Voxel-based Immersive Mixed Reality: A Framework for Ad-hoc Immersive Storytelling

Stuart Duncan*, Noel Park, Claudia Ott, Tobias Langlotz, Holger Regenbrecht

University of Otago, Dunedin, New Zealand

*uohci@sjmd.dev

Department of Information Science, University of Otago

Dunedin 9016, New Zealand

Abstract

Volumetric video recordings of storytellers, when experienced in immersive virtual reality, can elicit a sense of co-presence between the user and the storyteller. Combining a volumetric storyteller with an appropriate virtual environment presents a compelling experience that can convey the story with a depth that is hard to achieve with traditional forms of media. Volumetric video production remains difficult, time-consuming, and expensive, often excluding cultural groups who would benefit most. The difficulty is partly due to ever-increasing levels of visual detail in computer graphics, and resulting hardware and software requirements. A high level of detail is not a requirement for convincing immersive experiences, and by reducing the level of detail experiences can be produced and delivered using readily available, non-specialized equipment. By reducing computational requirements in this way, storytelling scenes can be created ad-hoc and experienced immediately—this is was we are addressing with our approach. We present our portable real-time volumetric capture system, and our framework for using it to produce immersive storytelling

Figure 1: Māori storyteller captured by our system at the physical marae (left); 3D reconstructed marae model (center); two remote users are meeting to experience playback of recorded storyteller a reconstructed marae virtually (right)

experiences. The real-time capability of the system, and the low data rates resulting from lower levels of visual detail, allow us to stream volumetric video in real-time to enrich experiences with embodiment (seeing oneself) and with co-presence (seeing others). Our system has supported collaborative research with Māori partners with the aim of re-connecting the dispersed Māori population in Aotearoa New Zealand to their ancestral land through immersive storytelling. We present our system in the context of this collaborative work.

1 Introduction

Storytelling is a common form of entertainment, education, and communication (Young et al., 2023), but it is also a natural way to remember history and maintain cultural identity. In stories of cultural significance the identity and character of the storyteller, the place where the story is told, and any artifacts used to support the story can all be extremely important. Such stories are often social experiences, where a group of people forms an audience to listen to the story and engage together with the storyteller. An example of this kind of storytelling is in Māori culture, where it is used to pass down mātauranga and tikanga (knowledge, protocols, and practices). Māori are the indigenous people of Aotearoa (New Zealand), whose arrival from Polynesia predates European settlers by some 500 years. Traditionally Māori groups are centered on a marae (a complex

of buildings at the center of a Māori community) and remain connected to their ancestral land and their community by close physical proximity and face-to-face interaction—including storytelling. Since European arrival, Māori have become increasingly dispersed and separated from their communities and their ancestral land, and increasingly disconnected from their culture. Re-connecting this diaspora with their geographical and cultural origins has been a topic of growing interest over the last decade. A common approach is to record and distribute stories using modern digital media such as audio and video recordings which are easy to capture, distribute, and experience. These forms of media can capture some of the characteristics of the storyteller, and the physical place where the story is told or where it takes place. However, a degree of separation between the physical space surrounding the audience and that of the storyteller persists. As a result audio and video may fail to invoke the Māori concept of tūrangawaewae—which is derived from tūranga (standing place) and waewae (feet) and often translated as 'a place to stand' or 'a place of belonging'—in the audience.

Immersive virtual reality (VR) systems have the capacity to elicit a sense of *presence*, or the feeling that the user is in a place other than their physical location. If that virtual place is a faithful representation of a real (but perhaps distant) Māori ancestral place, then the sense of presence can perhaps confer some degree of tūrangawaewae. Combining this sense of presence with audio-visual representations of a storyteller can allow the user to experience a sense of *co-presence* (cf Guven et al., 2009) with the storyteller, deepen their connection to the place, and introduce a connection to the community via the storyteller. Including user self-representation in the environment, where the user identifies with their representation, introduces an element of *embodiment* which supports and strengthens the illusion of presence. Allowing remote users to see and interact with one another via their self-representation introduces an element of *tele-co-presence* or *social co-presence*, where the users feel as if they are together Hauber et al., 2005, and further enhances the experience. An experience that combines a sense of presence, co-presence with a

storyteller, embodiment, and social co-presence between multiple users would represent a step closer to the natural form of face-to-face storytelling.

