Scientific Research Publishing

Journal of Software Engineering and Applications, 2019, 12, 171-197
http://www.scirp.org/journal/jsea
ISSN Online: 1945-3124
ISSN Print: 1945-3116

Resolutions and Network Latencies Concerning a Voxel Telepresence Experience

Noel J. W. Park, Holger Regenbrecht

Department of Information Science (HCI Lab), University of Otago, Dunedin, New Zealand
Email: noel.park@otago.ac.nz, holger.regenbrecht@otago.ac.nz

How to cite this paper: Park, N.J.W. and Regenbrecht, H. (2019) Resolutions and Network Latencies Concerning a Voxel Telepresence Experience. *Journal of Software Engineering and Applications*, **12**, 171-197.
https://doi.org/10.4236/jsea.2019.125011

Received: April 8, 2019
Accepted: May 28, 2019
Published: May 31, 2019

Abstract

Recent advancements in computing research and technology will allow future immersive virtual reality systems to be voxel-based, *i.e.* entirely based on gap-less, spatial representations of volumetric pixels. The current popularity of pixel-based videoconferencing systems could turn into true telepresence experiences that are voxel-based. Richer, non-verbal communication will be possible thanks to the three-dimensional nature of such systems. An effective telepresence experience is based on the users' sense of copresence with others in the virtual environment and on a sense of embodiment. We investigate two main quality of service factors, namely voxel size and network latency, to identify acceptable threshold values for maintaining the copresence and embodiment experience. We present a working prototype implementation of a voxel-based telepresence system and can show that even a coarse 64 mm voxel size and an overall round-trip latency of 542 ms are sufficient to maintain copresence and embodiment experiences. We provide threshold values for noticeable, disruptive, and unbearable latencies that can serve as guidelines for future voxel and other telepresence systems.

Keywords

Copresence, Embodiment, Mixed Reality, Telepresence, Voxels

1. Introduction

Recent advancements in technology [1] [2] [3] indicate that there is a good chance for future immersive systems to be voxel-based (volumetric pixel) visualizations. For example, Intel corporation uses voxels in their recent technology called "Intel True View" which allows users to watch recorded National Football League (NFL) games from any viewpoints in the stadium [3]. Because users have the freedom to watch the game from any viewpoint, users could watch the NFL

game from the field sideline or even from the middle of the field. Such advancements in technology allow other science fiction ideas to come true, such as telepresence. Telepresence technologies are an efficient way to communicate with remote friends, families, or colleagues, if reaching their location is impossible or hindered by factors such as high cost. They provide a shared virtual environment where users can communicate face-to-face with each other [4] like we would do in the real world.

To be effective, a telepresence system has to provide users with a sense of copresence, *i.e.* the user's sense of being with other people [5] [6] [7] [8]. The level of copresence depicts how convinced users feel that they are with another person. Without this sensation, users would ignore or treat others in the virtual environment as inanimate objects, which makes the system meaningless as a communication medium. Conventional 2D telepresence systems generally provide a convenient way to achieve this with video streams. While in the real world we can look at a person from multiple viewpoints (free viewpoints), 2D telepresence systems confine us to one (staring at a screen). Therefore, researchers have strived to develop 3D telepresence systems through immersive technologies to mitigate such problems. Common approaches for 3D telepresence systems provide users with a first person view using immersive head mounted displays (HMDs). They are often complemented with colour and depth (RGB-D) cameras to capture and render the real user (virtual body representation) [1] [9] [10] [11].

There are currently two popular methods for rendering three-dimensional environments. The first and most common is to use a polygon mesh, which is a surface geometry composed of many primitive shapes (commonly triangles). Although the visualization is currently coarse, the second method is to instead use voxels, which extends the pixel concept into three dimensions [1] [12] [13] [14]. Beck *et al.*'s work [10] provides a polygon mesh based mixed reality telepresence system involving multiple cameras with high quality rendering. However, work by Regenbrecht *et al.* [1] demonstrates that you don't necessarily need high quality rendering to achieve presence. While the sense of presence in our context refers to how convinced users feel that they are in a virtual environment [15] [16] [17], it is a related concept for embodied telepresence experiences. This is because the factors that influence the sense of presence (spatial-presence) and embodiment (self-location) are overlapping concepts [1] [18]. Therefore, we developed a telepresence system extending Regenbrecht *et al.*'s work who also provide arguments for the voxel concept over traditional polygon mesh or point cloud rendering: 1) achieve a convincing coherent mixed reality environment with less effort (but currently at the cost of visual quality); 2) avoid the uncanny valley[1] that polygon meshes might have with high quality humanoid representations; and 3) allows us to utilize pixel based algorithms with an extra dimension, so no need to invent new methods.

[1]The uncanny valley is the eerie human sensation that can possibly arise when interacting with humanoid things. A simple example is humanoid doll where some people can be put off by its close human resemblance.

We implemented a network protocol that allows two systems to exchange their local voxel body data between them. With this function, users are able to visually see other users in a remote location. However, when designing such a telepresence system, there are multiple factors that need to be considered. Steuer's work [15] describes two major quality of service factors in a telepresence system that influences the telepresence experience: vividness and interactivity. Vividness is the system's ability to produce a sensory rich virtual environment, and interactivity is the degree to which users can influence the form and content of a virtual environment. Therefore, we want to know if there are specific thresholds for these quality of service factors to guarantee a voxel telepresence experience. So, we investigate the questions: 1) is there a voxel size (vividness) for maintaining the copresence and embodiment experience; and 2) is there a threshold latency (interactivity) for maintaining the copresence and embodiment experience. These questions were solved as described in each Section: 1) we explored the related work; 2) developed and implemented a working voxel telepresence protocol; 3) ran a user study; 4) performed statistical analysis on our measures; and 5) discussed our findings.

2. Related Work

In this section, we discuss major literature related to our voxel telepresence research system and study. We first explore related user experience concepts in mixed reality: presence, telepresence, copresence, and embodiment. We then expand on the literature related to telepresence technologies with some working 3D telepresence system examples. Finally, the literature on the voxel-based mixed reality system and its related uncanny valley concept are explored.

2.1. Presence and Telepresence

Presence is generally defined as the sense of "being there" in a virtual environment, which is typically experienced through interactive 3D media like virtual reality content and 3D video games [15] [16] [17]. Schubert *et al.* [16] state that when users are present in a virtual environment, the location of the user's body is interpreted as being contained in the space rather than viewing it externally. In other words, this means that users feel that they are located within an environment when present. It is generally understood that this sensation emerges when users can devise possible actions regarding the environment they exist in [16] [19] [20]. Glenberg states that our ability to devise actions comes from two sources [19]: the environmental context and our memory (past experiences). The system's ability to allocate the user's attention towards virtual stimulus is a necessary condition for presence [16] [21]. Based on this concept, two key components of presence arise [1] [7] [16]: *spatial-presence*, and *involvement*. *Spatial-presence* is the sense we are located in and acting from within the virtual environment and *involvement* is the sense that we are concentrating on the virtual environment while ignoring the real environment (or conflicting stimuli). The

same literature also state that realism, the fidelity of the virtual environment to that of the real world, is another a key component for presence. We thus define presence to be comprised of three main components: spatial-presence, involvement, and realism. However, Schubert *et al.* [16] suggest that the realism factor contributes less compared to spatial-presence and involvement.

Telepresence was first introduced as a concept by Marvin Minsky in 1980 as a technology for remote controlled machines [22], generally called teleoperation. Telepresence in his context means that users would feel present in their remote work environment (via machinery) while physically being located elsewhere [7].

Steuer [15] generalizes telepresence as the sense of presence in an environment using a communication medium, or being present by some computer network model. Steuer identifies five factors that expands from vividness and interactivity: *breath* and *depth* for vividness; and *speed*, *range*, and *mapping*. *Breadth* and *depth* are respectively defined as the range of senses stimulated and the resolution of each sensory input provided by a telepresence system. *Speed* refers to the systems response time to user inputs. *Range* determines the number of virtual content and forms that can be manipulated by the user. *Mapping* refers to the way in which human actions are mapped within the virtual environment. Since the human perceptual system is optimized for visualization and interactions in the *real world* [15], these five factors influence how we visualize and interact with the virtual world in comparison.

2.2. Copresence and Embodiment

Copresence is defined as the sense of being with other people in the virtual environment. [5] [6] [7] [8]. Based on Zhao's works [5], copresence is said to consist of two dimensions: copresence as a mode of being with others, and copresence as the sense of being with others.

The first dimension is the form of human colocation in space-time that allows for real-time and mutual human contact. Different forms of copresence depend on the physical distance between interacting individuals, and the corporeal (bodily) presence at the colocation site. In our telepresence system both users are virtually present and are colocated in each other's virtual proximity, so we have a hypervirtual telecopresence mode. Additionally, users are provided with a system interface which they can use to interact with each other. Therefore Zhao [23] specifies four parameters that should be considered when designing such a system interface: 1) *embodiment* refers to the involvement of human bodies; 2) *immediacy* refers to the speed at which messages travel between copresence individuals; 3) scale refers to the number of people enabled by the interface; and 4) mobility refers to how well copresent individuals can interact while in locomotion. The ideal system interface would be one that incorporates the full human body, in real time, and one that allows one-to-many interactions while in locomotion, like how we interact in the real world.

