the *Point Cloud Library* (PCL)[7]. The GPU buffer is allocated at a fixed size, and once it becomes full it triggers a filtering event; this can take several seconds, and so this begins by copying all of the buffer's contents to a back buffer which is swapped forward so that the existing cloud can continue being rendered. A separate GPU buffer will begin receiving any newly captured points so that they can also continue being rendered and aren't lost once the main buffers are swapped back, making this process completely invisible to the user other than a slight drop in the application frame rate.

The filtering process itself involves storing the PCL cloud in a voxel grid structure, which divides the 3D space into uniformly-sized voxels and combines all points within each to remove duplicates. This also enforces a minimum density on the 3D reconstruction, drastically reducing the size of the full point cloud. The original GPU buffer's size is repeatedly doubled until it is large enough to contain this new cloud, which it then receives as well as any new points that were captured during the filtering process. This buffer is then swapped forward to be used as the main render object, just as it was before the filtering event was triggered.

Using this filtering technique, surprisingly large environments can be captured and rendered by the application while maintaining interactive frame rates, with the maximum recorded reaching 13 million points before the Android operating system refused to allocate the application more memory. Notably, this was an operating system restriction and not a hardware one, suggesting that even larger spaces could be captured on even the Lenovo Phab's modest hardware. This is also many more points than would reasonably be required during regular usage, as the two-storied building shown in Figure 5.5 only consisted of six million points in total. While a desktop-based system could in theory store a larger cloud, they could not hope to capture such a large area as they would be restricted by the length of their cables or the sizes of their outside-in tracking spaces.

[7] www.pointclouds.org

Figure 5.5: An example of a two-storied building that was reconstructed in real time using Mobileportation. Each floor is constructed separately for the sake of clarity, but the application is capable of producing this entire point cloud in one session, which would be difficult with traditional desktop systems tethered by cables and tracking spaces.

5.2.2 Panorama Acquisition

While 3D reconstructions can provide important spatial context to the environment, the low resolution, short range, and other inherent drawbacks of ToF sensors can result in crucial information being either missing or incomprehensible. For situations where this is undesirable or the remote user doesn't want to wait for incremental construction, Mobileportation also allows users to egocentrically view live capture from a 360° camera which can view the entire environment immediately with higher fidelity. Similar to the framework in chapter 3, this is achieved using a Ricoh Theta S connected via USB and controlled by the UVCCamera library[8].

Each image captured by the Theta is displayed on a sphere mesh with the principal point of the camera as its central point. This sphere has a radius of two metres, allowing any sufficiently close 3D data to remain visible and provide parallax to the user's movements. Images are captured in the dual-fisheye seen in Figure 3.12 and so again require projection to equirectangular format using the method previously described in subsection 3.4.5, but using a unit vector to each of the sphere's vertices for the calculation rather than the user's gaze vector. As only one method of panorama construction is supported, no intermediate buffer is required to store it and so this projection and the subsequent unprojection are now performed in the render shader.

Objects within the point cloud and the panoramic capture are unlikely to align when the remote user wanders too far from where they were captured, so to combat this the sphere is only rendered when the two users are sufficiently close. To aid in this transition the panorama's opacity a is exponentially increased along the curve

[8]https://github.com/saki4510t/UVCCamera

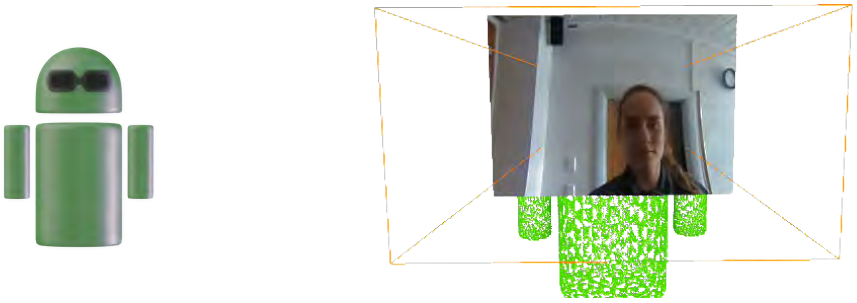


Figure 5.6: (Left): The model used to show each user's position and orientation within the virtual space. It is first converted to a point cloud through offline subsampling to maintain visual coherence with the rest of the environment. (Right): The model once rendered into the environment. The user's face is captured and overlaid to allow face-to-face communication, and a frustum is shown to indicate their gaze direction.

$a = \frac{e^{3d}-1}{e^3-1}$ as the distance d between the two lessens, losing full transparency at one metre and becoming fully opaque at 10cm. Any movements made toward each other will also be exaggerated using this same exponential curve to smoothly guide users into the egocentric viewpoint, with the two "snapping" together once 10cm apart to ensure any small movements don't accidentally return them to the exocentric view.

5.2.3 Avatar Creation

Each user's position and orientation within the environment is represented as the three-dimensional avatar in Figure 5.6, which is first converted to a point cloud offline using basic subsampling in order to maintain visual coherence with the rest of the environment. A frustum is also rendered in front of this avatar to give a general idea of where each user is facing; this frustum matches the field of view used for rendering and so will accurately show what is currently in that user's view, though is incapable of representing more fine-grained gaze information.

In traditional telepresence applications, each user typically only has one non-panoramic camera and so must make a choice between showing their environment or their face. Each has its advantages; views of collaborative task spaces can provide additional opportunities for conversational grounding (Fussell et al., 2000), while viewing the communication partner's face can provide helpful emotional and conversational cues (Flor, 1998). Mobileportation requires no such compromise as both will be visible to the 360° camera at all times. This facial capture is isolated from the rest of the scene and rendered as a quad overlaid on the user's avatar by using its centre as the point of unprojection for the algorithm outlined in subsection 3.3.2. If at any point users enter

the egocentric view this quad is instead displayed in the top-right corner of the screen to ensure it remains visible while their avatars are not.

Users will often be repositioning their phones to adjust their viewpoint and so it is unlikely that their face will always be within the same position relative to the camera. To combat this, each user's face is tracked using OpenCV's[9] implementation of Haar-cascade detection (Viola and Jones, 2004); to reduce latency this is performed on the unprojected fisheye image as it is unlikely that users will be viewing their device from an extreme enough angle to introduce any significant distortion. Once a bounding box containing the face is found, the latitude and longitude of its central point are calculated and used as the centre of unprojection for the algorithm in subsection 3.3.2. As it is likely that other faces will be captured by the 360° camera, it is assumed that the user's is the closest to the centre of the forward-facing camera, with subsequent frames always choosing the face closest to the previous frame's facial coordinates. This process is unfortunately quite slow and so is only performed every half second, though it is unlikely that users will move their head dramatically in this time so this proves sufficient.

5.2.4 Rendering

Most systems tend to create a mesh from captured point data (Piumsomboon et al., 2017), however I instead opt to keep the cloud in its unprocessed form and render each point as its own `GL_POINT` primitive. This keeps latency as low as possible as no processing is required from when the point is coloured to when it is rendered, but has the unfortunate side effect of leaving holes where insufficient data has been captured.

The other main limitation of this approach is the sheer number of primitives that will need to be rendered once a sufficiently large cloud has been captured. To mitigate this, a random noise texture is generated, with each texel assigned a random floating point number in the range $[0.1, 1]$. For each point in the vertex shader the texel at $(xz - \lfloor xz \rfloor, yz - \lfloor yz \rfloor)$ is then sampled using its absolute coordinate vector (x, y, z), ensuring each point samples the same value every frame. The sampled value is then multiplied by the distance to the far plane to give a maximum distance at which each point can be rendered. This gives the effect of "thinning" objects that are far away where detail becomes less important, though capping the minimum random value to 0.1 ensures that these objects will still be rendered in part and visible as long as they

[9] https://opencv.org/

are in front of the actual far plane. As the texture is only generated once and every point will always sample the same value, no shimmering is introduced as it would be by other random sampling methods.

To prevent this thinning from creating visible holes in distant objects, each point's rendered size is gradually increased as it gets further from the camera. This reduces the perceived resolution of these far objects, though the fixed pixel count of the display means this happens naturally anyway and so this optimisation is almost completely invisible to the user.

As well as standard first-person views, the system again supports use of a mobile HMDs. As there is now depth to the rendered environment, proper stereoscopy is supported, though this requires rendering the scene twice per frame and so has an unfortunate detriment to performance. The requirement that users walk around their physical space also made collisions with it common, so this feature was not made a focus of the system.

5.2.5 Networking

Networking is again performed through Google's implementation of WebRTC. The same server and matchmaking method are used as from subsection 3.2.3, though support for NAT traversal through STUN and TURN servers has been added to allow for connectivity over more strict connections such as mobile networks. This has resulted in the application successfully establishing international connections with Australia as well as a domestic connection between Dunedin and Auckland.

Mobileportation's networking architecture differs from the framework in chapter 3 in how the various channels are utilised. Each peer again has dedicated audio and video channels which transmit the user's voice and 360° video feed respectively, though this time no synchronisation with the data channel is required as the relationship between frames and their associated pose are not as temporally sensitive. This pose information is still transmitted via the data channel, though it now is also used to transmit each frame's point cloud. This is coloured and transformed into world-space before being sent to avoid unnecessary computation on the receiver's device.

5.3 Other Explored Features

Though I have shown that mobile phones are capable of many advanced features previously relegated to desktop computers, unfortunately their modest computational abil-

ities mean that they are still limited in what they can process in real time. With the amount of data captured by the various cameras included in Mobileportation's hardware configuration, much more could be done to further immerse users and provide better representation of their actions within the shared space, including changing their view based on the position of their head, providing more accurate indication of their gaze direction, allowing fully three-dimensional gestures, or even capturing and displaying their full body in place of their virtual avatar. These were explored and found infeasible on current consumer hardware, though their implementation is detailed here, both to prove that they are possible on a single hand-held device and so that these features may be included as soon as mobile phones become powerful enough to support them.

5.3.1 Gaze-Based Rendering

To spatially render each user's face, its position within the 360° video must be found and its latitude and longitude on the environment sphere calculated so that it can be correctly rendered without distortion. Using this information, it would also be possible to calculate a vector from the user's face to the display. This not only gives the angle at which they are viewing the device from, but also the angle from which they're attempting to view the environment. Rotating the render camera by this angle could thus allow the user to obtain novel views not just by rotating their device but by moving their head, letting them use their device as a "smart window" or "transparent display" (Andersen et al., 2016) into the environment, increasing their spatial understanding of the objects within it (Kruijff et al., 2010).

This angle could also be used to adjust the direction of the user's gaze indicator, giving a more accurate approximation of where they are currently looking; future improvements in camera quality could even make fine-grained eye tracking possible, reducing this indicator down to a single point. A rendition of how this could look is shown in Figure 5.7.

Though it sounds simple to calculate, this process is unfortunately made infeasible in real time due to how long it takes to retrieve information from GPU shaders. Calculating the latitude and longitude of the user's face is currently performed within the render shader, but for this information to be used elsewhere necessitates a separate compute shader at the beginning of the render loop. As this information would be needed in every subsequent render shader to keep the view direction consistent between them, the result would need to be copied to the CPU, which can take up to 15ms

Figure 5.7: An example of how the user's gaze direction could be used to affect their view of the environment. The angle between their face and the display is calculated and used to offset the angle of the render camera, allowing the display to be used as a "smart window" into the environment. Their gaze frustum is also offset by this same angle, allowing more precise indication of where they are currently looking.

for even small amounts of data. Though it sounds negligible, this extra computation would instantly halve the application's overall frame rate, even before any points have been captured.

Even if this calculation was fast enough, HAAR cascades are too computationally demanding for real-time use on modest mobile hardware. In the current implementation faces are only tracked twice per second, which would mean the user could only adjust their view at this same rate. Given the current 60fps tracking, such changes would be too jarring, either reducing the rate at which they can move from 60 times per second to only two, or making their view move twice per second for reasons that would probably seem arbitrary to them. Given this, it seems more sensible to wait for hardware revisions to make this process smoother and thus more desirable for the user.

5.3.2 Reintegration of Gestures

In the initial telepresence framework the user's hands were captured, isolated, and rendered over the environment to allow for rich gestural interaction between peers. This was afforded by the relatively simple two-dimensional environment presented to them, which meant 2D gestural capture was sufficient in ensuring these gestures were presented in a visually consistent manner.

The shift to 3D environments means that such simple captures are no longer acceptable. All pointing and representational gestures require the proper spatial context in order to be correctly interpreted, but two-dimensional capture would make this con-

Figure 5.8: A user performing a pointing gesture within 3D space. Their hand is found within the latest point cloud capture using colour thresholding and only rendered temporarily so that gestures can be used without affecting environmental reconstruction.

text impossible to preserve; displaying them on a plane in front of the user would make them difficult to see when looking at their target, and somehow projecting them into the environment could lead to incorrect assumptions about what the target of the gesture was and thus incorrectly portrayal of its meaning.