Within this cultural and technological context, we have developed a system for storytelling in immersive VR which supports embodiment, presence, co-presence, and social co-presence (the four experiential factors) and where the virtual environments and the storytellers represent real places and people. The experience is based on an audio-visual model of a real place (the virtual environment), and we use volumetric video recordings and live streams to represent the storyteller and the users respectively. In our main example, we have used an accurate model of the wharenui (see Figure 1), which is the Māori meeting house and is a significant place on the marae. Together with our Māori partners, we have worked to ensure that the places and people are represented appropriately, which has required being able to allow the relevant members of the community to experience the stories directly. Ad-hoc volumetric capture and immediate rendering or playback have played a critical role in that aspect of the work, and have paved the way to an anticipated handover of our system to our Māori partners. The goal of this handover is to allow them to produce and distribute storytelling experiences without our input and in a practical and economical way. Because of the specific requirements of Māori culture, we opted to co-design a tailored system instead of modifying an existing solution in particular in the "social VR" space. Platforms such as VRTogether[1] or projects like VRComm (Gunkel et al., 2021) would have lent themselves as starting points. While providing an elegant solution to host virtual meetings with photo-realistic participant representations (by way of encoding depth information into grayscale images) they do not allow volumetric video recordings and streams to be easily embedded to support the cultural goals of our application. Also, a part of the overarching Ātea project, which is not described here, is researching and developing mechanisms to ensure and maintain data, infrastructure, and technological sovereignty for Māori. This, in combination with needing to satisfy specific Māori cultural

[1] https://vrtogether.eu

requirements, has motivated us to develop a bespoke system rather than using an existing platform or infrastructure [2]. Our approach has allowed us to ensure that the technology and the media content, and as much of the infrastructure as possible, can be owned and controlled entirely by our Māori partners once our system is easily used by laypeople.

We present first our conceptual approach to storytelling, and what is required to present stories in a way which elicits the experiential factors in such a way as to connect Māori to their distant ancestral marae (Section 2). Our system relies on an accessible and portable method for producing a volumetric video which we use to record the storyteller on-site so that they may naturally refer to the carvings and other artwork which decorate the interior of the wharenui. We use the same system to stream volumetric video in real time to represent users to themselves, and to one another. We describe the implementation of the volumetric capture system in section 3, and then the platform by which we capture an audio-visual model of the environment and deliver the complete virtual storytelling experience in section 4. Finally, we describe how the work was received by the Māori community, what limitations were identified, and potential ways to address these limitations in the future in section 5.

While we have conducted formal user studies based on this system (e.g. N. Park et al., 2022; Regenbrecht et al., 2021), that is not the focus of this paper. Our work is directed towards developing a novel system to suit the specific needs of our Māori partners, and our publication here is based on our belief that this approach can nevertheless produce useful research in other areas and is therefore of interest to the broader research area of virtual and mixed reality. We aim to describe our system here in sufficient detail for others to replicate it, and we make the project available publicly[3].

[2] for example www.tno.nl/en/digital/digital-innovations/digital-infrastructures/smart-society/

[3] Source code and binaries can be downloaded from https://sjmd.dev/vimr

2 Approach

Our conception of immersive storytelling consists of a storyteller delivering a story to one or more members of an audience (users) in a virtual environment. The storyteller and the users are embodied, and the users are present in the environment and co-present with one another and with the storyteller. Users are free to explore and examine the environment, and to interact directly with one another, with elements of the environment, and perhaps also with the storyteller. Delivering a convincing VR experience with these components depends on four experiential factors:

- A sense of presence—users feel present in the virtual environment, rather than their physical location
- A sense of co-presence with the storyteller—users feel as if the storyteller is there with them
- A sense of embodiment—the user has a virtual representation in the environment and has a sense of ownership, agency, and self-location with respect to that representation
- A mutual sense of social co-presence between users—users feel as if they are co-present with one another, and have a sense of shared experience

To consider the requirements for supporting these experiential factors, we decompose the experience into several components: the storyteller, the virtual environment, virtual representations of the audience members, and methods for the audience to explore and interact with the experience and with one another. While the storyteller, the environment, and the audience representation can in practice be partly or even purely synthetic, in our case it is important that each of these are accurate representations of their real analogs. The remainder of this section describes these components and how they should be experienced by users in order to elicit the requisite experiential factors. Later, in sections 3

and 4, we describe the implementation details of the system which we have developed to satisfy these requirements.