Embodiment is defined as the "subjective experience of using and 'having' a

body" [18] [24]. This is formally defined as follows: where *E* is some alternative representation of a body, "*E* is embodied if some properties of *E* are processed in the same way as the properties of one's body" [24] [25]. To achieve this experience, Kilteni *et al.* suggest three components that influence the embodiment experience [18]: sense of self-location, agency, and body ownership.

The sense of "self-location is a determinate volume in space where one feels to be located" [18]. Regenbrecht *et al.* [1] state that spatial-presence and self-location, "refers to the same perception of spatially being part of an environment". They are indeed similar, but self-location as Kilteni *et al.* describes it is concerned with the relationship between one's self and body, whereas spatial-presence generally refers to the relationship between one's self and the environment; the "sense of self-location specifically refers to one's spatial experience of being inside a body" [18]. However, although self-location and spatial-presence address different spatial questions, they are considered complementary concepts that both constitute one's spatial representation [18] and therefore, can be considered equivalent.

The sense of agency refers to the overall sense of body control; specifically the subjective experiences of action, control, intention, motor selection, and conscious experience of will [18]. The sense of agency is said to arise when one's predicted sensory consequence matches the actual sensory consequence, or depend on the synchronicity of visuomotor correlations [18]. This means that one will generally feel in control of an alternative bodily representation when their bodily actions are coherent with their expected outcome.

The sense of body ownership refers to one's self-attribution of a body which "implies that the body is the source of the experienced sensations" [18]. The sense of ownership is proposed to emerge from the combination of human sensory and cognitive influences. The sensory influences refer to the sensory information that arrives at our brain (like visual, and tactile sensory input), and the cognitive influence refers to sensory information that may be modulated based on your thought process.

2.3. 3D Telepresence

A traditional telepresence system is the 2D teleconferencing systems which are accomplished using live 2D video streaming [26]. However, Kuster *et al.* [26] state that traditional telepresence systems lack realistic user experience in contrast to how we communicate in real life. In particular, they describe three main limitations of traditional telepresence systems. Firstly, users are required to sit in front of a computer, or at least carry a mobile device. This means we can't roam around and communicate like we can in the real world. Being stationary also restricts us from using bodily gestures, so we can't communicate effectively with both non-verbal and verbal communication. Secondly, eye contact doesn't come naturally like real world face-to-face communication [26] [27]. This is because they are generally looking at a screen where the remote user's face is rendered instead of directly at the camera capturing them. Thirdly, traditional telepre-

sence systems typically capture the upper body, so we get incomplete whole body language communication, such as incomplete posture and no depth cues [26].

3D telepresence systems try to mitigate the common problems with traditional telepresence systems. A common visual approach in telepresence systems other than 2D screens are head mounted displays (HMDs). Stereoscopic HMDs allow users to view the world in 3D just like in the real world and allow users to view the world in the first person perspective, which is known to have several benefits. For example, Hauber *et al.* found that the sensation of copresence increased when using 3D views on a modified 2D video-conferencing system [28]. With the 3D views, users gain depth cues which increases vividness (depth), and therefore increases the telepresence experience. Additionally, Kilteni *et al.* state that an egocentric visuospatial perspective (first person view) can enhance the sense of embodiment (self-location, ownership) [18]. For example, studies like Petkova *et al.* [29] showed that "physiological responses to a threat given an artifical body were greater in the first person perspective compared to the third person perspective" [18]. Although HMD technologies, such as the Oculus Rift[2] and HTC Vive[3], are attached to a computer, they provide some space for users to move around in, increases the range of possible user interactions. Additionally, using RGB-D (colour and depth) cameras like the Microsoft Kinects, we can capture people and display them in the virtual environment (mixed reality). This provides a way for non-verbal telecommunication through bodily gestures, and postures in telepresence systems. Furthermore, because we portray the real user in the virtual environment, their sense of embodiment (agency, ownership) can further be enhanced.

Most 3D telepresence combines both HMDs and RGB-D cameras for the above reasons. For example, Beck *et al.*'s [9] earlier work used Microsoft Kinect version 1 for 3D data acquisition which are projected onto a wall on the copy and paste issue fix with: recipient's side. This is further complemented with shutter glasses to provide users with depth perception. Maimone *et al.* [11] used multiple KinectFusion cameras for real-time volumetric 3D capturing of room-size scenes and uses a large parallax monitor for autostereoscopy. Beck *et al.* [10] later proposed a 3D data acquisition method using four Microsoft Kinect v2 cameras using polygon mesh rendering. Along with his sweep-based multi-kinect calibration, we can obtain a high visual quality 360-degree view of a person. Beck *et al.*'s [10] work also mentions compatibility with virtual reality HMD devices (*i.e.* Oculus Rift), which could further increase vividness and interactivity. Regenbrecht *et al.* [1] [30] provide a voxel-based approach using an Oculus HMD and Kinect camera.

2.4. A Voxel-Based 3D Telepresence Approach

As previously mentioned, Schubert *et al.* [16] suggest that high realism is a minor contributor to the sense of presence in virtual reality environments. Regenbrecht *et al.*'s work [1] propose a voxel-based mixed reality approach using a

[2]Link to Oculus Rift webpage: https://www.oculus.com/rift/#oui-csl-rift-games=mages-tale.
[3]Link to HTC Vive webpage: https://www.vive.com/nz/.

Kinect 2.0 camera and an Oculus Rift HMD. Their work showed that their system is able to provide a high sense of presence with low realism using voxels and provide a high sense of embodiment through a convincing virtual body representation. Their system captures the surface of real human bodies with the Kinect camera and virtually renders them in the virtual environment as voxels. Because everything is perceived in 3D (enhances vividness), it provides depth cues which traditional 2D telepresence systems lack. An obvious but a crucial quality of service factor is real-time interaction (speed of interactivity, immediacy, and sense of agency respectively) [5] [15] [18]. Since Regenbrecht *et al.*'s work [1] [30] provide a concrete system framework which mediates the sense of presence, embodiment and copresence (with recorded people), we extend their work to design and implement a voxel-based mixed reality telepresence system (**Figure 1**). Despite these advantages, using a close-view HMD obscures the user's face.

For systems that utilize human embodiment, it is important to be aware of the uncanny valley effect, because it could make systems unusable. The uncanny valley effect leads a user to feel anxious, disgusted, or eerie towards a humanoid object [31]. Mori explains that as humanoid objects become closer to 100% human likeness, our level of affinity (our natural liking for someone) increases until it suddenly drops as we near 100% human likeness. Voxel-based mixed reality systems may not be suspectible to the uncanny valley with high embodiment and copresence levels, which could equate to high human-likeness and high sense of affinity for the other person respectively.

3. System

In this section, we discuss all the concepts and design involved in implementing our working voxel-based mixed reality telepresence (vbMRT) system, including the following: 1) explain the mode of copresence and network design; 2) modelling measureable overall latency; and 3) study implementations for controlling voxel sizes and inducing network latency.

Figure 1. A working voxel-based 3D telepresence prototype based on Regenbrecht's *et al.* work [1] [30]. These pictures depict two users at different locations virtually shaking hands in a shared virtual environment. The middle picture shows an over the shoulder view from one user in the shared space. Pictures from Regenbrecht *et al.* [30].

3.1. Telepresence Implementation

Three specific functions have been implemented: 1) the network protocol for streaming voxels; 2) function to induced network latency; and 3) function to control voxel sizes.

Our mode of copresence is categorized as a hypervirtual telecopresence scenario following Zhao's copresence taxonomy [5]. Both users are in separate physical environments each with their own RGB-D Camera. The software responsible for capturing and voxelizing camera data sends voxel data to our system, where the user's virtual representations are rendered. We then extended the user's visual perception using a computer network so they are within the virtual proximity of each other. This configuration is depicted in Figure 2.