These gestures could be captured in 3D through the Lenovo's integrated ToF sensor, however this introduces the issue of how it should be rendered. Any process that attempts to detect or reconstruct any hands in the depth map need to be completed before the point cloud is recorded to prevent them being permanently shown in the environmental reconstruction, and so must be fast enough to prevent noticeable latency being introduced to the capture process. This rules out any mesh reconstructions such as used by Sodhi et al. (2013) as a complicated search through unsorted depth data would be far too slow for this narrow window, though faster hardware could allow this in the future.

The best method possible on current mobile devices is thus to simply render the hand as it is captured without further processing. This is problematic when there is only one capture point as parts of the gesture can be easily occluded by the back of the hand, in particular the indicating finger if the user attempts to point at too low an angle.

To determine whether this would be an issue, I integrated a proof-of-concept gesture tracker into Mobileportation that works similarly to that outlined in the original framework. When colouring the latest point cloud captured by the ToF sensor, each time a point is projected into the RGB image it is checked whether the resulting pixel falls within the skin-colour range defined in subsection 3.3.5 with the additional constraint that it has to be within a metre of the sensor. If these conditions were met, a flag bit within the byte used to store the point's alpha value is set, and the point

is copied to a separate buffer which is only temporarily rendered until the next point cloud is captured. This prevents the hand from being permanently recorded while keeping it visually consistent with the rest of the 3D reconstruction.

While this allows primitive gestures to be used in conversation, it was found that the user's hand is always too close to the device to properly align the depth and colour cameras, resulting in unaligned areas being permanently recorded in the environmental cloud. Pointing also has to be performed at an unnatural angle to ensure the indicating finger isn't occluded by the back of the hand, and as seen in Figure 5.8, the sparseness of the cloud and the absence of the occluded side of the hand makes many gestures too abstract to properly interpret, meaning gestures unfortunately cannot be supported with current hardware.

5.3.3 Full Body Capture

Another solution to incorporating gestures would be to capture the user's entire body, which would not only allow hand-based gestures but would also convey more subtle body language that can often aid in comprehension between interlocutors (Flor, 1998) and increase the user's sense of embodiment in the space (Fribourg et al., 2020). This has been attempted in a semi-mobile form factor in the past (Xu et al., 2019), though this required a 360° camera mounted on a head-worn cap to ensure the user is in view. This becomes easier with Mobileportation as the user is already carrying a 360° camera in front of them, making this body capture possible with the existing hardware setup.

Several methods of body capture were attempted. The first was to use OpenPose (Cao et al., 2018), a library that provides full body tracking on the CPU. This unfortunately proved too slow for real-time use, taking several seconds per image, and so was quickly abandoned. The second was TensorFlow Lite[10], which is a light-weight version of the TensorFlow deep learning framework intended for usage on mobile devices. This was slightly faster and has previously been demonstrated to work in real time, but still takes several seconds per image when integrated into Mobileportation due to the existing heavy contention for system resources.

If body tracking was feasible there are several ways it could be integrated. The user's body could be directly displayed, but since the 360° camera only captures 2D data it would need filled in in some way to maintain consistency with the 3D environment. The other would be to use the position of the user's limbs to orient those of their avatar,

[10]https://www.tensorflow.org/lite/

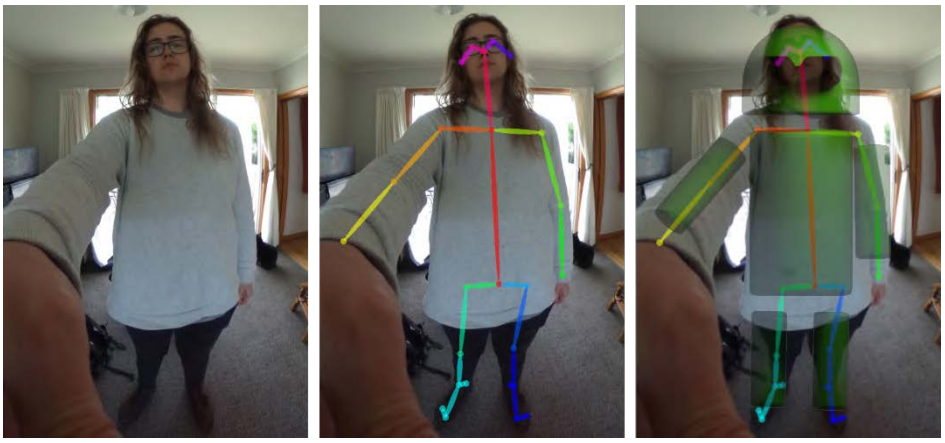


Figure 5.9: An example of how the user's body can be tracked within the 360° video captured by the Ricoh Theta. (Left): The user as seen by the 360° camera. (Center): The user's skeleton detected within this image using OpenPose. (Right): The position of the user's limbs are used to manipulate their 3D avatar, allowing their actions to be properly displayed within the environment.

as shown in Figure 5.9, though this would not provide the same sense of embodiment (Fribourg et al., 2020) and would obfuscate any subtle body language.

5.4 Evaluation of Requirements

As with the framework in chapter 3, we must now determine whether this new system satisfies the requirements laid out in subsection 2.3.6 that a telepresence system must meet in order to provide a satisfying, presence-inducing, and efficient means of communication with a remote party.

The first of these is that the environment presented to users must remain consistent and symmetrical between them, and that this environment must also facilitate views of both the users and the task space. As with the previous framework, all orientation matrices are shared between users and used as the basis for all reconstruction and rendering. Whenever they are altered, for example to facilitate the "snap in" mechanism, this is done before the matrix is transmitted or used for rendering so this alteration will be reflected on both ends of the connection. Mobileportation improves on the previous framework when it comes to incorporating the user into the environment, as the use of avatars and facial capture ensures that users will always be visible within the environment, making it so they can both be viewed simultaneously in as natural a way as possible.

Mobileportation also improves on the previous framework in regard to the second

requirement, which is that users should be able to obtain independent views within the environment. Previously each could rotate their device to explore panoramic spaces with 3DoF, but the remote user was always locked to the local user's position and thus did not have full freedom within the remote space. Mobileportation introduces translational movement thanks to its 3D reconstructions, which will provide more means of pulling information from the environment without intervention from the local user and thus a more presence-inducing experience for its users.

The third requirement is that users should always be aware of what their partner is doing through conveyance of their position and gaze direction. This proved simple for the previous framework as users were always co-located and so this information was already known to them without further indication. In Mobileportation this became somewhat more difficult as users could now freely change their position with the environment, so the 3D avatar was included to make this information clear. The avatar was made intentionally asymmetrical to give an indication as to which direction each user is facing, which is made even clearer by rendering their current field of view as a frustum in the direction of their virtual camera.

It is unfortunately in the fourth requirement where Mobileportation falters, as no means of performing gestures is provided to its users. A primitive method of doing so was attempted, as described in subsection 5.3.2, though it was found that the current resolution of the camera and computational capability of the mobile phone are insufficient in producing these gestures in 3D space in a convincing way. Perhaps in several years gestures will be possible, however for now they will have to unfortunately be omitted.

The fifth and final requirement is that all features of the proposed application be possible using purely mobile devices to maximise its availability and appeal to the wider population. This is again achieved, with all computation performed on the mobile phone itself, though an external 360° camera is required to capture the panoramic video and each user's face. This camera is still portable and affordable enough for general use, but is still an extra device that users would need to carry with them and so ease of use unfortunately suffers. However, recent trends in mobile phones have seen integration of both depth sensors and wide-angle lenses, so it feasible to believe that all of the features Mobileportation offers will soon be available in a single hand-held consumer-grade device.

Chapter 6

Evaluating the Mobile Experience

With almost all of the requirements for effective collaboration laid out in subsection 2.3.6 met in a mobile form factor, it remains to be seen whether this is sufficient for providing users an efficient and intuitive means of communication. A technical evaluation of this new application is thus detailed to evaluate whether this new interaction method is achievable in real-time on mobile hardware, as well as the results of a user experiment gathering feedback about the system's viability as a communications tool.

6.1 Technical Evaluation

One of the main goals when developing Mobileportation was to ensure real-time frame rates and negligible latency so that users would have the seamless experience they are accustomed to from existing videoconferencing solutions. As seen in Figure 6.1 this was achieved for all but the largest environments, with the application rendering at the full 60fps afforded by Android's enforced V-sync before any points have been captured. This may seem trivial given there's no environment to render, but the application must still provide positional tracking, face detection, 360° video capture, and rendering of these various components.

The application's frame rate decreases almost linearly as more points are captured, reaching 15fps at five million. This is approximately what was required to capture the two-storied building in Figure 5.5 which will likely be a rare case, so typical use will probably see average frame rates of 20-30fps. It's important to note that this is much higher performance than many desktop systems utilising multiple discrete GPUs such as PanoInserts (Pece et al., 2013), SLAMCast (Stotko et al., 2019), and the work by Gao et al. (2017).

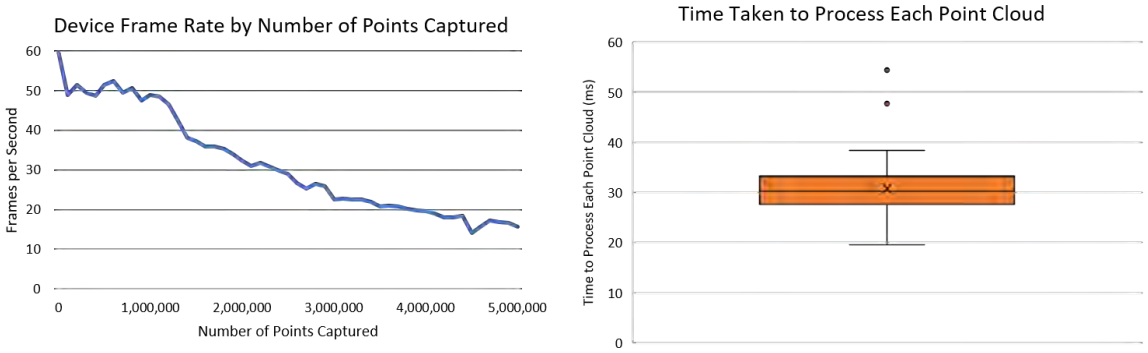


Figure 6.1: (Left): The application frame rate as a function of the number of points captured. More points result in a linear decrease in performance, though it remains interactive for all but the largest of captures. (Right): The time required to process each point cloud once it is captured. This averages 30ms in almost all cases, suggesting that 30fps capture would be possible with a more capable ToF sensor.

The Lenovo Phab 2 Pro's ToF sensor used for environmental capture is limited to only five captures per second, providing an upper bound on the application's overall capture rate. Each frame can theoretically contain up to 30,000 points but typically averages 5,000, and once captured takes 31ms to be transformed to world space, coloured, and uploaded to the GPU. This low processing time allows the application to easily process every captured frame, which in theory could provide 30fps environmental capture with a faster ToF sensor without requiring any other hardware or algorithmic changes. This also applies to the Ricoh Theta's limited 15fps capture, which could be rendered at 30fps given a more capable camera.

This short processing time also ensures that latency remains low, with only 31ms between when a point is captured by the ToF sensor and when it is ready to be rendered. The user's pose is also tied to this process, so any movement will require the same 31ms before being shown in the virtual space. The network still sees the same worst-case scenario of 256ms over local connections measured in section 4.1, so the remote partner will see the user's movements or captures within 287ms of when they make them. This latency will be longer over wide-area connections, though this is unfortunately unavoidable in current networks.

6.2 User Evaluation

With almost all requirements met, we can now determine whether this system proves a natural and presence-inducing means of communication for its users. The previous

framework proved capable in this regard, with previous work showing its ability to induce a sense of presence within the shared space (Jo and Hwang, 2013), and so even parity with the previous framework would prove this system's capabilities while opening up remote mobile communication to use cases never before possible.

I had the following hypotheses when developing Mobileportation:

1. The spatial presence induced by Mobileportation would not be significantly different from that induced by conventional 360° videoconferencing despite the reduction in visual fidelity and immediacy of information introduced by incremental 3D capture.

2. Showing each user's pose and facial capture via a 3D avatar would provide a higher sense of co-presence than unsituated facial capture alone.

3. Mobileportation would induce a higher sense of social presence and thus a more social experience for users than conventional 360° videoconferencing.

4. Mobileportation would be overall preferred to conventional 360° videoconferencing.

To test these hypotheses, a user study was conducted with novice users. As the system was developed with these users in mind, a large focus was placed on social scenarios, so to provide a more casual environment the usual industry-focused measures of task performance were omitted.