2.1 The Storyteller

Since storytellers are not always accessible—partly due to the dispersion previously mentioned—stories have been recorded and distributed using conventional forms of media, often audio or video recordings. Audio recordings capture stories in their oral form, are easily distributed and experienced, and capture some of the characteristics of the storyteller albeit only those expressed by their voice. Video recordings have similar practical advantages to audio and capture an additional set of visual characteristics, including facial expressions, gestures, and body poses. Video may also capture the surroundings of the storyteller and/or scenes which support the narrative of the story. These forms of media generally do not immerse the audience in the environment with the storyteller and rely entirely on the imagination of the audience for any sense of presence or co-presence. As a result, they may not invoke a connection to the places and people relevant to the story.
Volumetric video (or free-viewpoint video) is the extension of video to three dimensions, where the viewer is free to move around and explore the volumetric video and surrounding environment. When experienced using an immersive display system, appropriately scaled volumetric video of a storyteller can elicit in the user a sense of co-presence, where the user feels like they are sharing the experience with the storyteller. Volumetric video is accompanied by one or more audio recordings, which can be rendered as emanating from a particular position within the recording; The voice of the storyteller should emanate from the storyteller's recorded mouth, for example. Spatial rendering of positional audio should include spatial effects such as direction, attenuation over distance, and perhaps even a sound propagation and reverberation model of the virtual environment.
Production of volumetric video, however, remains expensive and time-consuming. Young et al., 2023 state factors related to hardware (cameras and other sensors, studio equipment)

and labor in post-production (experts using dedicated software tools). State-of-the-art systems such as VoluCap Schreer, Feldmann, Ebner, et al., 2019 use a studio consisting of 32 cameras surrounded by controlled lighting and background screens, and can capture an active action area of 3 meters in diameter. For each minute of volumetric video, VoluCap produces terabytes of data and requires 12 hours of post-processing time by trained operators. The cost, data volume, and time are partly due to the goal of producing photo-realistic reconstructions, which is a goal shared by most other volumetric capture systems, such as Microsoft's Mixed Reality Capture Studios, 8i, or Vologrames. All of those systems are capable of producing very high-quality volumetric videos but are everything but inexpensive, real-time, and ad-hoc in nature. Research has shown, however, that co-presence does not necessarily rely on very high levels of visual detail (e.g., Regenbrecht et al., 2017). It is therefore possible to relax the hardware and post-processing requirements by aiming for lower levels of detail.

Performance capture, on the other hand, has aimed to reduce the required number of cameras while still aiming for photo-realism by inferring or interpolating the missing data. Common approaches rely on a parametric model of the human body, where the general shape and structure are predefined and the RGBD data is used to adjust the model parameters and apply color to the model Loper et al., 2015. More recently, machine learning approaches have been used to infer missing parts of the reconstruction from a learned model of the data. In both cases, the complexity of the representation is limited by the features which are implicitly included in the learned or parametric model, and these typically do not include modeling skin wrinkles or creases in clothing which are included in directly captured data even if some detail is lost. In addition to this, generating missing details while aiming for photo-realism can sometimes produce uncanny results and the context of our work introduces cultural and philosophical aversions to partial or complete synthesis of models which represent real people. In our collaborative work, we can safely assume that audio and video recordings of people are often acceptable in Māori culture

since these formats are already widely used. However, parametric and generative models constitute a conceptually different approach where it is not clear that the visual representation is authentic, or that this is an acceptable way to represent real people. In Māori storytelling, storytellers are often kaumātua (respected elders), and misrepresenting them would not be appropriate (we discuss the appropriateness of storyteller representation by our system later in section 5). Conversely, incomplete recordings are acceptable so long as the data which is captured is accurate. Therefore, our approach is based on ensuring that the volumetric video is accurate and authentic and that it can be produced relatively easily, even if we have to sacrifice some visual detail, and if we must accept that the representation may be missing parts.

In addition to the oral component of the story, the storyteller may also refer to carvings and other artwork which support their stories. Our goal is to capture as much of the storyteller's natural references to these supporting artifacts, which includes gestures like pointing, and their gaze direction. Many of the artifacts are integrated into the interior structure of the wharenui, and so we should ideally capture volumetric video of the storyteller in situ so that they can naturally refer to them—that is, inside the wharenui rather than in a studio or in our research lab. Ultimately, we aim to have a system that can be operated without special training, in order to allow our Māori partners to freely produce and experience their own volumetric video and immersive experiences. Our system in its current state would already allow for that, but more documentation and general introductory training sessions are required before we can achieve this goal.

2.2 The Environment

Tūrangawaewae is important, particularly for dispersed Māori, and is related to the sense of presence in VR research. Our aim was to invoke the concept of tūrangawaewae with respect to a user's distant marae, by using the sense of presence to virtually transport the users to the marae.