3.2. Voxel Transmission

In our system, a voxel is represented as a 3D position and a colour. The voxel colour is represented using the RGB-A colour space. In total one voxel can be represented in 16 bytes: three floating numbers for the x, y, z positions (12 bytes), and one byte for each of the red, green, blue, and alpha channels (4 bytes in total). A voxel image is stored as a list of voxels. For transmitting voxels, we perform a lossless voxel data compression which reduces the data size from 16 to 9 bytes. More specifically we compress the voxel position data from 12 to 6 bytes. One unit in our system equals one metre, which means the voxel size of 8mm can be represented as 0.008 m (floating point value). Because the maximum decimal precision we require is only up to the 1/1000th decimal place, we multiply 1000.0 (float) to obtain a whole number, essentially we are converting

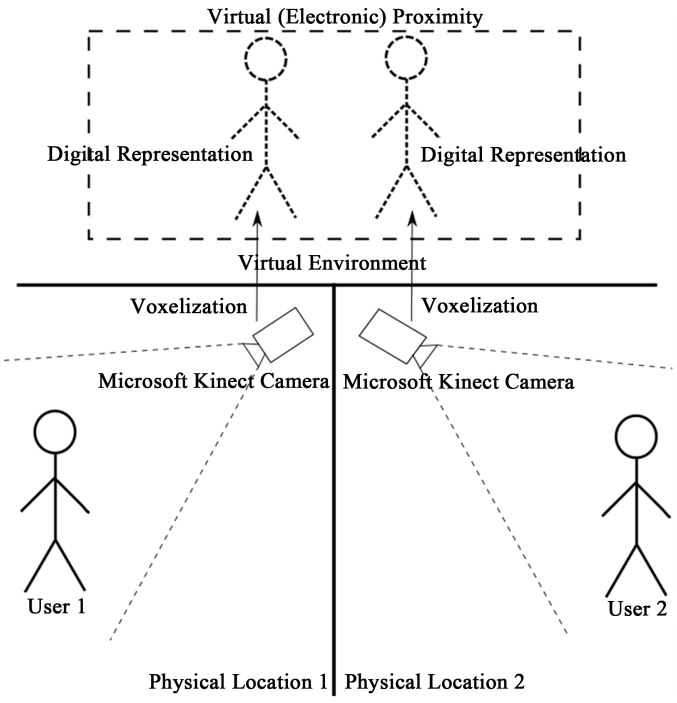

Figure 2. Depicts mode of hypervirtual telecopresence.

metres into millimetres. Then we store the whole number as a 16-bit integer (signed short). It should be noted that this compression only works if the compressed voxel position data size is within the C# short data range (−32,768 and 32,767). But because after compression our voxel positions can be expressed between −1536 and 1536 per axis, based on the 3.072^3 m^3 voxel space (3.072 m × 3.072 m × 3.072 m), we are well within the short data type range. Additionally, the alpha channel data is redundant because it is constant (255) and therefore we only need 3 bytes (RGB). With compression, voxel data consists of [int16 x, int16 y, int16 z, 1 byte red, 1 byte green, 1 byte blue] with a data packet sizes up to 166 voxels (*i.e.* each packet is 9 × 166 bytes large). We reverse our compression method to decompress the voxel data.

We chose UDP (on a 1 gigabit network) for our application because we can tolerate some packet loss assuming it doesn't significantly affect the visual appearance of the voxelized user. Our initial telepresence network protocol transmitted 9 × 7000 bytes per UDP packet (7000 voxels per packet), which is below the theoretical UDP maximum transmission unit (MTU) of 65,535 bytes. But initial tests with University College of London (UCL) over a Wide Area Network (WAN) fragmented the packets due to the ethernet MTU (1500 bytes). Although the IP layer handles packet reassembly in IPv4, the increased throughput with packet fragmentation caused packet drop. Therefore, we reduced the voxel packet size down to 9 × 166 bytes per UDP packet (166 voxels per packet) to avoid packet fragmentation. The average number of voxels per image is approximately 14,000 voxels for a person with average height. This means on average a voxel image frame would total around 208,000 bytes when rendering, or 126,000 bytes when transmitted through the network. Of course the voxel data could be optimized (e.g. with an octree data structure and bitwise encoding) to significantly reduce the frame data size on the network. But because we only use one Kinect camera per user, a list of voxel was sufficient for the study.

3.3. Sender Thread

```
1  while ThreadIsRunning do
2      if NewFrame then
3          obtain LatestFrame;
4          if VoxelCount > MaxVoxelsPerPacket then
5              set PacketBufferSize ← MaxBufferSize ;
6          else
7              set PacketBufferSize ← MaxBufferSize × VoxelCount ;
8          end
9          HasRemainingVoxels ← false ;
10         foreach Voxel in LatestFrame do
11             PacketBuffer ← Compressed VoxelData ;
12             if PacketBuffer is full then
13                 send UDP packet ;
14                 calculate RemainingVoxels in LatestFrame ;
15                 if RemainingVoxels < MaxVoxelsPerPacket then
16                     set PacketBufferSize ← MaxBufferSize × RemainingVoxels ;
17                     raise HasRemainingVoxels ← true ;
18                 end
19             end
20         end
21         if HasRemainingVoxels then
22             send RemainingVoxels ;
23         end
24         send EndOfFramePacket (1 byte packet) ;
25         set NewFrame ← false ;
26     end
27 end
```

Algorithm 1. Sender thread protocol.

The protocol for sending and receiving voxels run on two separate threads concurrently running with the main thread. When the sender thread receives a new frame from the Kinect, we segment the frame into smaller segments (166 voxels per packet). For each voxel, we compress the voxel data in the aforementioned method, write them into the packet buffer and send them to the receiver thread once the packet buffer is full. We continue until all remaining voxels are processed. When all voxels in a frame are processed, a final 1 byte packet is sent to indicate the end of a frame. Indicating an end of frame this way can double up in frames on the receiving end. However, packet loss was not an issue to conduct our study in a local area network (LAN). Additionally, initial tests with UCL over a WAN didn't indicate related issues to packet loss for one Kinect camera. The latest frame data in the sender thread is a critical section, especially if the render (main) thread overwrites the latest frame buffer while the sender thread is still processing it. The C# lock implementation was used to handle this.

3.4. Receiver Thread

```
1  while ThreadIsRunning do
2      receive VoxelPacket from RemoteAddress ;
3      if VoxelPacketLength==1 then
4          if UsingFrontBuffer then
5              BackBuffer ← LatestFrameBuffer ;
6          else
7              FrontBuffer ← LatestFrameBuffer ;
8          end
9          if NewFrame==false then
10             swap buffers ;
11             raise NewFrame flag ← true ;
12         end
13     else
14         while ByteIndex < VoxelPacketLength do
15             VoxelData ← read voxel bytes ;
16             LatestFrameBuffer ← Decompressed VoxelData ;
17         end
18     end
19 end
```

Algorithm 2. Receiver thread protocol.

A dual buffer was used for concurrent reading/writing between the render and receiver thread respectively. In the receiver thread, new voxel data is buffered in a local frame buffer until the 1 byte (end of frame) packet is received. If so, we allocate the latest frame buffer to the unused one in the render thread. The front and back buffer are swapped only when both are unused by the render and receiver thread. The new frame flag is then raised indicating the render thread to read the latest frames. The buffer is swapped using a boolean variable which toggles between the front and back buffer for the render thread to use. The new frame variable is also controlled by a C# lock implementation in case the buffers are prematurely swapped, while the render thread reads from it.

3.5. Voxel Size and Latency Control

We want to be able to change the voxel size and network latency to determine if there are threshold values for maintaining the copresence and embodiment ex-

perience. Figure 3 models the round-trip process from when user 1 initiates an action to when he/she perceives user 2's responding action. Each component represents the intermediate tasks the system performs which adds onto the overall latency. Some components can practically be measured such as the latency for sending one voxel packet, or can be predetermined such as the Kinect camera latency from the frame rate (30 Hz/33 ms). Other components like measuring the time it takes to transmit Kinect data across via the internal bus (hardware level) is a lot harder, so we assume that any hardware related measures are less than 1 ms.

Initial measures from Regenbrecht *et al.* [1] show that the overall standalone system latency is about 40 ms. They also measured that the overall latency doesn't vary until the system starts rendering 700,000 voxels, which we are well below. Although the measured 40 ms latency is based on the user's self-perception, it is still applicable in the telepresence context because the same rendering process for perceived self-actions occur symmetrically for user 2. Therefore by doubling the self-perceived latency, we get an estimate overall round-trip latency of 80 ms assuming the user responds instantly, which is unpredictable. The ping test was used to determine the round-trip latency for transmitting 166 byte packets. In our LAN environment, the ping test reported a round-trip latency of less than 1ms, and therefore we assume no latency for sending a voxel packet. The latency would vary the most when transmitting data in WAN (Figure 3), so we induce latency in the network for our study.

Our study will only explore symmetrical latency (where both users experience the same amount of latency). When controlling induced latency in the system, it's a common mistake to simply sleep the sender thread for a fixed amount of time. This would cause network stuttering instead of playback latency. We simulate latency using a dynamically sized ring buffer in the main thread. As you increase latency the ring buffer will get larger and therefore the cycle time will take longer. By cycle time, we mean the time taken to traverse through the ring buffer once from an initial starting point. The same ring buffer implementation was used to induce self-perceived latency. For real-time LAN telepresence, this ring buffer would be set to size 1, so it sends the most recent frame as soon as possible. Each ring buffer increment can be assumed to add about 33 ms in latency because of the Kinect frame rate (30 Hz).

A TCP socket was used to control the network latency symmetrically from one computer using three integer values: 0 indicating that nothing needs to be done; 1 to add latency; and 2 to reset the latency. Because the control message data size is small, we only needed a byte sized TCP packet. TCP was necessary for this instance because we wanted reliable network latency control. Otherwise, in the event of packet loss, both sides would have asymmetrical network latency.