Participants were asked to test the system over a live connection with a mediator placed in a remote unprepared environment. Mobileportation was compared to video-only 360° videoconferencing in its ability to induce spatial, social, and co-presence in its users. To ensure all other factors such as the network, performance, and video quality remained consistent between these two conditions, the 360° videoconferencing system was simply Mobileportation locked to an egocentric view with depth capture disabled. Users' faces and gaze indicators were still captured and displayed, and participants still had free view control within the live 2D environment, though they could not freely move around the space as their position was locked to the local user's. No mobile HMD was used for this experiment to make facial capture possible. A comparison between standard and 360° videoconferencing has been done before (Jo and Hwang, 2013) and so was omitted for this experiment.

6.2.1 Study Design

14 participants were recruited between the ages of 18 and 65, of which eight were female. Each was gifted a $10NZD supermarket voucher upon completion of the experiment. A within-subjects design was used with two conditions, where the independent variable was the application used: either video-only 360° videoconferencing or Mobileportation. The order of conditions was randomised for each participant to minimise potential learning effects.

Each condition consisted of an informal guided tour through one of two floors of the rental property in Figure 5.5. The floor explored was randomly assigned per condition for each participant to minimise the effects the contents of each may have had on the participant's engagement with the space. A reconstruction of the explored property is shown in Figure 5.5.

Spatial presence was measured using the IPQ questionnaire by Schubert et al. (2001), while social and co-presence were measured with questionnaires by Biocca et al. (2003), Bailenson et al. (Bailenson et al., 2005), and Hauber et al. (Hauber et al., 2006). Each consisted of statements about the user's experience while using the relevant system, with the participant noting the degree to which they agree with each statement on a 7-point Likert scale with varying anchors. Space was left at the end of the questionnaire for participants to leave free-form comments about their experience. An additional task of sketching the explored property and evaluating the mental workload to do so was also initially included, though pilot testing found that this distracted too much from the social aspect of the experiment and so was removed. These questionnaires can be seen in Appendix B.

To gauge which system participants preferred overall, a post-experiment questionnaire was also completed after both conditions had been tested. This consisted of the following questions, with space allocated after each so that participants could justify their decision:

1. Which of the two systems did you find easiest to use?
2. Which of the two systems made you feel more 'present' in the virtual environment?
3. Which of the two systems made it feel more like the remote partner was present with you?
4. Which of the two systems did you prefer overall?

This experiment was conducted with the approval of the University of Otago Human Ethics Committee (Non-Health).

6.2.2 Procedure

During each condition, participants were connected to a remote study mediator who was physically located within the rental property several kilometres away via Wi-Fi or 4G. They were first given a brief overview of how to operate the current condition's application, then given five minutes to familiarise themselves with it within a designated room of the property; the same room was used for both conditions to highlight the differences between the two applications. They remained connected to the study mediator for this so they could experiment with live video and orientation data, though their cameras and microphone were disabled so that their actions could not be seen and they could freely explore without worry of being observed.

Once comfortable with the system's operation, a brief and informal tour of a randomly selected floor of the property was conducted. The mediator would pretend the participant was interested in renting the property and show them through various rooms while describing their contents, though participants were encouraged to explore on their own and could (and often did) ignore the mediator entirely. All video, audio and tracking data was captured and transmitted in real time so that the mediator could appropriately react to participant comments or requests.

Participants were encouraged to ask the mediator to revisit areas they wished to see more of, particularly in the 360° videoconferencing condition where they couldn't revisit it themselves. No specific task was set other than experience the system in a realistic social scenario. Once the relevant floor had been extensively shown, taking approximately five to ten minutes, participants were asked to complete the presence questionnaire. This process was repeated for the remaining condition on the other assigned floor, after which participants completed the post-experiment preference questionnaire before being gifted the $10NZD supermarket voucher.

6.2.3 Results

The presence questionnaires indicated that the amount of spatial presence induced within the shared environment was rated similarly for both Mobileportation ($\mu = 4.30$, $\sigma = 0.86$) and 360° videoconferencing ($\mu = 4.27$, $\sigma = 0.79$), with a Wilcoxon signed rank test ($N = 14$, $\alpha = 0.05$) showing no significant difference between the two condi-

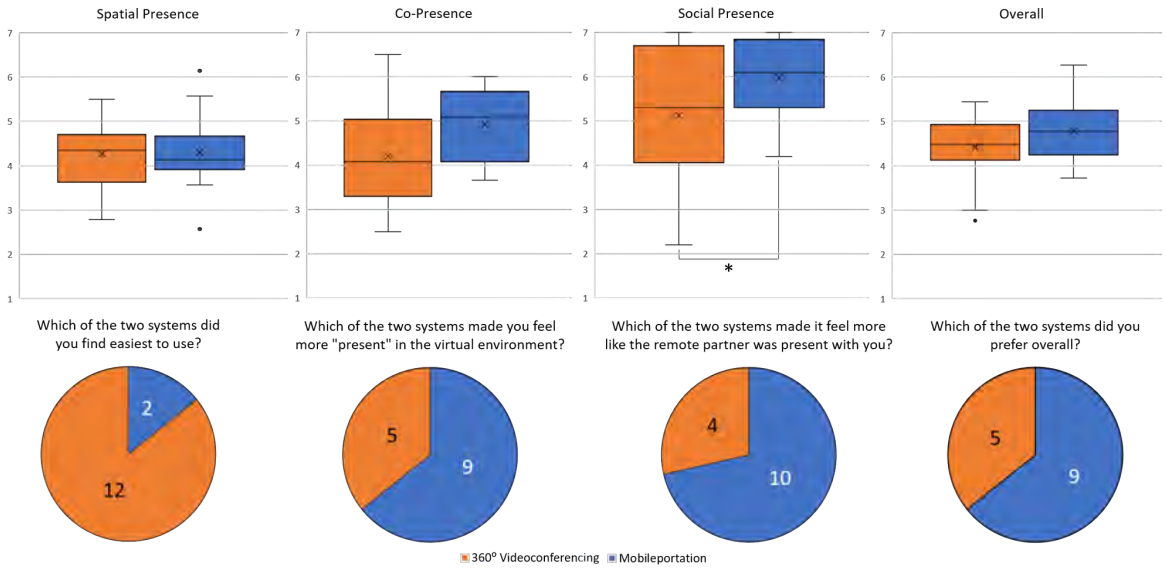


Figure 6.2: The results of the user study comparing Mobileportation to conventional 360° videoconferencing. The top row shows the results of the presence questionnaires for each condition, while the bottom shows the proportion of participants that preferred each system in various categories according to the post-experiment questionnaire.

tions ($p = 0.59$). The co-presence induced between communication partners was rated slightly higher in Mobileportation ($\mu = 4.93$, $\sigma = 0.81$) than in 360° videoconferencing ($\mu = 4.20$, $\sigma = 1.10$), though another Wilcoxon signed rank test revealed this difference to be insignificant ($p = 0.12$). A significant difference was however found in the social presence induced between the two parties ($p = 0.03$), with Mobileportation ($\mu = 5.99$, $\sigma = 0.90$) providing significantly more than 360° videoconferencing ($\mu = 5.13$, $\sigma = 1.57$).

The post-experiment questionnaire revealed that despite twelve of the fourteen participants (85%) finding Mobileportation more difficult to operate than 360° videoconferencing, nine (64%) still chose it as their preferred system overall. This may be due to how ten (71%) felt that it provided a better sense of being with their remote partner, or nine (64%) claiming that it made it feel more present within the shared environment.

6.3 Discussion

In developing Mobileportation I believed that it would be overall preferred to anything currently available on mobile devices, of which the most immersive experience is currently 360° videoconferencing. This proved to be the case, though not to the

degree expected as only 64% indicated as such. Those that did choose Mobileportation cited the fun and unique experience they had during the tour as well as the sense of presence it invoked within the shared environment. Conversely, every participant that chose 360° videoconferencing as their preferred system stated that the low visual fidelity was the main contributing factor, so future improvements in reconstruction quality could see more participants preferring the Mobileportation experience. Here I further explore how these two competing factors influence all other aspects of the user experience as well as the unique problems and opportunities presented by this new, untethered interaction method.

6.3.1 Spatial Presence

My first hypothesis was that the 6DoF exploration afforded by Mobileportation would induce a similar sense of spatial presence within the shared environment as 3DoF exploration of a panoramic one, which was confirmed by the experimental results. While this hypothesis may seem unambitious at first, it makes sense if the visual fidelity of each environmental representation is taken into account. With 360° cameras the entire environment is visible and explorable as soon as the application starts, and moving to new locations does not require this information to be recaptured. On the other hand, the incremental reconstruction required for exocentric viewing means that large parts of the environment will usually be absent, especially if the remote user tries to venture too far from the local one. All five participants that chose 360° videoconferencing as the system which induces more spatial presence specified the lack of immediacy of information as the main contributing factor, with some confused by the environment being "fragmented", "patchy" and "not fully rendered", one finding the system "hard to use since the image was often broken into pieces", and another stating that "3D would have been awesome if I could perceive my surroundings without the need to wait".

This interestingly contradicts the findings of the previous experiment, where it was found that a user's perception of presence within an environment shouldn't be affected by whether it is live and complete or static and incrementally constructed. It could be that the shift to 3D environments exacerbates the downfalls of incremental reconstruction as the areas not visible to the remote user become much larger. On the other hand, this could be further reinforcement of this previous finding, suggesting that it still holds true even when the incrementally built environment is more complex and explorable.

Participants also felt that the quality of the reconstruction was too low for them to

fully immerse themselves within the space. All five that preferred 360° video overall to Mobileportation indicated that this was the only reason for their decision, feeling that "video only was more reality [sic] than the 3D model" and that "seeing the quality of video image to 3D, it is very less fascinating than video only", but that "if the 3D was better for rendering then the in world feel would definitely be there".

The ability to transition to an egocentric view was originally implemented as a way to offset this lower fidelity by giving users a way to view the higher-resolution video. Despite this, participants performed this transition only sparingly, and only ever when entering a new room. As soon as the room had been partially reconstructed they would always favour viewing it from an exocentric position, and would not re-enter the egocentric view until they moved to the next room, consequently almost always seeing the environment in its 3D form. This could be due to many participants not remembering this ability to transition existed, as despite this functionality being extensively explained and tested by them during the training period, many became surprised if they accidentally triggered it when attempting to walk past the mediator.

Despite the lack of immediate, high-fidelity reconstructions, Mobileportation still provided a comparatively immersive experience to 360° videoconferencing. This is likely due to the ability it gives users to freely and independently explore the shared environment offsetting the reduction in visual fidelity, which led to an "overall more enjoyable", "more seamless", and "more fun and immersive" experience where participants would often abandon the study mediator in favour of self-guided navigation. One participant in particular would often wander off to look at a collection of video game consoles, which led them to believe that "my results may be biased, I was looking at the gaming stuff as opposed to looking at my partner". While problematic for the experiment, such an experience would be completely impossible with existing systems as this collection would likely fill the entirety of the explorable area. Another would often "forget that the world is virtual and that building structures and furniture don't actually exist. For example, I continued to walk around the bed until I realised that I didn't need to and I could just walk through it", suggesting a degree of autonomous interaction with the remote environment usually relegated to fully-immersive HMDs despite viewing the space through a non-immersive mobile phone display.

Participants that used Mobileportation in the first condition missed the freedom to explore when locked to an egocentric position in the second, with one stating that "without being able to physically move around I felt it's more restrictive. I prefer [Mobileportation] where the room is rendered out despite the video being clearer [in

this condition]," and another saying that "the thing I missed in this experience is that I do not have the freedom to move and interact with the virtual world [and thus] it felt more like a virtual tour." With this enhanced freedom to explore and interact within the remote space only inhibited by the quality of the explorable environment, it is feasible to believe that hardware improvements and subsequent increases in reconstruction quality could lead to Mobileportation inducing significantly more spatial presence than what is currently available on mobile phones with no changes required to the underlying algorithms.

6.3.2 Co-Presence

My second hypothesis was that spatially-rendered representations of each user within the environment through 3D avatars and facial capture would significantly increase the co-presence felt between them than statically-situated facial capture alone. Unfortunately, the presence questionnaires indicated that this is not the case, even though 71% of participants claimed otherwise in the post-experiment questionnaire.

The most likely explanation for this is that the 360° videoconferencing implementation shows constant representation of each user in the corner of the display, whereas during 6DoF exploration users can only see representation of their partner when they were directly looking at them. This led to participants having trouble locating the tour guide at times after abandoning them for independent exploration, with one participant having "difficulty as sometimes I was not able to catch [the mediator] or not able to spot him," and another finding it "hard when [he] moved without me knowing, finding him again was confusing at times" and that it was "easier to track where the orange box [the gaze frustum] is."

This lack of constant representation was exacerbated by an entirely novel problem introduced by this system: users would often lose each other. Participants would abandon the mediator for independent exploration, and so would often be in entirely separate rooms and lose track of where the mediator had gone. As there was no indicator to guide the two back together, a repair step would occur where the mediator would guide the participant to their location. This is a problem completely unique to this form of telepresent interaction: in existing systems with limited tracking spaces users may only ever be several metres apart from one another, so this new independent exploration could be somewhat of a double-edged sword in its current iteration. This could have been the cause of the lack of increase in co-presence as some participants spent long periods of time alone and only vocally interacting with the mediator.