Presence relies on appropriate interaction between the user and the scene (Section 2.4), the user's embodied relationship with the environment (Section 2.3), and the realism of the environment Schubert et al., 2001. Realism refers to how realistic the environment appears to the user, which includes the visual appearance, the audio characteristics, and the ambient noise of a physical place. The interior of the wharenui is decorated with intricate carvings and other artwork which should be captured with as much visual detail as possible since the details relate to the stories being told. Since the structure is static, we can use a high-fidelity static mesh model and use conventional methods to optimize the texture size and geometric complexity while preserving near-photo-realism if possible.

Aural characteristics of the experience are often neglected, however, intrusion of sounds from the user's physical location and the absence of appropriate ambient sound in the virtual environment weaken the sense of presence (sometimes even to the point of breaking the illusion Garau et al., 2008). To avoid this we must include the audio characteristics of the wharenui, and we should also model the reverberation characteristics of the physical space in order to accurately render audio such as the storyteller's voice.

2.3 User Representation

Users of our system should have embodied experiences—they should be able to look down (or in a virtual mirror) and see a virtual representation of their own body. This is normally only possible with augmented reality systems (cf Young et al., 2023). Additionally, they should be able to see and hear remote users who are sharing the immersive experience with them. Representing users to themselves and to one another has similar requirements to those of representing the storyteller, and so we choose to use the same volumetric video representation for all of the people in our system. The users' self-representation, the users' representation of one another, and the storyteller.

Embodiment has specific requirements over co-presence (either with recorded people or remote users). In embodied experiences the user is represented to themselves in the virtual

environment where embodiment can be separated into factors of ownership, self-location, and agency Kilteni et al., 2012. Self-location requires that the virtual body is seen from a first-person perspective, and agency requires that the virtual body responds appropriately to the movements of the user's physical body. To support a sense of agency, the movements of the virtual body should mirror those of the physical body with minimal delay and maximum accuracy. Ownership refers to the user identifying with the virtual body as their own. This is a composite factor that emerges when self-location and agency are sufficient and is strengthened when the virtual representation matches the user's physical appearance Waltemate et al., 2018.

Many existing systems use a rigged mesh model to represent the user and rely on a motion capture system to track the user's movements and animate the model Waltemate et al., 2018. Rigged mesh models can be generic humanoid models which are sometimes personalized manually or automatically by capturing images of the user Waltemate et al., 2018. In both cases, animating the model requires a bespoke motion capture system which normally uses reflective markers to track the user's joints to infer their body pose. Similar systems are used for facial tracking (like Weta Digital's FACET system). Motion capture systems are often expensive and cumbersome to transport. Furthermore, personalizing mesh models either requires a separate system to capture their likeness, or a time-consuming manual personalization process, or an auto-personalization process that often has the same authenticity concerns as performance capture systems. In particular, we aim to avoid any per-user setup steps (e.g. constructing or personalizing a model of the user, or affixing reflective markers to support motion capture), since this would limit the number of people that we are able to engage with—particularly outside of a laboratory context. Hence, approaches which require a setup or personalization step and rely on a motion capture system are not usable in our context where affordability and immediacy are primary overarching requirements.

Embodiment, like co-presence, does not depend strongly on the visual level of detail, and

rather on correct self-location and agency Kilteni et al., 2012. We can therefore accept a lower level of detail, and instead aim to support self-location and agency with a system that has minimal setup requirements per user and minimal additional hardware. Our alternative which satisfies this requirement is to re-use the volumetric capture system to produce a live volumetric video of the user which is rendered in real-time. Volumetric video can be precisely aligned with the VR system so that the user sees their volumetric representation from a first-person perspective. So long as the delay introduced by the volumetric capture system is minimal, the user will experience a sense of agency since the volumetric video directly captures their physical body in real-time. This does not require any per-user setup or personalization; users can simply use the HMD and engage with an embodied experience using a personalized virtual representation of their real body. In section 5 we discuss how this has been critical in engaging with a large and diverse audience. It has allowed us to share our work with the broader Māori community, which would not be possible if they had to travel to our lab, or if we had to personalize virtual bodies for all of the users.

2.4 Interactivity & Navigation

The sense of presence in a virtual place relies on the ability of the user to navigate the scene in a natural way, and in immersive VR this is supported by using a tracked HMD. To look around the user moves their head, and their view is updated appropriately. To move around the user can simply walk around their physical space, within the limits of the HMD tracking system. Virtual spaces, however, are often larger than the limits of the VR tracking system or the physical space available to the user. In these cases, the users should be able to explore the environment without moving around an equally sized physical space. VR systems commonly use a teleportation metaphor where the user retains their natural navigation (and thereby a sense of presence) for their immediate surroundings, but can also make large 'teleport' movements around the virtual space without physically moving. Users may also control the playback of the story to, for example, pause playback to more

closely examine relevant parts of the environment and resume playback once satisfied. Where multiple users are co-present with one another, the state of the environment should be consistent for all users; if one user pauses the story, then it should be paused for all other users sharing the experience. Likewise, the teleportation of users, and any other environment state, must be synchronized to all participating users so that they have a consistent shared experience.