We used the existing Kinect (data) Capture application with multiple executables representing different voxel sizes (or resolutions). Three different voxel sizes were used to for a uniform capture volume of 3.072^3 m^3. A gapless voxel

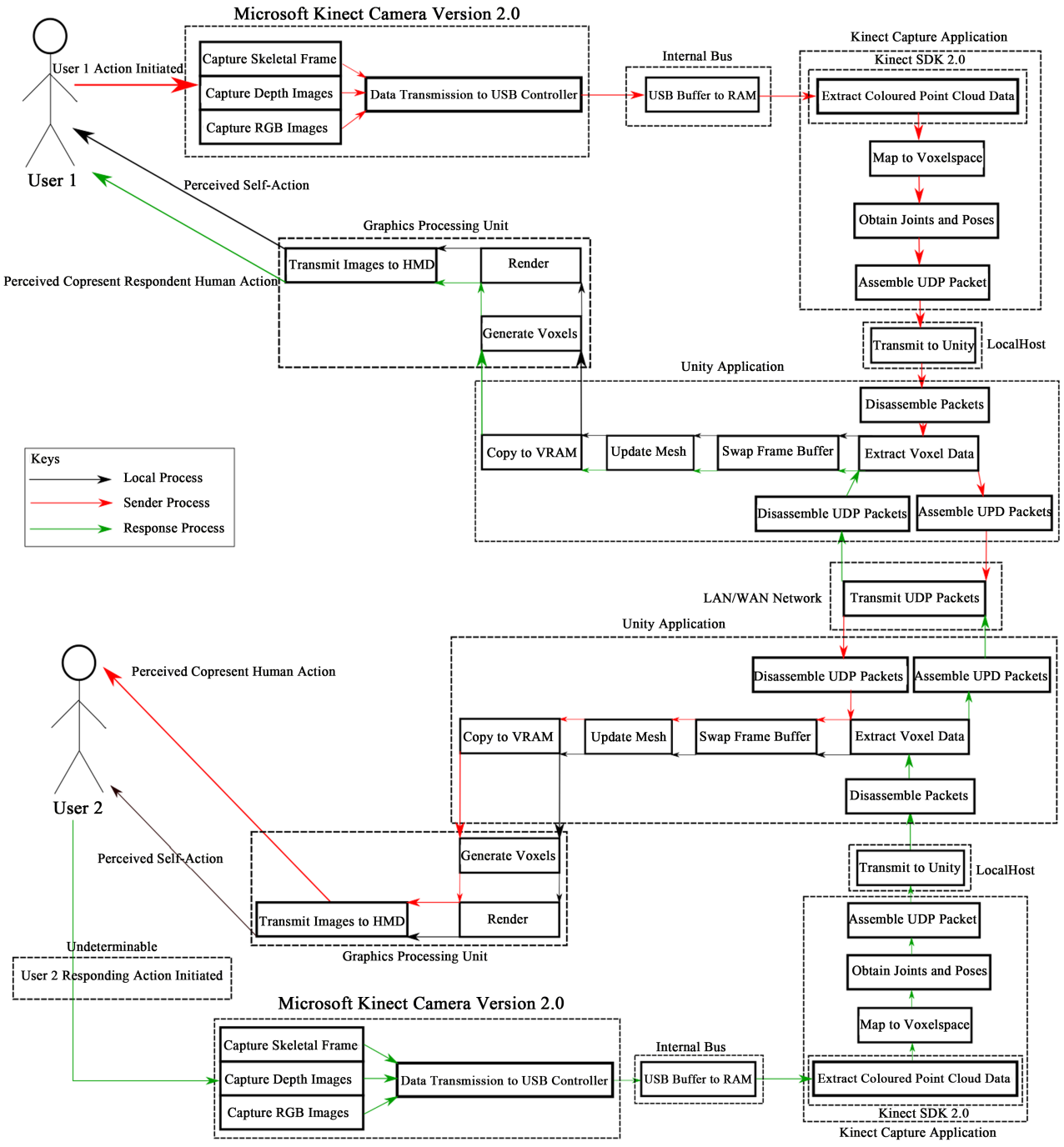

Figure 3. Overall perceived latency (round-trip latency) based on User 1. Following the read arrow from User 1 and then the responding green arrow from User 2 (and vice versa), we can trace the paths and sections of the system where latencies may occur.

representation must be maintained when changing voxel sizes. Therefore in the study, we are restricted to voxel sizes divisible by 3.072 m with no remainder. The visual outcomes for five different usable voxel sizes are shown in **Figure 4**.

4. Methodology

This chapter details the experimental design that was used to obtain our

(a) (b) (c) (d) (e)

Figure 4. Examples of five different voxel sizes. (a) 8 mm; (b) 12 mm; (c) 16 mm; (d) 24 mm; (e) 32 mm.

copresence, embodiment, network, and self-perceived latency measures. Three tasks are described for obtaining our measures in the following order: charades, social interaction, and self-perceived latency tasks.

4.1. Independent Variables

The first independent variable targets the different voxel sizes. 8 mm, 32 mm, and 64 mm voxel sizes are explored in this study. Although Regenbrecht *et al.*'s studies [1] [30] show that 8 mm voxel size can provide the copresence and embodiment experience, we want to further verify this in the telepresence context with coarser voxels. An initial pilot study indicated that changes in usable voxel sizes between 8 - 32 mm and 32 - 64 mm were not really noticeable. We also don't explore voxel sizes above 64 mm because it produced blocky body representations and thought that this would be threshold.

The second independent variable is the induced network latency, which simulates network latency by delaying the playback of the copresent individual. For example, if we induce network latency in the voxel streaming, then an action initiated by the local user will be perceived by the copresent user after some delay. The same method is used to induce latency on the system itself so that the user's virtual body perception is delayed (self-perceived latency). Network and self-perceived latency are induced separately for two different latency experiments (social interaction and self-perceived latency task).

4.2. Dependent Variables

The two main dependent variables measured are copresence and embodiment using scale questionnaires from Bailenson *et al.* [32] and Llorbera *et al.* [33]. Additionally, signs of discomfort were measured using simulator sickness questionnaires from Kennedy *et al.* [34]. For both the telepresent and non-telepresent tasks, the noticeable, disruptive, and unbearable latency values were obtained based on participants' observations.

Each participant completed four copresence and embodiment experience questionnaires: three for charades task and one at the end of the network latency

task. At the end of the study session, participants filled out the simulator sickness questionnaire and three follow-up questions which asked: 1) yes/no for the first task being difficult; 2) if yes for previous question, which round of charades would they rate the hardest (1, 2, or 3); and 3) rate voice communication quality on a scale between 1 to 5 (inclusive).

4.3. Experiment Design

We used a within-subjects design where all participants were exposed to all three voxel sizes in random order, and incremental network/self-perceived latencies. The study required two live participants per session. 18 males and 18 females, Mage = 21, SDage = 3.58) were selected from the University of Otago using email advertisements, and word of mouth. The order that participants were exposed to different voxel sizes were randomized using a true random number generator[4], which determined the participant numbers for balanced exposures. Participants were individually given $20 vouchers for their time.

4.4. Tasks

We produced three tasks for our measures: 1) Charades task; 2) social interaction (network latency) task; and 3) self-perceived latency task. The first two tasks were in a telepresent scenario while the last task wasn't.

For the Charades[5] task, we picked 40 Charade items (20 items per participant) that were assumed to be familiar actions or objects for people to easily act out (e.g. crocodile or swimming). One charade game consisted of two charade rounds. Three games were played for each voxel size exposure. Before the beginning the charade games, participants picked two charade items from their corresponding lists. Two rounds were necessary because the exposure duration was too short (only few seconds) for the participant to experience the voxel sizes. For each game, participants performed their two charade items in alternating turns.

In the social interaction task, both participants were asked to coordinate a *high five*, and report back on whether they noticed anything different in performing the high five. After each prompt latency was incremented until the noticeable, disruptive and unbearable network latency (NNL, DNL, and UNL respectively) thresholds were obtained. These were measured only when both participants came to a consensus. We defined the disruptive threshold as the point where participants start to notice their deteriorating ability to perform the task, while still tolerating the latency. Therefore, we believe that this would be the threshold for maintaining copresence and embodiment levels. The unbearable point would be when participants couldn't tolerate the latency to complete the task. The task needed to become difficult as latency increases. Additionally, the latency is best noticed if the task required both participants to coordinate in some way. A high five action was used for initially determining the network la-

[4]https://www.random.org/.

[5]A game where one or more players guess the acting of another to identify the action or object being portrayed.

tency thresholds because participants were able to identify increasing latency. Each participant took turns counting down from three before initiating a high five and waiting for the other's response. For the participant who initiated the high five, they should notice the other's response slowing down with increasing network latency. After all threshold values were determined, participants were asked to perform other social interactions such as hugging while influenced by their reported DNL value.

The self-perceived latency is the delayed perception of participant's virtual body. Because this latency is induced on the participants individually and relative to their own body movement, the effects are more easily noticed. The self-perceived latency was induced in a similar way as the network latency experiment. Participants were asked to wave their arms back and forth in front of them. As the self-perceived latency increased, they were asked to identify when they felt that that latency was noticeable, disruptive to their movement, and unbearable, respectively.

4.5. Procedure

Our procedure is segmented into five stages sequentially: the preliminary; three tasks; and post study questionnaires.