Though difficult to keep track of, participants still found that the spatial rendering of their partner's position was helpful in providing a sense of co-presence with them. Many stated it made it feel "like the remote partner was in the same room compared to just seeing the face all the time" and that it "allowed for a feeling of 'being there' with the person rather than being on call with them" as they "know what my partner looking at and where he is not just where his camera showing". Participants also felt that the mediator's avatar gave more indication as to their current actions and position relative to the environment, which could have also contributed to a sense of spatial presence as one participant noted that "I could walk around with [the mediator], which made me feel like I was there rather than just moving my hands most of the time."

With these benefits, it seems advisable for future systems to use a hybrid approach where simple facial captures are shown in the corner whenever the partner's avatar moves out of view, with the avatar becoming visible whenever the user looks towards their peer. An indicator such as was used in the previous framework could also be used, highlighting the edge of the screen to show the direction the user must move in order to find their partner. This would provide the benefits of spatially situated rendering, but would also serve as a constant reminder that the remote partner is there in the virtual space and make independent exploration less of a lonely experience.

6.3.3 Social Presence

My third hypothesis was that the enhanced interaction afforded by Mobileportation would provide a greater sense of social presence between its users than conventional videoconferencing. This was confirmed by experimental results, suggesting that free-form exploration of a space could provide a more social experience than that of a fixed viewpoint and reduce the sense that interaction between peers is computer-mediated.

The most common complaint participants had of 360° videoconferencing was that it felt too much like a "360 visual tool" or a "virtual tour," or that it was "a bit like a 'presentation"' or "like Google street view" in comparison. These complaints seem a little unfounded considering this is exactly what they were doing; taking a virtual tour and complaining that it felt too much like one seems a strange thing to complain about. However, taken in the context of the experiment, these comments reveal a rather surprising result: namely, while exploring the property participants forgot that they were on a tour at all and the experience became something entirely different. This led to participants engaging more with the 3D environment than the 360° video, in which one participant felt "it was easy to ignore and just listen, which would be the

same as a pre-recorded video," and another that it was more like "being there while they show you something" rather than actively participating in the tour.

It is also possible that this increase in social presence could be due to a shift in how the task was performed and perceived by participants between conditions. In 360° videoconferencing, users were forced to take the tour through the flat and had no means of ignoring it. In Mobileportation, participants could abandon the mediator and do whatever they wished within the space, as exemplified by the participant who was captivated by the video game collection. When users took the tour during this condition they did so of their own volition, giving them a degree of autonomy not possible with purely 3DoF interaction.

6.4 Summary

In these past two chapters, I have presented and evaluated a completely novel experience on mobile phones: the ability to freely explore a remote partner's location in real time with six degrees of freedom whilst interacting with them within it in a natural and intuitive way. This was facilitated by Mobileportation, an application that allows incremental 3D reconstruction and streaming of the local user's surroundings with some remote partner, who can freely roam around this virtual space by walking around their real one. Face-to-face communication is possible within this space through facial tracking and rendering over spatially-placed avatars, and the remote users can see high-resolution live and complete views of the environment by simply walking towards their partner.

Crucially, an evaluation of this application's performance showed that this experience is achievable in real time with purely mobile devices. Latency also remained low, even on a four-year-old smartphone, ensuring a smooth experience for users.

Experiments on novice users have also proven the application's feasibility as a natural and efficient communications tool. Spatial presence proved similar to 360° videoconferencing, which was previously the pinnacle of what was achievable on a mobile phone, and future improvements in hardware could see this increasing further to provide an even greater sense of presence in remote spaces than was previously possible with no changes in the system's underlying algorithms.

Co-presence was similarly rated, with users appreciating the ability to see their partner's location through their 3D avatar, but easily losing them due to the lack of indicator of their position once they went out of view. Despite this tendency for users

to lose each other inhibiting the possible co-presence between them and violating my second hypothesis, this shows the benefits of untethered exploration as interlocutors can inhabit entirely separate spaces where previously they could only ever be several metres apart.

Social presence was the only form of presence to see any significant increase due to the wealth of interactions afforded to users. Giving participants the ability to freely explore the shared space rather than be dragged through it provided a much more fun and social experience, resulting them in seeing the virtual tour as something else entirely. Users enjoyed the ability to ignore the mediator and explore the rental property at their own pace, with many ignoring the mediator in favour of self-guided exploration of the contents that interested them most.

Chapter 7

Conclusion and Future Work

Conversations require a certain degree of common ground to be established between speakers before they can be efficient. Without this shared knowledge, communication becomes meaningless and frustrating as each constantly clarifies the meaning of their previous statements that they assumed would be immediately understood. Talking about an object only works if all involved know exactly which object is being talked about, and giving instructions only works if all involved know who should perform them and what they should be performed on. If this knowledge is not already shared, conversation must come to a standstill as it is negotiated and explained; the more this happens, the more inefficient the conversation becomes.

This common ground does not always have to be explicitly and verbally negotiated. If the two speakers are co-located, one can simply point to the person or object they are referring to, keeping any verbal utterances short and vague without losing any meaning. Instead of asking where their partner is or what they are doing, a speaker can simply look at them to gain this information without their partner's intervention. Side-by-side communication is thus more efficient than its remote counterpart as there are more ways in which information can be shared, ensuring context is never missed and maximising the common ground shared by communicating parties.

For a remote communications system to provide a comparable experience it must thus emulate this free information sharing as closely as possible. Through a survey of experiments examining how this exchange occurs in realistic scenarios, I thus identified that such a system must allow each user free exploration within a symmetrical and temporally consistent environment, that during this exploration their current position and actions should be made known to their partner, and that each party involved may communicate to all others within a shared space through explicit gestures and subtle

body language.

Many systems exist that attempt to provide such interaction, though all that come close are severely limited by the desktop systems they are designed for. Views of the environment are often more important than views of a partner's face (Luff et al., 2003), but in targeting such a platform any environment will inevitably be limited by its stationary nature, restricting the freedom provided to users in their ability to freely explore and pull information from their partner's surroundings.

Targeting desktop systems also ignores a global shift toward mobile computing, meaning the proposed solution becomes immediately irrelevant in a world where more people own and prefer to use a smartphone than a desktop or laptop computer (Ofcom, 2018). This preference is due to the convenience and portability of these mobile devices, providing more opportunities for conversations that were previously either impossible or redundant by utilising otherwise dead moments in the person's day. Such opportunities are often taken advantage of for such purposes, with most people choosing to spend their commute talking to a distant acquaintance through their mobile device (Ofcom, 2018).

Given this, I imposed an additional requirement on future telepresence systems: that they operate purely on mobile hardware, allowing the vast majority of the population to immediately experience their benefits using hardware they already own. Not only could this drastically increase the potential adoption rate of new communications media, but it would also allow new communications media to be used whenever and wherever the user desires. This opens up previously inconceivable scenarios where a desktop system could never go, such as taking a house-bound relative to the top of a mountain or giving a tour through a museum.

7.1 Contributions of this Thesis

In this thesis I showed that mobile devices are fully capable of providing this immersive experience without relying on proprietary or external hardware, though can still benefit from its inclusion. This was done through identification of the previously described requirements such a system must meet, and then development of two separate frameworks that attempted to fulfil them while also ensuring real-time performance for a smooth and natural experience.

The first served to identify whether a "sweet spot" existed where increasing the fidelity and possible view independence within a reconstructed environment gave neg-

ligible gains in increased presence, providing a minimum floor an application must meet in these areas. The local user could create a panoramic representation of their space, which the remote user could freely look around by reorienting their device. The field of view of each was shared, and both could communicate through unmediated, spatially-rendered gestures for both pointing and representation.

Five means of creating this symmetric, consistent, and explorable environment were implemented, which between them covered the full continuum of view independence possible in such a panoramic space. Through a comparison of these by novice users, it was found that a static and incrementally constructed environment is just as effective at inducing a sense of spatial presence within it as a complete and live one such as provided through 360° video without compromising the co-presence felt between users.

Though this framework provided nominal independence between its users, it failed in two key areas: this independence was limited to rotation as users must always assume the same spatial position, and due to this co-location it was impossible for each to view the body language of their partner. As it was found that static and incrementally constructed environments are sufficient for maximising spatial presence, a second framework was developed that allows full 6DoF movement through a three-dimensional reconstruction of the local environment, a first for completely mobile hardware. Each user's position, orientation, and face are displayed via a virtual avatar, allowing conventional face-to-face communication to be enhanced through the context provided by the wider environment. The modest capabilities of mobile processors meant this came at the expense of visual fidelity, and so to see a full and high-resolution capture of the space users could transition to an egocentric view of it as captured by an external 360° camera. As only mobile devices are required for this experience, to the best of my knowledge this is the first system where truly arbitrarily large environments can be explored without being restricted by cables or limited tracking spaces.

In realising this new experience other compromises unfortunately had to be made, resulting in one of the identified requirements not being met: that hand and body language be freely shared between users. The 2D hand segmentation used in the previous framework proved insufficient as gestures became visually inconsistent with the rest of the space, and other methods were too computationally demanding due to resource contention caused by environmental reconstruction. Alternative methods such as 3D segmentation and full body capture were tested, though found infeasible in real time on current mobile devices.

Despite this reduction in fidelity, users still found this new framework induced a

high degree of spatial presence within the shared environment. Though the additional freedom to explore was not enough to significantly distinguish this system from conventional 360° videoconferencing in this regard, all participants noted that the low resolution of the reconstruction was the main deciding factor, suggesting that future hardware improvements will result in a much more presence-inducing experience than anything currently available on mobile devices. Co-presence similarly saw no significant increase, though this was due to a completely novel issue never before seen in similar systems: in their independent exploration within the large shared environment, users would often lose each other. Though an obvious limitation of the application, this is something completely impossible in existing systems where they could only ever be a few metres apart.

7.2 Future Work

Future improvements and innovations in mobile hardware could further revolutionise the ways in which communication can happen remotely. Not only could the applications outlined in this thesis run at higher frame rates, with higher fidelity, and more immersively through more sophisticated HMDs, but increases in computational capabilities could see new features becoming available on future mobile phones, and the inclusion of extra sensors and other hardware could open the opportunity for features never before possible on a mobile device.

7.2.1 Future Design Space

In section 5.3 I outlined additional features developed for Mobileportation that were unfortunately discarded due to the inability of mobile devices to handle them with acceptable performance. These were full-body capture and tracking, which could allow users themselves to be present in shared spaces rather than virtual avatars, 3D-rendered gestures, which would further ease conversation, and gaze-based rendering, which would allow the device to act as a "smart window" into the shared environment and provide more accurate indicators of each user's current object of attention. While too demanding for current hardware, these features could be reintroduced to the application within a few years, or further optimisations could be found that allow their use on current mobile devices.

Future innovation in the hardware included in mobile phones could also bring

changes to pervasive communications. For instance, the recently announced iPad Pro[1] includes a LiDAR scanner in place of the usual ToF camera. This change allows much higher fidelity 3D reconstructions than ever before possible with a hand-held device, and if this sensor's inclusion becomes popular it could result in almost everyone having an enterprise-grade depth scanner in their pocket.

Despite users tending toward mobile devices for their communication needs, future work could also explore how this relatively modest hardware can be combined with stationary systems to create a more immersive experience. For instance, while out and about users could interact using only their mobile phone, but when at home they could link it to a dedicated outside-in capturing system such as used by Park et al. (2019) to replace their virtual avatar with a live 3D scan of their body. They could alternatively create a large, low-fidelity scan of their environment to provide wider context outside of the captured area, then return to the outside-in system which would act as the designated interaction space and be captured in higher fidelity.

7.2.2 Future Research

Future work could also further hone the ideas presented in this thesis. In particular, a more extensive comparison of the various modes of interaction outlined in chapter 3 could be performed. Currently only three modes have been directly compared: Live Spatial Video Calling, Incremental Panoramic Calling, and Live Panoramic Video Calling. While this proved sufficient in identifying how independently explorable a space must be to induce a sufficient sense of spatial presence within it, a similar sweet spot for co-presence and social presence has yet to be defined. We know that the degree of view independence has some kind of effect on co-presence due to the significant differences seen between Incremental Panoramic Calling and Live Panoramic Video Calling, it is not known whether this increase is due to the environment being immediately available to fully explore, as in Panoramic Calling with Live Inserts, or the environment being captured live, as introduced by Live Panoramic Video Calling.

The experiments described in this thesis also had a heavy emphasis on socialisation in shared spaces and thus excluded the usual industry-focused metrics of task completion time, task accuracy, and task efficiency. Future experiments could reintroduce these measures: for instance, they could be evaluated for each of the modes of interaction in chapter 3 to determine how much of an effect view independence has on the

[1] https://www.apple.com/ipad-pro/

time taken to complete a collaborative task, or how complete a shared environment must be in order to complete a task within it.