Interaction between users requires that they can see and hear each other, so their virtual representations should be visible to one another in the shared space. With current volumetric video systems, this is typically only possible with co-located users using an AR system to support the experience (e.g. MS HoloLens with Vologrames). Sharing the dynamic parts of the experience, such as the users' voice, their physical movements, and their teleported position, as well as the playback state of the recorded story, is essential to achieve a coherent sense of social co-presence between users.

2.5 Summary

Combining the experiential components described above creates social storytelling experiences as illustrated in figure 2. The storyteller is represented by a recorded volumetric video, and users are represented by similar live volumetric video streams. Users can see one another and can speak to one another naturally with audio that is rendered spatially as emanating from their position in the environment as they navigate by walking and by teleportation. The recorded story is also shared and the playback state is synchronized between all users, and any of them can control the shared story playback state.

To support this experience we have designed a volumetric video capture and streaming system based on the above requirements, while also prioritizing ad-hoc capture and portability so that the system can be used in situ to both produce and experience volumetric story recordings and experiences. We combine this with a framework for

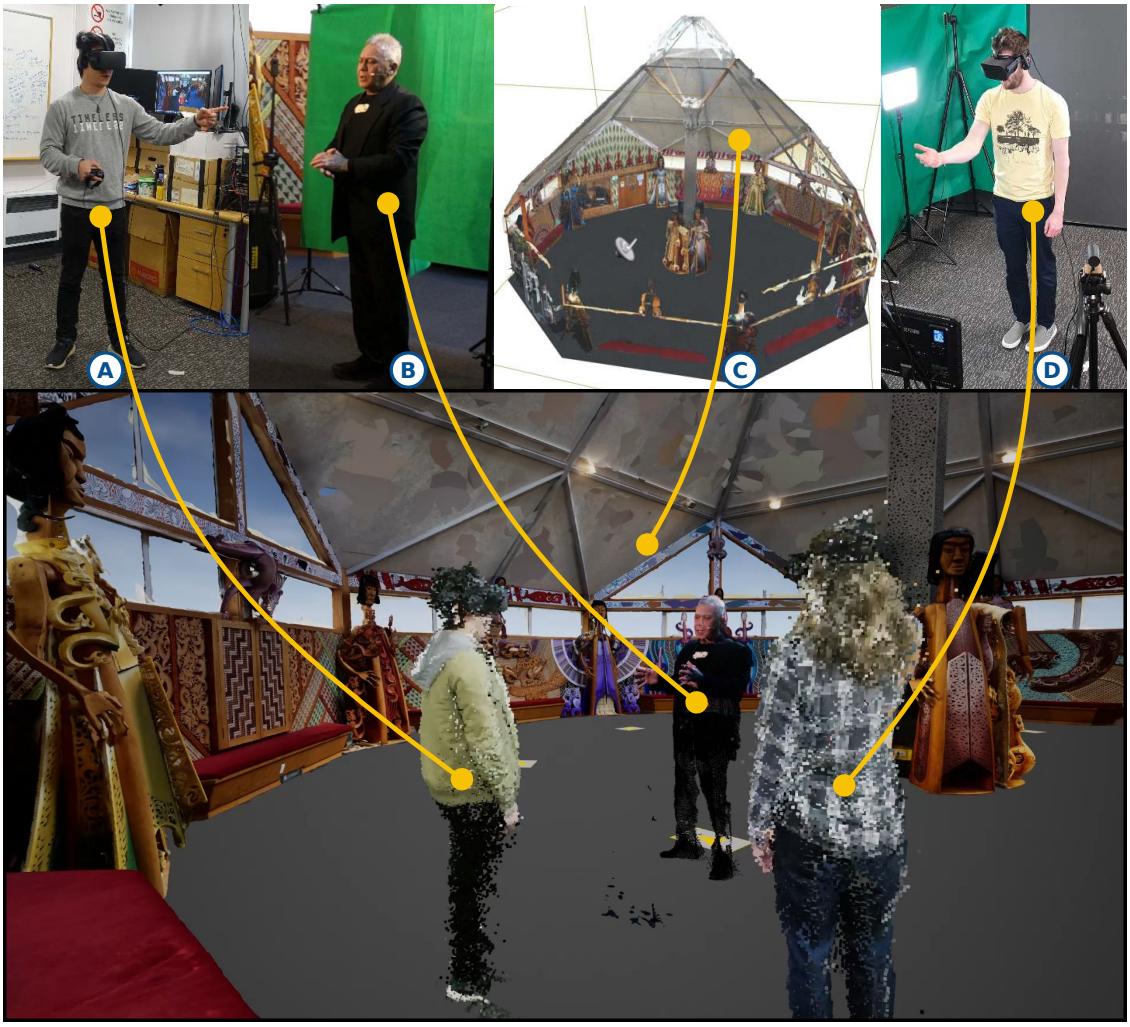

Figure 2: Immersive storytelling with local and remote users (A and D), where the shared experience includes a recording of the storyteller (B) and an audio-visual model of the environment (C). Users are represented by live volumetric video streams and can see and hear each other. The storyteller is represented by a volumetric video recording, accompanied by an audio recording of their voice. Voice audio, including the users' voices, is rendered spatially as emanating from the mouth of the person speaking.