Two experimenters were required because participants were placed in two adjacent rooms separated by a closed door. The leading experimenter coordinated the whole experiment and assisted the participant in the primary room, while the other experimenter assisted the participant in the secondary room. When both participants arrived, they were led into the primary room for them to: read a study information sheet; to sign a consent form; and fill out a demographics sheet. They were then informed on how to fill the questionnaires, complete the charades task and then one participant was led into the secondary room. Participants were expected to fill the first three question sheets after each charade game (per voxel size exposure), and the last was filled after the social interaction tasks. In each room, experimenters helped each participant wear the Oculus Rift HMD. Participants were then introduced to the virtual environment where they should see the other participant and be able to talk through the voice chat. We used LAN enabled TeamSpeak for communication.

Charades task (5 - 10 mins): After participants were introduced to the system, they were then asked to pick their first two Charades items for the first game[6]. The participants took alternating turns where they each acted their chosen Charade items for the other to guess. If participants took too long (more than a minute), the charade item acted out was revealed. When participants completed their turns, they were asked to fill the first question sheet, while the next voxel size exposure was prepared for the second Charade game. This was repeated until all three voxel sizes were exposed.

[6]It should be noted that it would have been more elegant to provide the Charade item list inside their HMD view, but that would involve participants using the Oculus touch controllers (hand-held device input). For the study purpose, we wanted to produce a hands-free telepresence application.

Social interaction task (10 - 15 mins): After participants finished the third question sheet, the high five and social tasks were explained back in the primary room. They were informed that network delays will be simulated during this task. Then the participants were lead back to their corresponding rooms and fitted with the HMD again. The latency was increased one ring buffer frame each time (33 ms) and participants were asked every time if they thought they reached the thresholds. After it was measured, their identified DNL value was set and then were asked to try other social interactions from our list: hugging, fist bumping, handshaking, patting the other's shoulder, head, and pinching their cheeks. They then filled out the fourth question sheet based on their social interaction experiences, while influenced by their reported DNL.

Self-perceived Latency task (5 - 10 mins): The system was configured for the last experiment while the participants filled their last question sheet. Because the last task is non-telepresent, they were run separately in each room. They were asked to identify the same latency threshold values but based on their self perception: noticeable, disruptive and unbearable self-perceived latency (NSL, DSL and USL respectively). When participants were fitted with the HMD again, latency was slowly increased while they waved their arms back and forth at a steady pace. Each reported threshold was recorded before adding more latency until the USL threshold. In the end, they were asked to fill out the post study question sheet (simulation sickness and follow up questions).

Post study: When both participants finished filling their post study question sheet, they were brought back to the primary room. The leading experimenter checked all questionnaires for completeness, while the other experimenter gave them their $20 vouchers for their time. Then participants were asked not to discuss their question choices with others if they personally knew other participants in the study. Afterwards, they were thanked for their time and released. The whole procedure approximately took 25 - 45 minutes depending on how fast participants completed their tasks.

5. Results

We report on: 1) overall embodiment and copresence; 2) Network latency; and 3) Self-perceived latency.

5.1. Statistical Analysis

Data analysis was performed using Microsoft Office 365 Excel and R. Excel was used for recording/pre-processing data and normality/significance tests were performed using R with 95% confidence. Reported likert scale embodiment (B) measures ranged between 1 - 10 and copresence (C) range between 1 - 6. There were four embodiment and three copresence questions on a question sheet. For each participant, the overall individual embodiment and copresence average was computed based on the question sheet measures. We label these per participant copresence and embodiment averages: 1_B; 2_B; 3_B; 4_B; 1_C; 2_C; 3_C; and

4_C. Then the overall copresence and embodiment average were obtained based on the per participant averages (suffixed with -_AV). The numerical prefixes (1, 2, 3, 4) denotes the four (8 mm, 32 mm, 64 mm voxel sizes, and the DNL) exposures respectively.

Initially the Kolmogorov-Smirnov (KS) test was used for testing data normality, then the Shapiro-Wilk (SW) was used to verify the KS test p-values. Because both tests showed different results, a third normality test was performed using the Anderson-Darling (AD) test. The KS test demonstrated that all the data was normally distributed and the SW and AD tests reported differently to it. Therefore the skewness, kurtosis values, and QQ plots were used to verify the normality test claims. The QQ plots reported that our sample data was left-skewed, which is consistent with the skewness values. This also makes sense because most of our measures were above the likert scale midpoint. It was also hard to interpret data distribution tailedness with the QQ plots, but the excess kurtosis values indicate a mixture of light and heavy tailedness.

Specifically, the 1_B, 3_B, 1_C, and 3_C datasets were measured to be platykurtic (less-tailed). 2_B, 4_B, 2_C, and 4_C datasets were measured to be leptokurtic (more-tailed). For these reasons, we cannot assume that all the data is normally distributed (reject assumption for parametric data). Therefore, we used the Wilcoxon signed-rank test with continuity correction using R, because it compares significant differences between measures based on their overall median (see Table 1 for results). Compared to mean based tests, using median significant difference tests is considered more powerful for non-parametric data.

5.2. Overall Copresence and Embodiment Results

All per participant copresence and embodiment measures were compared in pairs for significant differences. It could be considered that 1_B and 1_C measures were obtained under default the system voxel size (8 mm) and network

Table 1. Significant difference test results table. The results from a dependent Wilcoxon signed-rank test with continuity correction (computed using R). P-values computed comparing per participant averages for each exposure pair and their midpoint.

Comparisons	p-values	Midpoint Comparisons	p-values
1_B vs 2_B	250.008219	1_B vs Mid = 5.0	1.862e−07
1_B vs 3_B	9.923e−06	2_B vs Mid = 5.0	1.683e−06
1_B vs 4_B	0.6569	3_B vs Mid = 5.0	0.01666
2_B vs 3_B	0.0006768	4_B vs Mid = 5.0	3.644e−07
1_C vs 2_C	0.003595	1_C vs Mid = 3.0	2.338e−07
1_C vs 3_C	0.0006586	2_C vs Mid = 3.0	2.394e−06
1_C vs 4_C	0.2169	3_C vs Mid = 3.0	4.012e−06
2_C vs 3_C	0.03155	4_C vs Mid = 3.0	1.862e−07

latency, so we expect these measurements to have the highest copresence and embodiment levels. Consequently, we compared 4_B and 4_B to these. All overall measurements were compared to their midpoint following the likert scale analysis examples from [1] [7] [23] [35] [36]. All significant test results and effect sizes can be found in Table 1 and Table 2 respectively.

Embodiment results: Based on overall averages (Figure 5), the highest level of embodiment was measured with 8 mm voxel size (M = 7.65, SD = 1.26). 32 mm measured the second highest embodiment level (M = 7.09, SD = 1.54) and

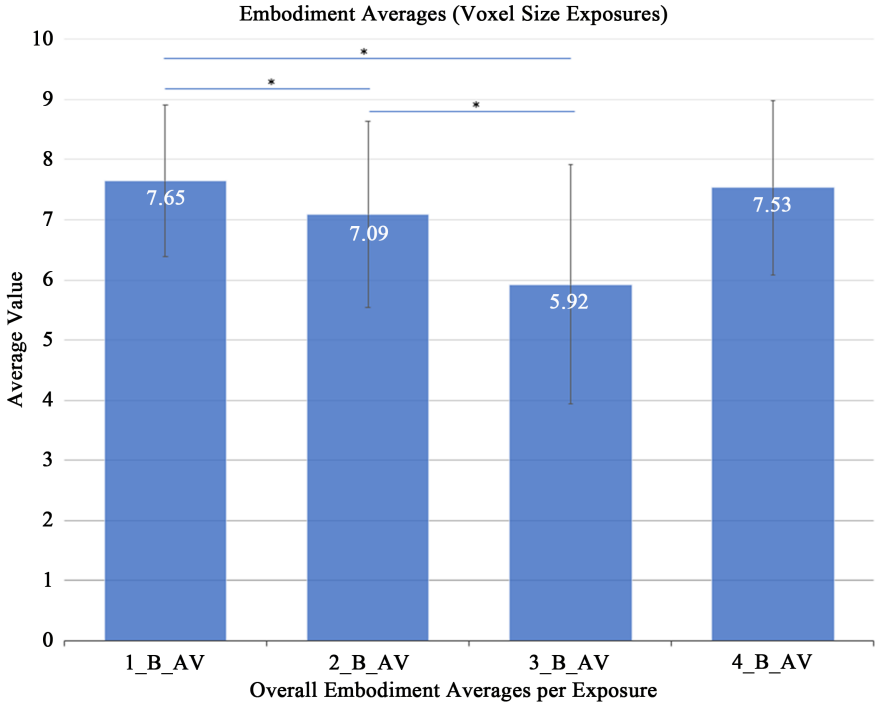

Figure 5. Overall embodiment averages based on a 10-point Likert scale. The asterisks indicate significant differences between different exposures.

Table 2. Effect size results table. This table reports the effect size for each Wilcoxon signed rank test. The Pearson's r was compared using the effect size thresholds for the *absolute values* of r defined by Cohen: $0.1 \leq r < 0.3$ (Small); $0.3 \leq r < 0.5$ (Medium); $0.5 \leq r$ (Large).