Pilot tests of Mobileportation included a set task where users would have to sketch the environment presented to them after the experiment, with the workload required to do so evaluated using the NASA Task Load Index (TLX)[2]. This was originally excluded from the final experiment as it distracted too much from its social aspect, though it could be reintroduced in future to determine whether participants have more spatial awareness within an incomplete but three-dimensional environment than they do a complete but two-dimensional one.

7.3 Conclusion

With these findings I have showed that mobile phones are a capable platform for future telepresence research. Any discoveries will have immediate benefits for the vast majority of the population, who will be able to experience them in a convenient and inexpensive form factor using hardware they already own and prefer to use for such purposes. Though not yet computationally capable enough, current trends in mobile computing suggest that mobile phones will soon be more than capable of providing a truly immersive and natural experience whenever and wherever the user desires, eventually creating a new paradigm in the way we communicate in our daily lives.

[2]https://humansystems.arc.nasa.gov/groups/TLX/

References

Agarwala, A., Zheng, K. C., Pal, C., Agrawala, M., Cohen, M., Curless, B., Salesin, D., and Szeliski, R. (2005). Panoramic Video Textures. In *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05, pages 821–827, New York, NY, USA. ACM.

Al-Tairi, Z., Rahmat, R., Iqbal Saripan, M., and Sulaiman, P. S. (2014). Skin Segmentation Using YUV and RGB Color Spaces. *Journal of Information Processing Systems*, 10(2):283–299.

Andersen, D., Popescu, V., Lin, C., Cabrera, M. E., Shanghavi, A., and Wachs, J. (2016). A hand-held, self-contained simulated transparent display. In *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pages 96–101.

Anjos, R. K. d., Sousa, M., Mendes, D., Medeiros, D., Billinghurst, M., Anslow, C., and Jorge, J. (2019). Adventures in hologram space: Exploring the design space of eye-to-eye volumetric telepresence. In *25th ACM Symposium on Virtual Reality Software and Technology*, VRST '19, New York, NY, USA. Association for Computing Machinery.

Arminen, I. and Weilenmann, A. (2009). Mobile presence and intimacy—Reshaping social actions in mobile contextual configuration. *Journal of Pragmatics*, 41(10):1905–1923.

Bailenson, J. N., Swinth, K., Hoyt, C., Persky, S., Dimov, A., and Blascovich, J. (2005). The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. *Presence: Teleoperators and Virtual Environments*, 14(4):379–393.

Bauer, M., Kortuem, G., and Segall, Z. (1999). "Where Are You Pointing At?" A Study of Remote Collaboration in a Wearable Videoconference System. In *In: Proceedings of the 3rd International Symposium on Wearable Computers*, pages 151–158, San Francisco, California. IEEE.

Biber, P., Fleck, S., and Duckett, T. (2005). 3D Modeling of Indoor Environments for a Robotic Security Guard. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, 3:124–124.

Biocca, F., Harms, C., and Burgoon, J. K. (2003). Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria. *Presence: Teleoperators and Virtual Environments*, 12(5):456–480.

Campos-Castillo, C. and Hitlin, S. (2013). Copresence: Revisiting a building block for social interaction theories. *Sociological Theory*, 31(2):168–192.

Cao, Z., Hidalgo, G., Simon, T., Wei, S., and Sheikh, Y. (2018). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1812.08008.

Clark, H. and Brennan, E. (2008). Grounding in Communication. *En.Scientificcommons.Org*, pages 127–149.

Clark, H. H. and Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1):62–81.

Diverdi, S., Wither, J., and Hllerert, T. (2008). Envisor: Online Environment Map Construction for Mixed Reality. In *Virtual Reality Conference, 2008. VR '08. IEEE*, pages 19–26, Reno, NE, USA. IEEE.

Dourish, P. and Bellotti, V. (1992). Awareness and Coordination in Shared Workspaces. In *Proc. Intl. Conf. on Computer-Supported Cooperative Work*, pages 107–114.

Fanello, S., Rhemann, S. O.-e. C., Dou, M., Tankovich, V., Loop, C., and Chou, P. (2016). Holoportation: Virtual 3D Teleportation in Real-time. *Chi*, pages 741–754.

Flor, N. (1998). Side-by-side collaboration: a case study. *Internation Journal of Human-Computer Studies*, 49(3):201–222.

Fribourg, R., Argelaguet, F., Lécuyer, A., and Hoyet, L. (2020). Avatar and sense of embodiment: Studying the relative preference between appearance, control and point of view. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1.

Fussell, S. R., Kraut, R. E., and Siegel, J. (2000). Coordination of Communication: Effects of Shared Visual Context on Collaborative Work. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, pages 21–30.

Fussell, S. R., Setlock, L. D., Yang, J., Ou, J., Mauer, E., and Kramer, A. D. I. (2004). Gestures over Video Streams to Support Remote Collaboration on Physical Tasks. *Human-Computer Interaction*, 19(3):273–309.

Gao, L., Bai, H., Lee, G., and Billinghurst, M. (2016). An Oriented Point-Cloud View for MR Remote Collaboration. In *SIGGRAPH ASIA 2016 Mobile Graphics and Interactive Applications on - SA '16*, pages 1–4.

Gao, L., Bai, H., Lindeman, R., and Billinghurst, M. (2017). Static Local Environment Capturing and Sharing for MR Remote Collaboration. *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications on - SA '17*, pages 1–6.

Gauglitz, S., Lee, C., Turk, M., and Höllerer, T. (2012). Integrating the Physical Environment into Mobile Remote Collaboration. In *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services*, MobileHCI '12, pages 241–250, New York, NY, USA. ACM.

Gauglitz, S., Nuernberger, B., Turk, M., and Höllerer, T. (2014). World-stabilized Annotations and Virtual Scene Navigation for Remote Collaboration. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, pages 449–459, New York, NY, USA. ACM.

Hauber, J., Regenbrecht, H., Billinghurst, M., and Cockburn, A. (2006). Spatiality in videoconferencing: Trade-offs between efficiency and social presence. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, CSCW '06, pages 413–422, New York, NY, USA. ACM.

Hauber, J., Regenbrecht, H., Hills, A., Cockburn, A., and Billinghurst, M. (2005). Social Presence in Two-and Three-Dimensional Videoconferencing. In *Proceedings of ISPR Presence 2005*, pages 189–198.

History.com Editors (2009a). Alexander Graham Bell. `https://www.history.com/topics/inventions/alexander-graham-bell`. Accessed: 2020-02-07.

History.com Editors (2009b). Morse Code & the Telegraph. `https://www.history.com/topics/inventions/telegraph`. Accessed: 2020-02-07.

Huang, W. and Alem, L. (2013). HandsinAir: A Wearable System for Remote Collaboration on Physical Tasks. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion*, CSCW '13, pages 153–156, New York, NY, USA. ACM.

Jo, H. and Hwang, S. (2013). Chili: Viewpoint Control and On-Video Drawing for Mobile Video Calls. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13*, pages 1425–1430, New York, New York, USA. ACM Press.

Kasahara, S., Nagai, S., and Rekimoto, J. (2014). LiveSphere: Immersive Experience Sharing with 360 degrees Head-mounted Cameras. In *Proceedings of the adjunct publication of the 27th annual ACM symposium on User interface software and technology*, pages 61–62.

Kasahara, S. and Rekimoto, J. (2014). JackIn: Integrating First-Person View with Out-of-Body Vision Generation for Human-Human Augmentation. In *Proceedings of the 5th Augmented Human International Conference*, pages 46:1–46:8, Kobe, Japan. ACM.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., and Lilienthal, M. G. (1993). Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *The International Journal of Aviation Psychology*, 3(3):203–220.

Kennedy, R. S., Stanney, K. M., and Dunlap, W. P. (2000). Duration and Exposure to Virtual Environments: Sickness Curves During and Across Sessions. *Presence*, 9(5):463–472.

Kim, S., Lee, G., Sakata, N., and Billinghurst, M. (2014). Improving Co-Presence with Augmented Visual Communication Cues for Sharing Experience through Video Conference. In *ISMAR 2014 - IEEE International Symposium on Mixed and Augmented Reality - Science and Technology 2014, Proceedings*, pages 83–92.

Kirk, D., Crabtree, A., and Rodden, T. (2005). Ways of the Hands. In *Proceedings of the Ninth European Conference on Computer-Supported Cooperative Work*, pages 1–21.

Kirk, D. and Stanton Fraser, D. (2006). Comparing Remote Gesture Technologies for Supporting Collaborative Physical Tasks. In *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, pages 1191–1200.

Komiyama, R., Miyaki, T., and Rekimoto, J. (2017). Jackin space: Designing a seamless transition between first and third person view for effective telepresence collaborations. In *Proceedings of the 8th Augmented Human International Conference*, AH '17, pages 14:1–14:9, New York, NY, USA. ACM.

Kratz, S., Avrahami, D., Kimber, D., Vaughan, J., Proppe, P., and Severns, D. (2015). Polly Wanna Show You: Examining Viewpoint-Conveyance Techniques for a Shoulder-Worn Telepresence System. In *MobileHCI 2015 - Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, pages 567–575, Toronto, Canada.

Kratz, S. and Ferreira, F. (2016). Immersed Remotely: Evaluating the Use of Head Mounted Devices for Remote Collaboration in Robotic Telepresence. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 638–645. IEEE.

Kratz, S., Kimber, D., Su, W., Gordon, G., and Severns, D. (2014). Polly: "Being There" through the Parrot and a Guide. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '14*, pages 625–630, New York, New York, USA. ACM Press.

Kraut, R. E., Gergle, D., and Fussell, S. R. (2002). The Use of Visual Information in Shared Visual Spaces: Informing the Development of Virtual Co-Presence. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work - CSCW '02*, pages 31–40.

Kruijff, E., Swan, J., and Feiner, S. (2010). Perceptual issues in augmented reality revisited. pages 3 – 12.

Kuzuoka, H. (1992). Spatial Workspace Collaboration: A SharedView Video Support System for Remote Collaboration Capability. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, Monterey,:533–540.

Kuzuoka, H., Oyama, S., Yamazaki, K., Suzuki, K., and Mitsuishi, M. (2000). GestureMan. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work - CSCW '00*, pages 155–162, New York, New York, USA. ACM Press.

Leithinger, D., Follmer, S., Olwal, A., and Ishii, H. (2014). Physical telepresence: Shape capture and display for embodied, computer-mediated remote collaboration. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, page 461–470, New York, NY, USA. Association for Computing Machinery.

Lombard, M. and Ditton, T. (1997). At the Heart of It All: The Concept of Presence. *Journal of Computer-Mediated Communication*, 3(2). JCMC321.

Luff, P., Heath, C., Kuzuoka, H., Hindmarsh, J., and Oyama, S. (2003). Fractured Ecologies: Creating Environments for Collaboration. *Human-Computer Interaction*, 18(1):51–84.

Müller, J., Langlotz, T., and Regenbrecht, H. (2016). PanoVC: Pervasive Telepresence using Mobile Phones. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–10.

Newzoo (2020). Newzoo Analytics. `https://platform.newzoo.com/key-numbers`. Accessed: 2020-01-24.

Ofcom (2018). The Communications Market 2018: Interactive report - Ofcom. `https://www.ofcom.org.uk/research-and-data/multi-sector-research/cmr/cmr-2018/interactive`. Accessed: 2020-01-22.

Opensignal (2019). The State of Mobile Network Experience- Benchmarking 5G (PDF) Report Report | Opensignal. `https://www.opensignal.com/reports/2019/05/global-state-of-the-mobile-network`. Accessed: 2020-01-26.

Pacha, A. (2013). *Sensor Fusion for Robust Outdoor Augmented Reality Tracking on Mobile Devices*. Diploma thesis, University of Augsburg (Institut für Software & Systems Engineering).

Park, N., Mills, S., Whaanga, H., Mato, P., Lindeman, R. W., and Regenbrecht, H. (2019). Towards a Māori Telepresence System. In *2019 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6.

Pece, F., Steptoe, W., Wanner, F., Julier, S., Weyrich, T., Kautz, J., and Steed, A. (2013). Panoinserts: Mobile Spatial Teleconferencing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems SE - CHI '13*, pages 1319–1328.

Piumsomboon, T., Day, A., Ens, B., Lee, Y., Lee, G., and Billinghurst, M. (2017). Exploring enhancements for remote mixed reality collaboration. In *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications*, SA '17, pages 16:1–16:5, New York, NY, USA. ACM.

Poelman, R., Akman, O., Lukosch, S., and Jonker, P. (2012). As if Being There: Mediated Reality for Crime Scene Investigation. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, page 1267–1276, New York, NY, USA. Association for Computing Machinery.