producing immersive storytelling experiences based on real people and places, and for integrating live and pre-recorded volumetric videos with models of real environments to produce an immersive, social, storytelling experience. We describe in detail the volumetric video capture system (Section 3) which we use to record storytellers and represent users in the virtual environment, and the integrated platform we use to deliver the experience (Section 4). Finally, we discuss how our work was received by our Māori partners, outline relevant technical limitations, and propose future work to address them (Section 5).

3 Volumetric Video Capture System

We have developed a system for capturing relatively low fidelity volumetric video in real-time, using readily available RGBD (RGB color and depth) cameras, generic processing hardware, and a consumer-grade VR system (an Oculus[4] CV1). Our system—voxel-based immersive mixed reality (VIMR)—uses an octree structure to encode voxel occupancy grids that represent voxelvideo frames (our volumetric video encoding). A corresponding serial octree encoding enables recording, playback, and live streaming of voxelvideos over network connections. We use instances of this system ranging from a single computer with a single camera, to multiple multi-camera instances networked together for tele-co-presence. In this section, we describe the technical implementation details of VIMR, along with some performance metrics such as voxelvideo data rates and reconstruction latency. We first describe the distributed and loosely coupled architecture of the system (Section 3.1), and the in-memory and serial-encoded octree representation of voxelvideo frames (Section 3.2). In section 3.3 we describe the camera registration method that we use to align data from multiple cameras to a single coherent coordinate system and to register the voxelvideos in the VR coordinate system. Finally, we describe how we stream voxelvideos over the network (Section 3.4), and how the streams are recorded with accompanying audio for later playback (Section 3.5).