Comparisons	Effect Size	Midpoint Comparisons	Effect Size
1_B vs 2_B	Medium	1_B vs Mid = 5.0	Large
1_B vs 3_B	Large	2_B vs Mid = 5.0	Large
1_B vs 4_B	None	3_B vs Mid = 5.0	Large
2_B vs 3_B	Large	4_B vs Mid = 5.0	Large
1_C vs 2_C	Medium	1_C vs Mid = 3.0	Large
1_C vs 3_C	Large	2_C vs Mid = 3.0	Large
1_C vs 4_C	Small	3_C vs Mid = 3.0	Large
2_C vs 3_C	Medium	4_C vs Mid = 3.0	Large

then followed by embodiment at 64 mm (M = 5.92, SD = 1.99). However, when examining all overall embodiment averages for each exposure, the embodiment measured with DNL actually obtained the second highest levels (M = 7.53, SD = 1.45). Embodiment levels were significantly different when comparing 8 mm voxel size (MDN8 = 7.75) to 32 mm (MDN32 = 7.38, p = 0.008219, r = −0.44), and 64 mm (MDN64 = 6.00, p = 9.923e−06, r = −0.74). Embodiment levels were also significantly different when comparing 32 mm to 64 mm (p = 0.0007, r = −0.57). However, the embodiment levels were not significantly different when comparing 8 mm voxel size to DNL measurements (MDNDNL = 7.50, p = 0.66, r = −0.07). All embodiment levels were significantly different compared to the midpoint (5.0): (1_B) p = 1.862e−07, r = −0.87; (2_B) p = 1.683e−06, r = −0.80; (3_B) p = 1.862e−07, r = −0.87; (4_B) p = 1.862e−07, r = −0.87.

Copresence results: Based on overall averages (Figure 6), the highest level of copresence was measured with 8 mm voxel size (M = 5.10, SD = 0.79). 32 mm measured the second highest copresence level (M = 4.76, SD = 0.88) and then followed by copresence at 64 mm (M = 4.48, SD = 1.12). However, when looking at all copresence averages from all experiments, the copresence measured with DNL actually obtained the second highest levels (M = 4.82, SD = 1.32). Copresence levels were significantly different when comparing 8 mm voxel size (MDN8 = 5.00) to 32 mm (MDN32 = 4.67, p = 0.0036, r = −0.49), and 64 mm (MDN64 = 4.83, p = 0.00066, r = −0.57). Copresence levels were also significantly different when comparing 32 mm to 64 mm (p = 0.032, r = −0.36). However, the copresence levels were not significantly different when comparing 8 mm voxel size to DNL measurements (MDNDNL = 5.00, p = 0.22, r = −0.21).

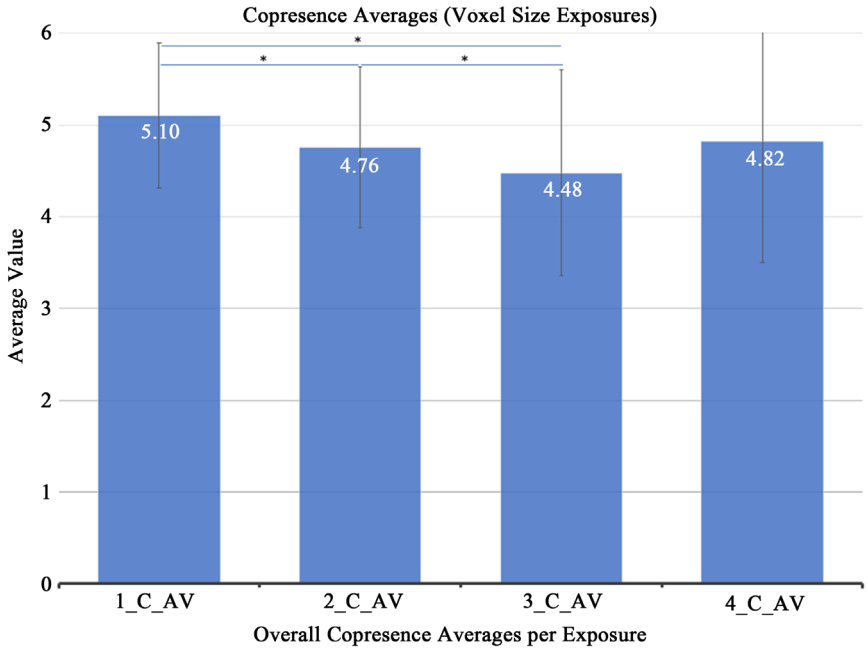

Figure 6. Overall copresence averages based on a 6-point Likert scale. The asterisks indicate significant differences between different exposures.

All copresence levels were significantly different compared to the midpoint (3.0): (1_C) p = 2.34e−07, r = −0.86; (2_C) p = 2.42e−07, r = −0.86; (3_C) p = 2.394e−06, r = −0.79; (4_C) p = 4.012e−06, r = −0.77.

5.3. Latency Threshold Values

Network latency results: The overall average, and medians for NNL, DNL, and UNL were 224 ms, 462 ms, 726 ms, and 224 ms, 462 ms, 677 ms respectively. There were 2 outliers (627 ms for NNL and 1386 ms for UNL) detected using box and whisker (BW) plots (Figure 7). For NNL values, 50% fall within the 272 ms upper quartile (UQ) and 165 ms lower quartile (LQ) bounds with a 107 interquartile range (IQR). The top 25% values fall within the 272 ms UQ and 330 ms maximum. The bottom 25% values fall within the 165 ms LQ and 99 ms minimum. Overall, we have a range of 231 for all observed NNL values. For DNL values, 50% falls within the 561 ms UQ and 363 ms LQ bounds with a 198 IQR. The top 25% values fall within the 561 ms UQ and 693 ms maximum. The bottom 25% values fall within the 363 ms LQ and 264 ms minimum. Overall, we have a range of 429 for all observed DNL values. For UNL values, 50% fall within the 833 ms UQ and 553 ms LQ bounds with a 280 IQR. The top 25% values fall within the 833 ms UQ and 1221 ms maximum. The bottom 25% values fall within the 553 ms LQ and 363 ms minimum. Overall, we have a range of 858 for all observed UNL values.

Self-perceived latency results: The overall average and medians for NSL, DSL, and USL were 135 ms, 255 ms, 47 ms, and 132 ms, 264 ms, 462 ms respectively. There were 2 outliers reported, one from NSL (297 ms) and DSL (495 ms) BW plots (Figure 8). For NSL values, 50% fall within the 165 ms UQ and 99 ms LQ bounds with a 66 IQR. The top 25% values fall within the 165 ms UQ and 264 ms maximum. The bottom 25% values fall within the 99 ms LQ and 0ms minimum. Overall, we have a range of 264 for all observed NSL values. For DSL values, 50% falls within the 297 ms UQ and 198 ms LQ bounds with a 99 IQR.

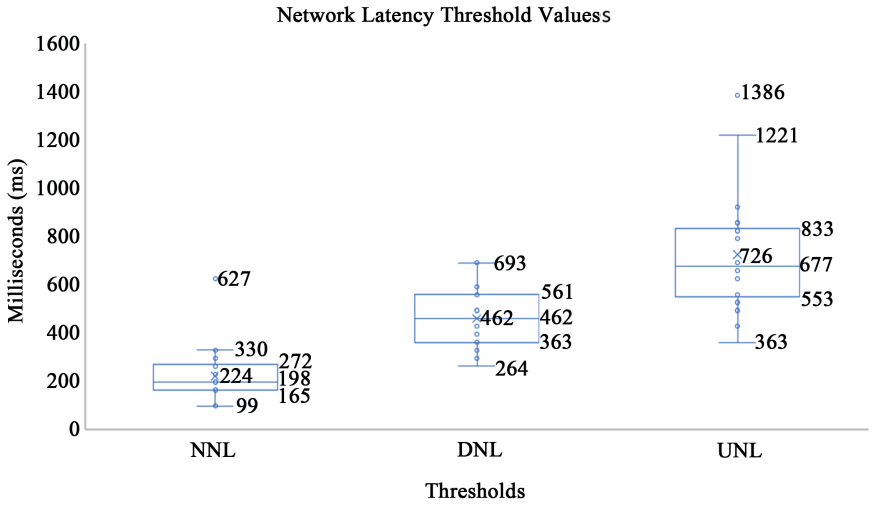

Figure 7. Network latency threshold values.

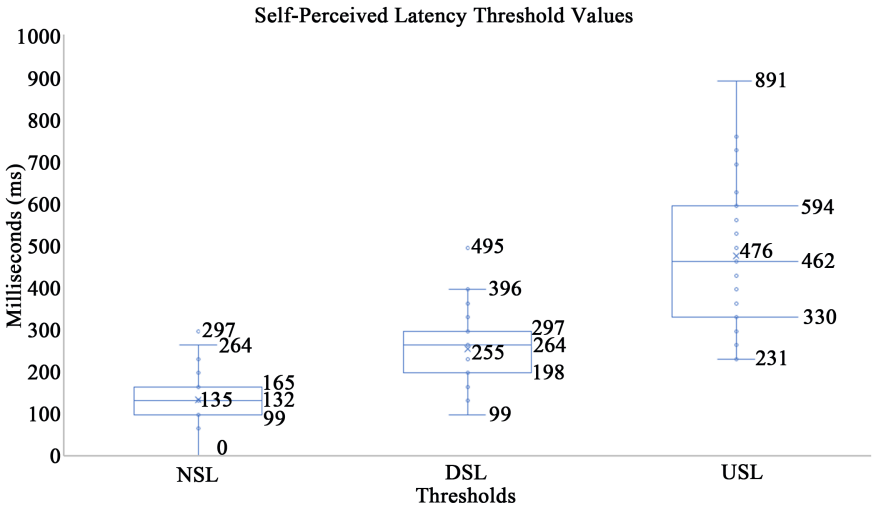

Figure 8. Self-perceived latency threshold values.