Regenbrecht, H., McGregor, G., Ott, C., Hoermann, S., Schubert, T., Hale, L., Hoermann, J., Dixon, B., and Franz, E. (2011). Out of reach? — a novel ar interface approach for motor rehabilitation. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 219–228.

Regenbrecht, H. T., Schubert, T. W., and Friedmann, F. (1998). Measuring the Sense of Presence and its Relations to Fear of Heights in Virtual Environments. *International Journal of Human–Computer Interaction*, 10(3):233–249.

Ritchie, H. and Roser, M. (2019). Technology adoption. *Our World in Data*.

Rother, C., Kolmogorov, V., and Blake, A. (2004). "GrabCut": Interactive Foreground Extraction Using Iterated Graph Cuts. *ACM Transactions on Graphics*, 23(3):309.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2564–2571.

Sakong, K. and Nam, T.-j. (2006). Supporting Telepresence by Visual and Physical Cues in Distributed 3D Collaborative Design Environments. In *CHI '06 extended abstracts on Human factors in computing systems - CHI EA '06*, page 1283, New York, New York, USA. ACM Press.

Schubert, T., Friedmann, F., and Regenbrecht, H. (2001). The Experience of Presence: Factor Analytic Insights. *Presence: Teleoperators and Virtual Environments*, 10(3):266–281.

Sodhi, R. S., Jones, B. R., Forsyth, D., Bailey, B. P., and Maciocci, G. (2013). BeThere: 3D Mobile Collaboration with Spatial Input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 179–188, New York, NY, USA. ACM.

Stotko, P., Krumpen, S., Hullin, M. B., Weinmann, M., and Klein, R. (2019). Slamcast: Large-scale, real-time 3d reconstruction and streaming for immersive multi-client live telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):2102–2112.

Tait, M. and Billinghurst, M. (2015). The Effect of View Independence in a Collaborative AR System. *Computer Supported Cooperative Work (CSCW)*, 24(6):563–589.

Tang, A., Fakourfar, O., Neustaedter, C., and Bateman, S. (2017). Collaboration in 360° Videochat: Challenges and Opportunities. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, pages 1327–1339, Edinburgh, United Kingdom. ACM.

Taylor, P., Kraut, R. E., Fussell, S. R., and Siegel, J. (2009). Visual Information as a Conversational Resource in Collaborative Physical Tasks. *Human-Computer Interaction*, 0024(917808986):13–49.

Teo, T., Lawrence, L., Lee, G. A., Billinghurst, M., and Adcock, M. (2019). Mixed reality remote collaboration combining 360 video and 3d reconstruction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 201:1–201:14, New York, NY, USA. ACM.

Viola, P. and Jones, M. J. (2004). Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137.

VSee (2011). A Missing Link in the History of the Videophone - VSee. `https://vsee.com/blog/a-missing-link-in-the-history-of-the-videophone/`. Accessed: 2020-02-07.

Wang, R. and Quek, F. (2010). Touch & talk: Contextualizing remote touch for affective interaction. In *Proceedings of the Fourth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI '10, page 13–20, New York, NY, USA. Association for Computing Machinery.

Worldometer (2020). World Population Clock: 7.8 Billion People (2020) - Worldometer. `https://www.worldometers.info/world-population/`. Accessed: 2020-01-24.

WorldTimeZone (2019). 4G map LTE World Coverage Map - LTE WiMAX HSPA 3G GSM Country List. `https://www.worldtimezone.com/4g.html`. Accessed: 2020-01-26.

Xenophon (2005). *The Expedition of Cyrus*. Oxford.

Xu, W., Chatterjee, A., Zollhöfer, M., Rhodin, H., Fua, P., Seidel, H.-P., and Theobalt, C. (2019). Mo2cap2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):2093–2101.

Yang, J. and Waibel, A. (1996). A Real-Time Face Tracker. In *Proceedings Third IEEE Workshop on Applications of Computer Vision. WACV'96*, pages 142–147.

Young, J., Langlotz, T., Cook, M., Mills, S., and Regenbrecht, H. (2019). Immersive telepresence and remote collaboration using mobile and wearable devices. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):1908–1918.

Young, J., Langlotz, T., Mills, S., and Regenbrecht, H. (2020). Mobileportation: Nomadic telepresence for mobile devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(2).

Acronyms

DoF Degrees of Freedom.

FoV Field of View.

fps frames per second.

GPU Graphics Processing Unit.

HMD Head-Mounted Display.

JNI Java-Native Interface.

SLAM Simultaneous Localisation and Mapping.

ToF Time of Flight.

Glossary

co-presence The degree to which two people feel mutual entrainment with each other, where entrainment is a synchronisation of mutual attention, emotion, and behaviour..

conversational grounding A process by which interlocutors establish common knowledge about the topic of conversation and its relevant context. Grounding takes place through a series of turns, each of which serves to right incorrect assumptions an slowly converge on some shared understanding between peers..

deictic reference A verbal utterance such as "this one" that refers to a specific person or object.

egocentric Within a central position. In the context of collaborative systems, a remote user takes an egocentric viewing position when they and the local user occupy the exact same position in space..

exocentric External to a central position. In the context of collaborative systems, a remote user takes an exocentric viewing position when they and the remote user occupy separate spaces within the shared environment..

local user A user of a collaborative system who is physically located within the environment to be shared..

pointing gesture A hand or body gesture with specific directionality that attempts to highlight a person or object of interest..

pulling Pulling information is a process of retrieving information from a conversational partner's environment or mental model without intervention on their part, such as by independently looking at or moving through their physical space..

pushing Pushing information is a process of delivering it to a conversational partner without intervention or effort on their part, such as by speaking to them or touching them..

remote user A user of a collaborative system who is physically situated in some remote location and wishes to be virtually transported to the shared one..

representational gesture A hand or body gesture meant to represent an object or action, such as waving a hand over the object to manipulate or mimicking the action to be performed by a remote collaborator..

social presence The degree to which conversation feels unmediated; the sense of "being there with another.".

spatial presence The degree to which a user feels physically located within a real or virtual environment; the sense of "being there" rather than viewing the space through some surrogate..

telepresence A field that attempts to make two remote speakers feel as if they are really together than communicating through some intermediate medium..

Appendix A

Questionnaire for the 2D Telepresence Study

The following are the information sheet and questionnaire presented to participants after each condition of the study used to evaluate the system presented in chapter 3 as well as the relevant ethics application. Questions evaluating spatial presence within the environment were adapted from the questionnaire designed by Schubert et al. (2001), those evaluating co-presence with the communication partner were adapted from a questionnaire designed by (Biocca et al., 2003), and those evaluating simulator sickness are from Kennedy et al. (1993).

Form Updated: December 2017

UNIVERSITY OF OTAGO HUMAN ETHICS COMMITTEE APPLICATION FORM: CATEGORY B

(Departmental Approval)

Please ensure you are using the latest application form available from:
http://www.otago.ac.nz/council/committees/committees/HumanEthicsCommittees.html

1. University of Otago staff member responsible for project:
Langlotz Tobias (Dr)

2. Department/School:
Department of Information Science/School of Business

3. Contact details of staff member responsible *(always include your email address)***:**
Telephone number: 479 8096
Email address: tobias.langlotz@otago.ac.nz

4. Title of project:
Towards Immersive Telepresence and Remote Collaboration using Mobile and Wearable Devices

5. Indicate type of project and names of other investigators and students:

Staff Research	X	**Names**	Dr Tobias Langlotz Prof Holger Regenbrecht Dr Steven Mills
Student Research	X	**Names**	Jacob Young Oliver Reid
Level of Study (e.g. PhD, Masters, Hons)			PhD BSc
External Research/		**Names**	
Collaboration			
Institute/Company			

6. When will recruitment and data collection commence?

11th April 2018

When will data collection be completed?

30th November 2018

7. Brief description in lay terms of the aim of the project, and outline of the research questions that will be answered (approx. 200 words)**:**

The purpose of this study is to evaluate the usefulness of a proposed system for use as an immersive telepresence application. The system connects two users, who are potentially remotely located in regards to each other, who can then communicate with each other within some shared real-world environment created through use of either a mobile phone or an external 360° camera. The system will be evaluated on its ability to evoke a sense of spatial presence (the sense of "being there") within the remote environment and co-presence (mutual entrainment) with the communication partner. The sense of presence induced through either condition will be evaluated via several industry-standard questionnaires completed by each participant retrospectively.

8. Brief description of the method. Include a description of who the participants are, how the participants will be recruited, and what they will be asked to do and how the data will be used and stored *(Note: if this research involves **patient data or health information** obtained from the Ministry of Health, DHBs etc please refer to the UOHEC(H) Minimal Risk Health Research - Audit and Audit related studies):-*

Participants will be split into two groups: a *live group,* in which participants will be communicating live with a study mediator, and a *recorded group,* in which they will be viewing a previously recorded communication session. The experiment is a within-subjects design where the independent variable is the system participants use, ie. the application users use to communicate with the study mediator. Both groups will participate in the same three conditions, which are:

1. Spatial Videoconferencing: The participant is shown video recorded using a mobile phone. This is either streamed live from a study mediator (in the case of the live group) or pre-recorded (in the case of the recorded group). The position of the phone is also recorded, which is used to project each video frame into a virtual spherical environment based on the rotation of the device when that frame was recorded. This allows the participant to infer spatial context from the positioning of each frame, as positions of objects in the video stream will match their relative locations in the real remote environment. Participants view this environment through a *head-mounted display* (HMD) to further immerse the participant within it, and can track this moving video stream or look elsewhere by rotating their head. The participant's own camera stream is similarly captured and projected, however theirs is processed so that only their hands are visible, providing embodiment in the environment and allowing them to point to objects of interest. This is shown to the remote study mediator (or local one for the recorded group) so that they can react to the participant's behaviour, and the two can also communicate through voice.
2. Incremental panoramic videoconferencing: Similar to the first condition, however as the local user (or pre-recorded video) moves around the environment it is recorded there, creating a 360° panorama over time. This allows the remote user to independently view previously visited areas.

3. Live panoramic videoconferencing: Also similar to condition 1, but an external 360° camera is used in place of the mobile phone's inbuilt camera. This allows for the full environment to be updated in real-time.

In all three conditions, the local user will be the same mediator whose performance will not be evaluated. Participants in the recorded group will be shown the same pre-recorded scenario for each condition.

Participants must be between the age of 18 and 65, have normal or corrected to normal vision, and must speak English in order to answer questionnaires. The study only requires one session from each participant and should take no longer than one hour. Participants will be recruited via the attached advertisement.

Procedure

Should participants agree to take part in this project, the will be asked to participate in a study where they will interact with either a remote mediator or a pre-recorded communication session using the described system.

Participants will first be presented with a demographics questionnaire regarding their age, gender, ethnicity, vision impairments, prior experience with *Virtual Reality* (VR) applications, and susceptibility to simulator sickness. Any participant that indicates they are susceptible to simulator sickness will be immediately excluded with no disadvantage to them. Any participant who begins to feel the effects of simulator sickness during the course of the study will similarly be immediately excluded. Participants will be encouraged to notify the study mediator if they begin to feel any symptoms of simulator sickness.

Once the participant has completed the demographics questionnaire they will be shown how to operate the system used in the first condition. To mitigate order bias, the order each condition is tested in will be randomised for each participant. Once the participant has completed the demographics questionnaire the first condition will be explained to them. They will then have two minutes in which to familiarise themselves with the system's operation; this will be done with a pre-recorded scenario rather than a live connection with the study mediator to allow them time to familiarise themselves with the system without external pressure and to avoid prior familiarity with the evaluated environment. For participants in the recorded group, the location they are shown will differ from the one used in the actual study condition.

Once the two minutes have expired, participants in the live group will be connected to the study mediator, or the pre-recorded session will begin for participants in the recorded group. They will then be given two and a half minutes to communicate within the shared environment. Participants will be told to actively participate in the conversation, for example to point to and ask about landmarks which they wish to know more about.

Participants will be notified once the time for the current condition has expired. They will then be given the presence questionnaires and asked to complete them at their leisure. The procedure will then be repeated for the remaining two conditions. Upon completion of all three, participants will be asked to complete another questionnaire about any simulator sickness experienced during the experiment, and upon its completion will be reimbursed for their time with a $20 New World supermarket voucher.

9. Disclose and discuss any potential problems and how they will be managed: (For example: medical/legal problems, issues with disclosure, conflict of interest, safety of the researcher, etc)

Simulator Sickness

The study comprises of standard off-the-shelf mobile phones similar to that which each participant will likely own. The use of an immersive VR HMD may induce *Simulator Sickness* (SS), particularly in those who have had little exposure to such devices in the past. The demographics questionnaire completed by each participant will ask their familiarity with VR systems and their susceptibility to SS; those who self-identify as being prone to severe SS will be excluded from the study with no further disadvantage to themselves. Additionally, all participants will be told they are able to withdraw from the study at any time if they feel they are being affected by SS. The lack of movement through the virtual environment and short exposure time to the virtual reality hardware minimise the risk of simulator sickness, even in inexperienced users.