[4]https://oculus.com

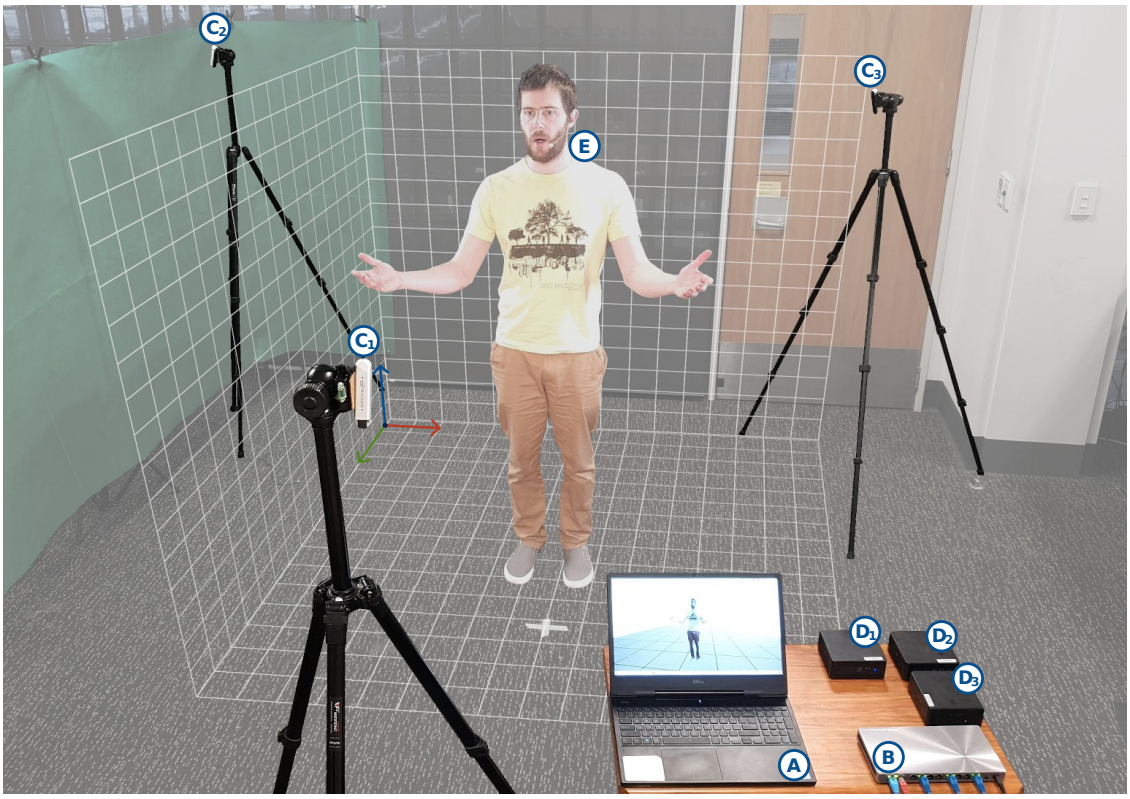

Figure 3: An example of the portable capture system in our lab, showing a laptop (A) rendering live volumetric video captured using three RealSense cameras (C_1-C_3). Each camera is connected to a mini NUC computer (D_1-D_3), and all of the computers are connected via a dedicated network switch (B). A voice recording is captured using a wireless headset microphone (E).

3.1 System Architecture

Figure 3 shows the physical components of a typical three-camera instance of VIMR which we use to produce voxelvideos. The instance depicted consists of a laptop computer for recording and rendering volumetric video, and three RealSense D415 cameras for capturing the point cloud data. Each camera is connected to a NUC[5], and the NUCs are connected over Ethernet to the laptop via the switch. The network connection allows the reconstruction software components running on each NUC to communicate with the rendering & recording components running on the laptop. Reconstruction components acquire data from cameras, convert it to one octree-encoded voxel grid per RGBD frame,

[5]https://intel.com/content/www/us/en/products/details/nuc.html

and then to a serial-encoded octree. Serial data is then sent over UDP[6] to the laptop where data from all three cameras is combined to a single octree for rendering. The combined octree is also re-serialized for recording, and so that it can be sent to a remote party as a representation of the local user to support tele-co-presence.

Peer-to-peer (P2P) Network connections between components are negotiated by first connecting to a central server (VNet), and repeatedly broadcasting a pairing request message to all other components connected to the server. The pairing request message is defined by the canonical ID of the component, and the ID of it's desired peer. When a component receives a pairing response message, request broadcast is stopped and the components can begin transceiving data. VNet operates as a STUN[7] server, allowing VIMR components to connect to one another when running behind some firewall configurations which block incoming connections. This allows negotiation of tele-co-presence connections over institutional internet connections where firewalls such as this are pervasive.

In addition to reconstruction and rendering & recording components, VIMR also has a control component that exposes runtime controls such as the voxel size, desired frame rate, and toggles for starting and ending voxelvideo recordings and enabling or disabling network streams. The control component is accessible as a web app served over the loop-back network device.

Internally, each software component implements a triple-buffered asynchronous data processing pipeline that processes data and converts data representation (e.g. between RGBD point cloud and octree, between octree and serial octree, etc). The triple buffering arrangement allows stages of the pipeline to operate concurrently in order to minimize latency. Processing stages are configured to discard data when a downstream stage is slower, in order to prioritize latency over lossless processing.

This arrangement has proved reliable, flexible, and allowed us to prepare and calibrate on-site volumetric capture systems for recording and for embodied experiences with about

[6] a networking protocol without error checking or corrections

[7] a gateway networking protocol for real-time video etc., applications.

two hours of total setup time.

3.2 Voxelvideo Frames

Voxels—cubes or rectangular volumes arranged in a gapless 3D grid—are the three-dimensional extension of pixels which are (conceptually) a gapless 2D grid of square or rectangular areas. Voxel grids have been used for spatial compression Kammerl et al., 2012, and to improve performance Lindlbauer and Wilson, 2018, however, the result is almost always converted to a textured mesh before rendering. In VIMR, we use the voxel grid as the native data representation as well as the visual primitive (see section 4.2). While there are existing systems to efficiently represent a voxel grid (e.g. GVDB Hoetzlein, 2016), we have used a bespoke pointer-based octree structure for in-memory octrees, and a bitfield-occupancy encoding for serializing the octree. This way we avoid requiring GPU hardware to manipulate the in-memory tree (as would be required for GVDB), and we avoid the complexity of using a more sophisticated encoding (e.g. MPEG-based containers). In any case, we eschew photo-realism and use a bitfield-encoded octree which is sufficiently compact for streaming and recording voxelvideos.