The top 25% values fall within the 297 ms UQ and 396 ms maximum. The bottom 25% values fall within the 198 ms LQ and 99 ms minimum. Overall, we have a range of 297 for all observed DSL values. For USL values, 50% falls within the 594 ms UQ and 330 ms LQ bounds with a 264 IQR. The top 25% values fall within the 594 ms UQ and 891ms maximum. The bottom 25% values fall within the 330 ms LQ and 231 ms minimum. Overall, we have a range of 660 for all observed USL values.

6. Discussion

The results show significant differences between each embodiment measured under all three voxel size exposures with medium to large effect sizes. All embodiment levels were significantly higher than the scale midpoint (including DNL embodiment measures), so increasing the voxel size did decrease embodiment levels. Additionally with these results, we can show embodiment levels are still maintained even at 64 mm. Most participants preferred the 8 mm voxel size based on comments such as: "it can get really blocky at times but for most of the experience it wasn't that distracting"; "I felt as though I was actually there with the other person when the video was higher quality"; and "Very realistic, especially 2nd round (8 mm)". Unexpectedly, some preferred the 32 mm over 8 mm: "the other person appeared more complete" (gapless) and "I found a lower quality to cause my arms to feel like mine because of fewer holes [sic]". To clarify, the "lower quality" here refers to changing voxel size from 8 to 32 mm. With lower voxel sizes, the gaplessness is limited by Kinect camera limitations, such as depth camera resolution. Although there are significant differences between 8 - 32 mm embodiment measures, this is one reason why the overall average embodiment levels for 32 mm voxels similarly high as 8 mm. Participants mostly commented on how "pixelated", or "blocky" their virtual body representations were at 64 mm, and a reason for lower embodiment level compared to 8 and 32 mm. Consequently, there were a few participants who couldn't guess what the

other person was trying to act out in Charades. However, this is not always true because a few participants were generally rough at the game, which is why the post study questions asked if any of the Charade games were difficult. This was to verify if any charade items weren't difficult to act out. It can also be implied that participants were rough at the game from being semi-uncomfortable acting out in an unfamiliar environment.

The results also show significant differences between copresence measured under all three voxel size exposures. Although empirically the overall copresence levels don't differ much, comparisons between them show medium to large effect sizes. Additionally, because all copresence levels were significantly higher than the scale midpoint (including DNL exposure measures), we can show copresence levels are still maintained even at 64 mm. This also shows that increasing voxel sizes decrease copresence levels. One participant commented: "Really cool experience. Never had experienced seeing myself or others, only had used for other virtual characters [sic]". So we could imply that the system generally achieves a high copresence experience.

There were no significant differences when comparing overall embodiment and copresence levels under the DNL and 8 mm voxel size exposure. Considering 8 mm voxel size achieved the highest levels, the no significant difference means that DNL measured just as high levels as 8 mm. Although empirically we see a small difference between the averages (lower DNL measures), the results imply that disruptive latency was not the lowest threshold for maintaining copresence and embodiment levels.

The average self-perceived latency threshold measures were lower than the network latency thresholds: 135 ms (NSL); 255 ms (DSL); and 476 ms (USL). This was expected because users would notice latency easier with their own movement.

We proposed a new concept and implemented a prototypical 3D telepresence system. In our voxel telepresence system, users felt copresent in a virtual environment while they were physically in other locations. We then experimented with the voxel size (vividness) and network latency (interactivity) to see if there were thresholds where the sense of embodiment and copresence were maintained. We found that a large 64 mm voxel size still maintained the sense of embodiment and copresence in users. We also found that with a DNL of 462 ms maintained the sense of embodiment and copresence in users. But because a 64 mm voxel size still maintains embodiment and copresence levels, it is possible to explore even larger voxel sizes. Additionally, we found out network latency threshold values (NNL, DNL, and UNL) to be 224 ms, 462 ms, and 726 ms that maintained copresence and embodiment experience levels. More specifically, based on the overall latency model (**Figure 3**), the overall round-trip threshold latency would be 304 ms, 542 ms, and 806 ms respectively (add 80 ms) assuming the immediate response time from the responding user. But because at a 542 ms round-trip latency the embodiment and copresence levels were high, it is possi-

ble to explore above the disruptive latency threshold. These threshold values may be externally valid for voxel telepresence systems where users perform full body social interactions. For more competitive tasks, such as telepresence gaming, the threshold values could be lower. Our reported self-perceived latency measures 135 ms (NSL), 255 ms (DSL), and 476 ms (USL) are lower than our network latency thresholds.

There are many possible further developments to consider. For example, we could align our architecture more towards Steuer's telepresence view; a server and client model which forms a centralized virtual environment for users. This design would provide an easier way to control objects influenced by user inputs compared to our peer-to-peer approach; where changes in the world would be maintained by both client systems. Additionally with a centralized virtual environment, we could investigate if there are standard environmental aesthetics that are required to maintain copresence. This aligns with Campos-Castillo *et al.*'s work [6] who state that the virtual environment's visual aesthetics also increase copresence. We could perform this same study with UCL in a WAN to obtain results in a realistic scenario. Because in this study we are using participants that live in different geolocations, we are enforced to use random participant pairs. In later studies, we could extend a pair of participants to multiple groups in different geolocations.

We could continue to refine our voxel size and threshold values by performing the study again on larger voxel sizes, and use our measured UNL. This would show if we are able to reach the point where embodiment and copresence are no longer maintained. We could also design a similar study for asymmetrical and random latency, which would have better external validity. Additionally, we could investigate methods to improve the photorealism (smaller voxel sizes), and see how if it improves our current embodiment and copresence levels. Smaller voxel sizes are hard to achieve with today's technologies, however we will investigate methods to improve the system's default overall network latency below approximately 80 ms. Because with the Kinect camera at least a 33 ms delay comes from its frame rate (30 Hz), we will investigate other RGB-D camera devices, such as the Intel RealSense[7], which can provide higher frame rates (60 Hz).

More complex future works could investigate incorporating haptic feedback to improve the sensory vividness in the system. Other areas to improve sensory vividness could be looking into HMD alternatives, that don't occlude facial expressions. We could also try improving the sensory interactivity by investigating hands-free user interactions methods so users can manipulate objects in their virtual world.

7. Conclusion

In summary, we provided implementation details on a voxel-based telepresence system, demonstrated its effectiveness, and determined resolution and latency

[7]Link to the Intel RealSense webpage: https://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html.

measures influencing the underlying defining concepts of copresence and embodiment.

Acknowledgements

We thank all participants who volunteered for our study. We specifically thank Jacob Young for helping us run the study, and Sebastian Friston (from UCL) for allowing us to test our system on a WAN (between London and New Zealand). We additionally thank the Information Science Department Staff members (Gail Mercer and Heather Cooper) for helping us prepare participants and vouchers. Finally, we thank the HCI Lab group in Otago for providing resources and feedback.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Regenbrecht, H., Meng, K., Reepen, A., Beck, S. and Langlotz, T. (2017) Mixed Voxel Reality: Presence and Embodiment in Low Fidelity, Visually Coherent, Mixed Reality Environments. *Proceedings of the* 2017 *IEEE International Symposium on Mixed and Augmented Reality*, Nantes, 9-13 October 2017, 90-99. https://ieeexplore.ieee.org/document/8115408/ https://doi.org/10.1109/ISMAR.2017.26

[2] Hoetzlein, R.K. (2016) GVDB: Raytracing Sparse Voxel Database Structures on the GPU. *Proceedings of High Performance Graphics*, Goslar Germany, Germany, 109-117. https://dl.acm.org/citation.cfm?id=2977351

[3] Corporation, I. (2018) NFL+Intel True View. https://www.intel.com/content/www/us/en/sports/nfl/overview.html

[4] Buxton, W.A.S. (1992) Telepresence: Integrating Shared Task and Person Spaces. In: Booth, K.S. and Fournier, A., Eds., *Proceedings of the Conference on Graphics Interface*, Morgan Kaufmann Publishers Inc., San Francisco, CA, 123-129. https://dl.acm.org/citation.cfm?id=155309

[5] Zhao, S. (2003) Toward a Taxonomy of Copresence. *Presence*: *Teleoperators and Virtual Environments*, **12**, 445-455. http://www.mitpressjournals.org/doi/10.1162/105474603322761261 https://doi.org/10.1162/105474603322761261

[6] Campos-Castillo, C. (2012) Copresence in Virtual Environments. *Sociology Compass*, **6**, 425-433. https://doi.org/10.1111/j.1751-9020.2012.00467.x

[7] Nowak, K.L. and Biocca, F. (2003) The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments. *Presence*: *Teleoperators and Virtual Environments*, **12**, 481-494. http://www.mitpressjournals.org/doi/10.1162/105474603322761289 https://doi.org/10.1162/105474603322761289