Data Protection

All data to be collected will be anonymized. To aid in this, consent forms and completed questionnaires will be stored in different locations in a randomised order. Any digitized information will only be stored locally on one device, and will be password protected to ensure data remains safe in the case of the device being lost or stolen. This also makes the information inaccessible to anyone but the research team. No material that allows for identification of the participant will be recorded, and all material transmitted between the two mobile devices is strongly encrypted on a private password-protected network to prevent interception by a third party.

Data Analysis

It is possible that the scope of the data to be collected will exceed what is currently outlined in this application. This is unlikely, but if this is the case we will apply to the ethics committee for a revision of our ethics approval to allow for this potentially undetermined data to be analysed.

Reporting Sheet for use ONLY for proposals considered at departmental level

---

***Applicant's Signature:** .....

Name (please print): ……..

Date: 10/04/2018

**The signatory should be the staff member detailed at Question 1.*

ACTION TAKEN

☐ Approved by HOD ☐ Approved by Departmental Ethics Committee

☐ Referred to UO Human Ethics Committee

Signature of **Head of Department*:* .....

Name of HOD (please print): ……..

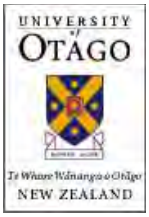
Date: .....

**Where the Head of Department is also the Applicant, then an appropriate senior staff member must sign on behalf of the Department or School.

Departmental approval: *I have read this application and believe it to be valid research and ethically sound. I approve the research design. The research proposed in this application is compatible with the University of Otago policies and I give my approval and consent for the application to be forwarded to the University of Otago Human Ethics Committee (to be reported to the next meeting).*

IMPORTANT NOTE: As soon as this proposal has been considered and approved at departmental level, the completed form, together with copies of any Information Sheet, Consent Form, recruitment advertisement for participants, and survey or questionnaire should be forwarded to the Manager, Academic Committees or the Academic Committees Administrator, Academic Committees, Rooms G22, or G26, Ground Floor, Clocktower Building, or scanned and emailed to either gary.witte@otago.ac.nz. or jane.hinkley@otago.ac.nz

[Reference Number: *as allocated upon approval by the Human Ethics Committee*]
10/04/18



TOWARDS IMMERSIVE TELEPRESENCE AND REMOTE COLLABORATION USING MOBILE AND WEARABLE DEVICES
INFORMATION SHEET FOR PARTICIPANTS

Thank you for showing an interest in this project. Please read this information sheet carefully before deciding whether or not to participate. If you decide to participate we thank you. If you decide not to take part there will be no disadvantage to you and we thank you for considering our request.

What is the Aim of the Project?

We aim to evaluate a proposed prototype for remote communication between two people. The system will be evaluated on its ability to evoke a sense of spatial presence (the sense of "being there") within the communicating partner's environment, as well as its ability to provide co-presence with that communicating partner (the sense of being present in that environment and having someone else there with you). The amount of presence evoked will be measured using retrospective questionnaires after experiencing this proposed system for yourself.

What Types of Participants are being sought?

Participants are sought mainly from staff and students at the University of Otago, from various disciplines, and with an age of between 18 and 65. We aim to recruit 25 participants.

What will Participants be asked to do?

Should you agree to take part in this project, you will be asked to participate in a study where you will interact with another person within a shared virtual environment. This environment will be constructed from that other user's real surroundings, effectively transporting you to that remote location. This study comprises one session only and should take no longer than one hour.

Participants will be divided into two groups: a *live group,* in which the video seen is recorded live, and a *recorded group*, in which all video seen is pre-recorded.

You will first be presented with a demographics questionnaire for collecting information on age, gender, ethnicity, potential vision impairments, prior *Virtual Reality* (VR) experience and susceptibility to simulator sickness.

The system to be used for this study comprises of two mobile phones, one of which will be used by yourself and the other by the remote party you are to communicate with. The mobile phone will be viewed through a mobile *Head-Mounted Display* similar to that available from most electronics stores. The study will comprise of three parts, the order of which will be chosen randomly.

For each of the three parts you will be introduced to the system you will be using, then given two minutes to familiarise yourself with its use. This will be done with a pre-recorded scenario rather than a live communication partner so that you can experiment in a pressure-free environment.

When comfortable in operating the system, you will then be connected to the remote communication partner; you will be informed before this happens. You will then have five minutes to interact with this partner and immerse yourself within their environment; we encourage you to take active part in this conversation, asking about and indicating toward environmental features. Once two and a half minutes have passed we will ask you to complete a questionnaire evaluating your experience with and sense of presence within the system; this process will then be repeated for the remaining two phases.

After all three phases have been completed you will be presented with a final questionnaire evaluating the extent to which you experienced simulator sickness during the experiment. Once completed, you will be awarded a $20 New World supermarket voucher to reimburse you for your time.

This study carries a risk of experiencing simulator sickness. This may induce nausea, headaches, dizziness or other symptoms. We do not expect this to be a problem for this particular study, but if you experience these or any other symptoms please inform the study mediator and you will be withdrawn without any disadvantage to yourself.

Please be aware that you may decide not to take part in the project without any disadvantage to yourself.

What Data or Information will be collected and what use will be made of it?

The demographic data collected includes age, gender, ethnicity, and potential vision impairments, as well as familiarity with similar systems and technologies. A paper-based simulator sickness questionnaire is administered to attain data relevant to the user's experiences in VR. A Presence questionnaire is used to measure user's sense of Presence. The researcher may take notes during the experiment, and participants will be encouraged to vocalise their thoughts regarding their experience. No personally identifiable data will be collected beyond that included in the demographic questionnaire, and every effort will be made to ensure that no data can be linked to any individual participant. Only the study coordinator will be able to see the actions you perform within the shared communication environment; the required camera images will not be stored or transmitted anywhere other than where required for the purpose of the study. The transmitted video is securely encoded such that it cannot be viewed or intercepted by a third party.

The data collected will be securely stored in such a way that only those mentioned below will be able to gain access to it. Data obtained as a result of the research will be retained for **at least 5 years** in secure storage. Any personal information held on the participants may be destroyed at the completion of the research even though the data derived from the research will, in most cases, be kept for much longer or possibly indefinitely.

The results of the project may be published and will be available in the University of Otago Library (Dunedin, New Zealand) but every attempt will be made to preserve your anonymity.

Can Participants change their mind and withdraw from the project?

You may withdraw from participation in the project at any time and without any disadvantage to yourself.

What if Participants have any Questions?

If you have any questions about our project, either now or in the future, please feel free to contact either:-

Reporting Sheet for use ONLY for proposals considered at departmental level

---

Jacob Young, Department of Information Science, University of Otago. Telephone Number: 479 5420,
Email Address: youja802@student.otago.ac.nz.

Oliver Reid, Department of Information Science, University of Otago. Telephone Number: 479 5420,
Email Address: reiol787@student.otago.ac.nz

Holger Regenbrecht, Department of Information Science, University of Otago. Telephone Number: 479 8322,
Email Address: holger.regenbrecht@otago.ac.nz.

Tobias Langlotz, Department of Information Science, University of Otago. Telephone Number: 479 8096,
Email Address: tobias.langlotz@otago.ac.nz

This study has been approved by the Department stated above. However, if you have any concerns about the ethical conduct of the research you may contact the University of Otago Human Ethics Committee through the Human Ethics Committee Administrator (ph +643 479 8256 or email gary.witte@otago.ac.nz). Any issues you raise will be treated in confidence and investigated and you will be informed of the outcome.

Study Questionnaire

In the following document are 18 Questions to answer in response to your experience with the application.

The questions follow the format of a statement, followed by a 7-Point Likert scale, similar to the one below. You will be asked to show how much you agree with each statement by circling the appropriate response on the scale below, with a score of ‘1’ indicating that you strongly disagree with the statement, and a score of ‘7’ indicating that you strongly agree.

An example of the scale is below:

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Disagree

Please ensure you completely understand this now, as once you begin answering the questions you cannot ask for any more help.

Please answer each question to the best of your ability, and to your own interpretation.

1) I had the feeling that I was in the middle of the action, rather than merely observing

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Disagree

2) I felt like I was part of the presented environment

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Disagree

3) I felt like I was actually there in the presented environment

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Disagree

4) I felt like the objects in the presented environment surrounded me

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Disagree

5) It was as though my true location had shifted into the presented environment

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Disagree

6) It seemed as though my self was present in the presented environment

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Disagree

7) I felt as though I was physically present in the presented environment

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Disagree

8) It seemed as though I actually took part in the actions of the presented environment

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Disagree

9) I had control of my perspective of the remote environment

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Disagree

10) I had a sense of being with the other person

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Disagree

11) I felt like there was someone else with me

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Disagree

12) I felt like the other user was aware of my presence

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Disagree

13) I was comfortable participating

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Disagree

14) Any additional comments?

No:_____ Date:_____

Simulator Sickness Questionnaire

Kennedy, Lane, Berbaum, & Lilienthal (1993)[1]

Instructions: Circle how much each symptom below is affecting you right now.

1. General discomfort	None	Slight	Moderate	Severe
2. Fatigue	None	Slight	Moderate	Severe
3. Headache	None	Slight	Moderate	Severe
4. Eye strain	None	Slight	Moderate	Severe
5. Difficulty focusing	None	Slight	Moderate	Severe
6. Increased salivation	None	Slight	Moderate	Severe
7. Sweating	None	Slight	Moderate	Severe
8. Nausea	None	Slight	Moderate	Severe
9. Difficulty concentrating	None	Slight	Moderate	Severe
10. Fullness of head*	None	Slight	Moderate	Severe
11. Blurred vision	None	Slight	Moderate	Severe
12. Dizziness (eyes open)	None	Slight	Moderate	Severe
13. Dizziness (eyes shut)	None	Slight	Moderate	Severe
14. Vertigo†	None	Slight	Moderate	Severe
15. Stomach awareness‡	None	Slight	Moderate	Severe
16. Burping	None	Slight	Moderate	Severe

* Fullness of the head refers to a sensation of pressure in the head without any pain, like that experienced when upside down.
† Vertigo is a loss of orientation with respect to verticality, like the sensation felt at great heights.
‡ Stomach awareness is usually used to indicate a feeling of discomfort which is just short of nausea.

[1] Original version: Kennedy, R.S., Lane, N.E., Berbaum, K.S., & Lilienthal, M.G. (1993). Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness. International Journal of Aviation Psychology, 3(3), 203-220.

Appendix B

Questionnaires for the 3D Telepresence Study

The following are the questionnaires and information sheet presented to participants to evaluate the system detailed in chapter 5 as well as the relevant ethics application. The first was completed after each condition to determine the amount of spatial, social and co-presence experienced by participants and was derived from questionnaires by Schubert et al. (2001), Biocca et al. (2003), Hauber et al. (2006), and Bailenson et al. (2005). The second was given to participants after both conditions had been completed and was used to determine if they had a preference for either experienced application.

Form Updated: November 2018

UNIVERSITY OF OTAGO HUMAN ETHICS COMMITTEE APPLICATION FORM: CATEGORY B

(Departmental Approval)

Please ensure you are using the latest application form available from: http://www.otago.ac.nz/council/committees/committees/HumanEthicsCommittees.html

1. University of Otago staff member responsible for project:
Langlotz, Tobias (Assoc. Prof.)

2. Department/School:
Department of Information Science/School of Business

3. Contact details of staff member responsible *(always include your email address)***:**
Telephone number: 479 8096
Email address: tobias.langlotz@otago.ac.nz

4. Title of project:
Pano3D: 6DoF Telepresence on Mobile Devices

5. Indicate type of project and names of other investigators and students:

Staff Research	X	**Names**	Assoc. Prof. Tobias Langlotz Prof. Holger Regenbrecht Dr Steven Mills
Student Research	X	**Names**	Jacob Young Noel Park
Level of Study (e.g. PhD, Masters, Hons)			PhD PhD
External Research/		**Names**	
Collaboration			
Institute/Company			

6. When will recruitment and data collection commence?

Monday 5th August 2019

When will data collection be completed?

Wednesday 5th August 2020

7. Brief description in lay terms of the aim of the project, and outline of the research questions that will be answered (approx. 200 words)**:**

The purpose of this study is to evaluate a novel mobile telepresence system which connects two potentially remote users within some shared real-world environment through use of mobile hardware. The system comprises of two mobile phones and two 360° cameras (one for each user) which are used to capture a three-dimensional representation of one user's surroundings and share it with their remote communication partner. The application will be evaluated on its ability to increase participants' spatial awareness within this shared environment with little cognitive load required as well as provide a heightened sense of physical presence within it (spatial presence) and mutual emotional connection with their communication partner (co-presence). Spatial awareness will be measured using a novel subjective measure, while cognitive load, spatial presence and co-presence will be measured using several industry-standard questionnaires that will be completed retroactively.

8. Brief description of the method. Include a description of who the participants are, how the participants will be recruited, and what they will be asked to do and how the data will be used and stored *(Note: if this research involves **patient data or health information** obtained from the Ministry of Health, DHBs etc please refer to the UOHEC(H) Minimal Risk Health Research - Audit and Audit related studies):-*

Participants will be asked to go on a "guided tour" through an indoor environment using two mobile teleconferencing systems. Both comprise of a 360° camera attached to the rear side of a mobile phone using a purpose-built mount. The experiment will make use of a within-subjects design where the independent variable is which of these two systems is in use during each condition. The two systems being evaluated are described as follows:

1. The first system is a conventional videoconferencing application. The 360° video captured by the attached camera is streamed from a study mediator to the participant's device in order for that mediator's environment to be shared. The participant may then obtain novel viewpoints within this video by rotating their own device in the desired direction, however no means of performing translational movement is provided. Audio communication capabilities are also included so that the mediator and participant may speak to one another. Such systems are well understood and so this is included as a point of comparison.
2. The second system is a novel implementation and so will be unfamiliar to all participants. The study mediator's mobile phone contains a depth sensing device that is used to create a three-dimensional reconstruction of their surroundings. This data is streamed live to the participant's device, and the participant can then freely move around within this virtual environment by walking about their own space; a novel first-person view of the reconstruction is shown to the participant based on their current location within it. The location of both parties is shown to the other via a virtual avatar, which will also show live video of their face as captured by the attached camera; the participant's face will always be visible to the mediator due to the 360° nature of the camera and integrated face tracking. If either user moves within a small distance of the other, the application will transition to a view similar to the first system where the video from the attached camera will be visible as well as any three-dimensional data that is within a

short distance from the participant. Audio capabilities will again be enabled to allow the participant and mediator to talk to one another.

Participants must be between the ages of 18 and 65, have normal or corrected to normal vision, and speak English in order to answer questionnaires. Only one session is required of each participant which will take no longer than one hour. Participants will be recruited via the attached advertisement.

Procedure
Should participants agree to take part in this experiment, they will be tasked with performing an informal guided tour through two real-world environments.

Participants will first be presented with a demographics questionnaire regarding their age, gender, ethnicity and vision impairments. They will then be instructed on how to operate the system used for the first condition, which will be randomly selected from the previously described applications in order to reduce potential learning effects.

Once participants feel comfortable with operating the system, they will be connected to the remote study mediator who will give them a guided tour through an indoor environment. There will be no set task for participants to perform during this tour other than to familiarise themselves with the space. Upon completion of the tour, participants will be asked to sketch the layout of the shown environment to the best of their ability; this sketch will later be examined to determine its similarity to the real environment. They will then be asked to complete several industry-standard questionnaires to determine the cognitive load experienced while producing this sketch as well as their perceived spatial presence within the shared environment and co-presence with the study mediator. This process will then be completed for the remaining condition.

Both conditions will take less than one hour to complete. Upon completion of both, participants will receive a $20 New World supermarket voucher.

9. Disclose and discuss any potential problems and how they will be managed: (For example: medical/legal problems, issues with disclosure, conflict of interest, safety of the researcher, safeguards to participant anonymity if open access to data is proposed etc)

Simulator Sickness
Similar systems often utilise a head-mounted virtual reality display in order to increase the participant's sense of immersion within the virtual environment. Such displays introduce the risk of participants experiencing simulator sickness, particularly in those unfamiliar with these devices. Our system chooses to instead use the mobile phone's integrated display and so no such risk is present.

Data Protection
All data to be collected will be anonymized. To aid in this, consent forms and completed questionnaires will be stored in different locations in a randomised order. Any digitized information will only be stored locally on one device, and will be password protected to ensure data remains safe in the case of the device being lost or stolen. This also makes the information inaccessible to anyone but the research team. No material that allows for identification of the participant will be recorded, and all material transmitted between the two mobile devices is strongly encrypted on a private password-protected network to prevent interception by a third party.

Data Analysis
It is possible that the scope of the data to be collected will exceed what is currently outlined in this application. This is unlikely, but if this is the case, we will apply to the ethics committee for a revision of our ethics approval to allow for this potentially undetermined data to be analysed.

Reporting Sheet for use ONLY for proposals considered at departmental level

---

***Applicant's Signature:** .....

Name (please print): Tobias Langlotz

Date: 05.08.2019

**The signatory should be the staff member detailed at Question 1.*

ACTION TAKEN

☐ Approved by HOD ☐ Approved by Departmental Ethics Committee

☐ Referred to UO Human Ethics Committee

Signature of **Head of Department: .....

Name of HOD (please print): ……..

Date: .....

**Where the Head of Department is also the Applicant, then an appropriate senior staff member must sign on behalf of the Department or School.

Departmental approval: *I have read this application and believe it to be valid research and ethically sound. I approve the research design. The research proposed in this application is compatible with the University of Otago policies and I give my approval and consent for the application to be forwarded to the University of Otago Human Ethics Committee (to be reported to the next meeting).*

IMPORTANT NOTE: As soon as this proposal has been considered and approved at departmental level, the completed form, together with copies of any Information Sheet, Consent Form, recruitment advertisement for participants, and survey or questionnaire should be forwarded to the Manager, Academic Committees and Services, **(1st Floor, Scott/Shand House, 90 St David's Street (Rooms 1.05 and 1.08)**.gary.witte@otago.ac.nz, or Senior Administrators Jo Farron de Diaz, jo.farrondediaz@otago.ac.nz or Ruth Sharpe ruth.sharpe@otago.ac.nz .

[Reference Number: *as allocated upon approval by the Human Ethics Committee*]
[*Date*]

Pano3D: 6DoF Telepresence on Mobile Devices
INFORMATION SHEET FOR PARTICIPANTS

Thank you for showing an interest in this project. Please read this information sheet carefully before deciding whether or not to participate. If you decide to participate we thank you. If you decide not to take part there will be no disadvantage to you and we thank you for considering our request.

What is the Aim of the Project?

We aim to evaluate a proposed prototype for remote communication between two people. The system will be evaluated on its ability to provide a sense of spatial awareness and presence (the sensation of "being there") within a shared remote environment as well as social presence and co-presence (mutual emotional connection) with a remote communication partner as well as the cognitive load required on your part to complete the prescribed tasks. These factors will be evaluated after experiencing the system using retrospective questionnaires.

What Types of Participants are being sought?

Participants are sought mainly from staff and students of the University of Otago from various disciplines, though others unaffiliated with the university will not be excluded from participating. Participants should be aged between 18 and 65 and speak English well enough to complete several questionnaires.

What will Participants be asked to do?

Should you agree to take part in this project, you will be asked to interact with a remote communication partner within a shared virtual environment. This will be constructed from the real-world location of the remote party, effectively transporting you to that remote location. This study consists of one session only and should take no longer than one hour.

The system used in this experiment consists of two mobile phones, one of which will be used by yourself and the other by your remote communication partner. Each phone is held within a purpose-built mount that also houses a standard 360° camera. Two separate applications will be used during the experiment in random order; both will be explained to you when relevant.

You will first be presented with a demographics questionnaire that collects information on age, gender, ethnicity and potential vision impairments. You will then be instructed on how to use the application used during the first condition and then connected with a remote

communication partner who will take you on a guided tour through a remote indoor environment. We encourage you to actively communicate with your partner during this time. Upon completion of the tour you will then be asked to provide a rough sketch of the area you were guided through, then complete several questionnaires to evaluate your experience while using the application and the cognitive load required on your part to complete the aforementioned sketch. This process will then be repeated for the remaining application, after which you will be given a $20 New World gift voucher.

Please be aware that you may decide not to take part in the project without any disadvantage to yourself.

What Data or Information will be collected and what use will be made of it?

The demographic data collected includes age, gender, ethnicity, and potential vision impairments. A presence and cognitive load questionnaire will be used to attain data relevant to your experience within the virtual environment. Your sketch of each virtual environment visited will be kept and used to determine the degree of spatial awareness afforded by each application.

The researcher may take notes during the experiment, and participants will be encouraged to vocalise their thoughts regarding their experience. No personally identifiable data will be collected beyond that included in the demographics questionnaire, and every effort will be made to ensure that data cannot be linked to any one participant. Only the study coordinators will be able to see the actions you perform within the shared virtual environment; any video captured by the system will not be recorded and is not transmitted anywhere other than is required for the purposes of the experiment. Any video transmitted is securely encoded such that it cannot be viewed or intercepted by a third party.

The data collected will be securely stored in such a way that only those mentioned below will be able to gain access to it. Data obtained as a result of the research will be retained for **at least 5 years** in secure storage. Any personal information held on the participants may be destroyed at the completion of the research even though the data derived from the research will, in most cases, be kept for much longer or possibly indefinitely.

No material that could personally identify you will be used in any reports on this study. Results of this research may be published and will be available in the University of Otago Library (Dunedin, New Zealand) but every attempt will be made to preserve your anonymity.

Can Participants change their mind and withdraw from the project?

You may withdraw from participation in the project at any time and without any disadvantage to yourself.

What if Participants have any Questions?

If you have any questions about our project, either now or in the future, please feel free to contact either:-

Jacob Young, Department of Information Science, University of Otago. Telephone Number: 479 5420, Email Address: youja802@student.otago.ac.nz

Noel Park, Department of Information Science, University of Otago. Telephone Number: 479 5420, Email Address: parju458@student.otago.ac.nz

Holger Regenbrecht, Department of Information Science, University of Otago. Telephone Number: 479 8322, Email Address: holger.regenbrecht@otago.ac.nz.

Tobias Langlotz, Department of Information Science, University of Otago. Telephone Number: 479 8096, Email Address: tobias.langlotz@otago.ac.nz

This study has been approved by the Department stated above. However, if you have any concerns about the ethical conduct of the research you may contact the University of Otago Human Ethics Committee through the Human Ethics Committee Administrator (ph +643 479 8256 or email gary.witte@otago.ac.nz). Any issues you raise will be treated in confidence and investigated and you will be informed of the outcome.

Date: Participant #: Condition:

Presence Questionnaire

In the following document are 26 questions to answer in response to your experience with the application.

The questions follow the format of a statement followed by a 7-Point Likert scale, similar to the one below. You will be asked to show how much you agree with each statement by circling the appropriate response on the scale below.

1	2	3	4	5	6	7
Fully Disagree			Neither Agree nor Disagree			Fully Agree

Please ensure you completely understand this now, as once you begin answering the questions you cannot ask for any more help.

Please answer each question to the best of your ability and to your own interpretation.

1) In the computer generated world I had a sense of “being there.”

1	2	3	4	5	6	7
Not At All						Very Much

2) Somehow I felt that the virtual world surrounded me.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

3) I felt like I was just perceiving pictures.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

4) I did not feel present in the virtual space.

1	2	3	4	5	6	7
Did Not Feel Present						Felt Present

5) I had a sense of acting in the virtual space, rather than operating something from outside.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

6) I felt present in the virtual space.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

7) How aware were you of the real world surrounding you while navigating in the virtual world? (i.e. sounds, room temperature, other people, etc.)?

1	2	3	4	5	6	7
Extremely Aware						Not Aware At All

8) I was not aware of my real environment.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

9) I still paid attention to the real environment.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

10) I was completely captivated by the virtual world.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

11) How real did the virtual world seem to you?

1	2	3	4	5	6	7
Completely Real						Not Real At All

12) How much did your experience in the virtual environment seem consistent with your real world experience?

1	2	3	4	5	6	7
Not Consistent						Very Consistent

13) How real did the virtual world seem to you?

1	2	3	4	5	6	7
About as Real as an Imagined World						Indistinguishable from the Real World

14) The virtual world seemed more realistic than the real world.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

BAIL 15) Even when the other person was present, I often felt alone in the virtual space.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

16) I felt like there was someone else in the virtual space with me.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

17) I felt like the other person was aware of my presence in the virtual space.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

NMM 18) I hardly noticed the other individual.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

19) The other person hardly noticed me.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

20) I think the other person often felt alone.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

21) I often felt as if I was alone.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

Hauber

22) I was always aware that my partner and I were at different locations.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

23) I was always aware of my partner's presence.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

24) It was just like being face-to-face with my partner.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

25) It felt as if my partner and I were in the same room.

1	2	3	4	5	6	7
Fully Disagree						Fully Agree

26) Any additional comments about your experience?

Date: Participant #:

Post-Experiment Questionnaire

1) Which of the two systems did you find easiest to use?

3D with Video Video Only

2) Why?

3) Which of the two systems made you feel more “present” in the virtual environment?

3D with Video Video Only

4) Why?

5) Which of the two systems made it feel more like the remote partner was present with you?

3D with Video Video Only

6) Why?

7) Which of the two systems did you prefer overall?

3D with Video Video Only

8) Why?