[8] Schroeder, R. (2002) Copresence and Interaction in Virtual Environments: An Overview of the Range of Issues. *Presence* 2002: 5*th Annual International Workshop on Presence*, Porto, Portugal, 9-11 October 2002, 274-295. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.105.3059

[9] Beck, S., Kunert, A.A., Kulik, A. and Froehlich, B. (2013) Immersive Group-to-Group Telepresence. *IEEE Transactions on Visualization and Computer Graphics*, **19**, 616-625. https://ieeexplore.ieee.org/document/6479190/ https://doi.org/10.1109/TVCG.2013.33

[10] Beck, S. and Froehlich, B. (2017) Sweeping-Based Volumetric Calibration and Registration of Multiple RGBD-Sensors for 3D Capturing Systems. 2017 *IEEE Virtual Reality*, Los Angeles, CA, 18-22 March 2017, 167-176. https://ieeexplore.ieee.org/document/7892244/ https://doi.org/10.1109/VR.2017.7892244

[11] Maimone, A. and Fuchs, H. (2012) Real-Time Volumetric 3D Capture of Room-Sized Scenes for Telepresence. *Proceedings of the* 2012 3*DTV-Conference: The True Vision—Capture, Transmission and Display of* 3*D Video* (3*DTV-CON*), Zurich, 15-17 October 2012, 1-4. http://ieeexplore.ieee.org/document/6365430/ https://doi.org/10.1109/3DTV.2012.6365430

[12] Maldonado, S.A.A., Maldonado, D.M.A., Quinde, C.P., Freire, E.X.G. and Vaca, G.K.A. (2017) Voxel: A 3D Rendering Proposal. *Proceedings of the* 2017 *Conference on Information and Communication Technology* (*CICT*), Gwalior, 3-5 November 2017, 1-5. http://ieeexplore.ieee.org/document/8340601/ https://doi.org/10.1109/INFOCOMTECH.2017.8340601

[13] Kaufman, A., Yagel, R. and Cohen, D. (1993) Volume Graphics. *Computer*, **26**, 51-64. http://ieeexplore.ieee.org/document/274942/ https://doi.org/10.1109/MC.1993.274942

[14] Zucker, S.W. and Hummel, R.A. (1981) A Three-Dimensional Edge Operator. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (*PAMI*), **3**, 324-331. http://ieeexplore.ieee.org/document/4767105/ https://doi.org/10.1109/TPAMI.1981.4767105

[15] Steuer, J. (1992) Defining Virtual Reality: Dimensions Determining Telepresence. *Journal of Communication*, **42**, 73-93. https://academic.oup.com/joc/article-abstract/42/4/73/4210117?redirectedFrom=fulltext https://doi.org/10.1111/j.1460-2466.1992.tb00812.x

[16] Schubert, T., Friedmann, F. and Regenbrecht, H. (2001) The Experience of Presence: Factor Analytic Insights. *Presence*: *Teleoperators and Virtual Environments*, **10**, 266-281. http://www.mitpressjournals.org/doi/10.1162/105474601300343603 https://doi.org/10.1162/105474601300343603

[17] Witmer, B.G. and Singer, M.J. (1998) Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence*: *Teleoperators and Virtual Environments*, **7**, 225-240. http://www.mitpressjournals.org/doi/10.1162/105474698565686 https://doi.org/10.1162/105474698565686

[18] Kilteni, K., Groten, R. and Slater, M. (2012) The Sense of Embodiment in Virtual Reality. *Presence*: *Teleoperators and Virtual Environments*, **21**, 373-387. https://ieeexplore.ieee.org/document/6797786/ https://doi.org/10.1162/PRES_a_00124

[19] Glenberg, A.M. (1997) What Is Memory for. *Behavioral and Brain Sciences*, **20**, 1-19. https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/what-memory-is-for/46E0F728DAA67D54AA44D2C4634D9556

[20] Zahorik, P. and Jenison, R.L. (1998) Presence as Being-in-the-World. *Presence*: *Teleoperators and Virtual Environments*, **7**, 78-89.

http://www.mitpressjournals.org/doi/10.1162/105474698565541
https://doi.org/10.1162/105474698565541

[21] Bystrom, K.-E., Barfield, W. and Hendrix, C. (1999) A Conceptual Model of the Sense of Presence in Virtual Environments. *Presence*: *Teleoperators and Virtual Environments*, **8**, 241-244. https://doi.org/10.1162/105474699566107
http://www.mitpressjournals.org/doi/10.1162/105474699566107

[22] Minsky, M. (1980) Telepresence. *OMNI Magazine*, 44-52.
https://philpapers.org/rec/MINT

[23] Guadagno, R.E., Blascovich, J., Bailenson, J.N. and McCall, C. (2007) Virtual Humans and Persuasion: The Effects of Agency and Behavioral Realism. *Media Psychology*, **10**, 1-22.
https://www.tandfonline.com/doi/full/10.1080/15213260701300865?scroll=topf&gneedAccess=true

[24] Bovet, S., Debarba, H.G., Herbelin, B., Molla, E. and Boulic, R. (2018) The Critical Role of Self-Contact for Embodiment in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics*, **24**, 1428-1436.
http://ieeexplore.ieee.org/document/8283639/
https://doi.org/10.1109/TVCG.2018.2794658

[25] De Vignemont, F. (2011) Embodiment, Ownership and Disownership. *Consciousness and Cognition*, **20**, 82-93. https://doi.org/10.1016/j.concog.2010.09.004
https://www.sciencedirect.com/science/article/pii/S1053810010001704

[26] Kuster, C., Ranieri, N., Zimmer, H., Bazin, J.C., Sun, C., Popa, T. and Gross, M. (2012) Towards Next Generation 3D Telconferencing Systems. *Proceedings of the* 2012 3*DTV-Conference*: *The True Vision—Capture*, *Transmission and Display of* 3D *Video* (3*DTV-CON*), Zurich, 15-17 October 2012, 1-4.
https://ieeexplore.ieee.org/document/6365454/
https://doi.org/10.1109/3DTV.2012.6365454

[27] Regenbrecht, H. and Langlotz, T. (2015) Mutual Gaze Support in Videoconferencing Reviewed. *Communications of the Association for Information Systems*, **37**, 965-989. http://aisel.aisnet.org/cais/vol37/iss1/45/
https://doi.org/10.17705/1CAIS.03745

[28] Hauber, J., Regenbrecht, H., Billinghurst, M. and Cockburn, A. (2006) Spatiality in Videoconferencing: Trade-Offs between Efficiency and Social Presence. *Proceedings of the* 2006 20*th Anniversary Conference on Computer Supported Cooperative Work*, ACM, New York, 413-422. https://dl.acm.org/citation.cfm?id=1180937
https://doi.org/10.1145/1180875.1180937

[29] Petkova, V.I., Khoshnevis, M. and Ehrsson, H.H. (2011) The Perspective Matters! Multisensory Integration in Egocentric Reference Frames Determines Full-Body Ownership. *Frontiers in Psychology*, **2**, 35-41.
http://www.ncbi.nlm.nih.gov/pubmed/21687436
https://doi.org/10.3389/fpsyg.2011.00035

[30] Regenbrecht, H.T., Park, N.J.-W., Ott, C., Mills, S., Cook, M. and Langlotz, T. (2019) Preaching Voxels: An Alternative Approach to Mixed Reality. *Frontiers in ICT*, **6**, 7. https://doi.org/10.3389/fict.2019.00007
https://www.frontiersin.org/articles/10.3389/fict.2019.00007/full

[31] Mori, M., MacDorman, K.F. and Kageki, N. (2012) The Uncanny Valley. *IEEE Robotics and Automation Magazine*, **19**, 98-100.
https://ieeexplore.ieee.org/document/6213238/
https://doi.org/10.1109/MRA.2012.2192811

[32] 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**, 379-393. http://www.mitpressjournals.org/doi/10.1162/105474605774785235 https://doi.org/10.1162/105474605774785235

[33] Llobera, J., Sanchez-Vives, M.V. and Slater, M. (2013) The Relationship between Virtual Body Ownership and Temperature Sensitivity. *Journal of the Royal Society Interface*, **10**, 1-11. http://rsif.royalsocietypublishing.org/content/10/85/20130300 https://doi.org/10.1098/rsif.2013.0300

[34] 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**, 203-220. https://www.tandfonline.com/doi/abs/10.1207/s15327108ijap0303_3 https://doi.org/10.1207/s15327108ijap0303_3

[35] Alghamdi, M., Regenbrecht, H., Hoermann, S., Langlotz, T. and Aldridge, C. (2016) Social Presence and Mode of Videocommunication in a Collaborative Virtual Environment. *Pacific Asia Conference on Information Systems* 2016 *Proceedings*, 126. https://aisel.aisnet.org/pacis2016/126

[36] MacDorman, K.F., Green, R.D., Ho, C.C. and Koch, C.T. (2009) Too Real for Comfort? Uncanny Responses to Computer Generated Faces. *Computers in Human Behavior*, **25**, 695-710. http://www.ncbi.nlm.nih.gov/pubmed/25506126 https://doi.org/10.1016/j.chb.2008.12.026