Our choice of voxels for visual representation is to present the data in as close to its native form as practical. The native form of RGBD cameras is a depth image, or '2.5D' representation, however, the data is almost always used to construct a 3D point cloud. This conversion preserves the discrete sampled nature of the data, but many systems process the data further to construct a textured mesh, modifying it to have a smooth appearance. This process requires inferring missing parts of the mesh in order to produce a watertight model. Generation or interpolation of data is conceptually different from audio or video recordings, where the user experiences data in the same form that it is captured in without using a model to infer missing parts.

On the other hand, research has demonstrated the efficacy of point clouds Gamelin et al., 2021 and voxel grids Regenbrecht et al., 2017; Regenbrecht et al., 2019 in supporting

embodiment and co-presence in VR.

Voxel grids, in addition to presenting performance advantages, encode a consistent level of detail over the whole grid while the spatial density of point clouds captured by RGBD cameras varies with distance from the camera. In VIMR we use a voxel grid for its performance advantages, for representing a consistent level of detail regardless of the distance to the camera, and to avoid generating or inferring data that is not directly captured by the camera. Nevertheless, we aim to preserve the contiguity of surfaces imaged by the cameras, and so we choose a 'gapless' voxel size to match the lowest expected spatial density of the point cloud. With this voxel size, the contiguous surfaces in reality correspond to contiguous 'shells' of occupied voxels in the grid.

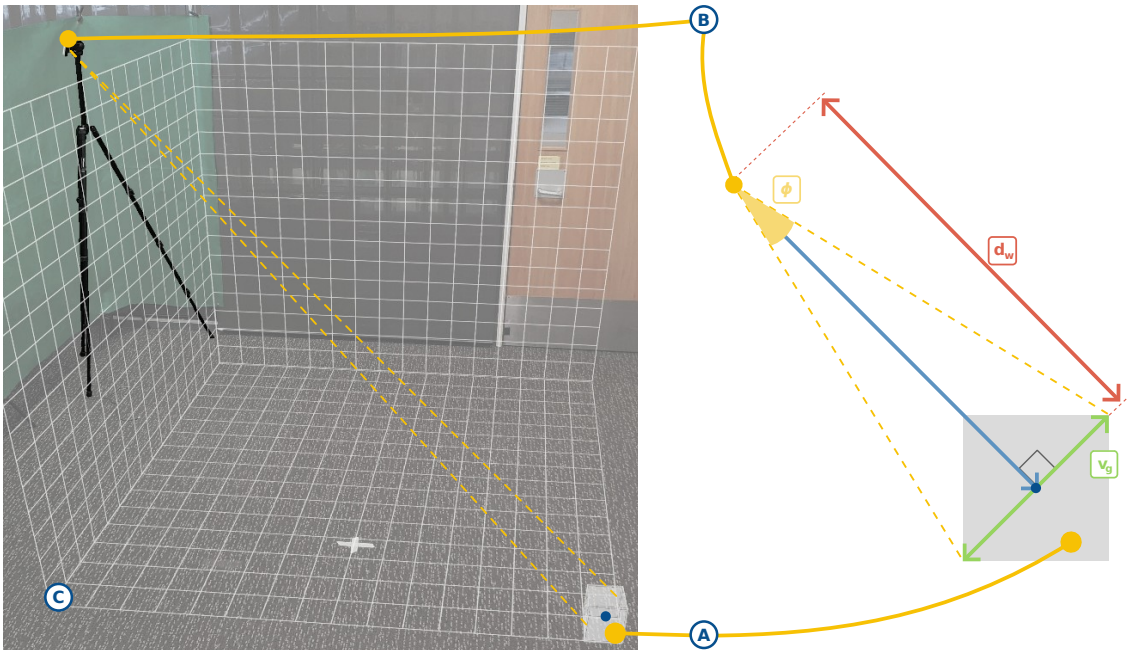

Figure 4: Physical measurements used to calculate v_g for the furthest voxel (A) from the camera (B) within the extents of the capture volume (C).

For a given camera, we compute the lower bound, v_g of the voxel size for which its point cloud will be gapless

$$v_g = 2\sqrt{2}\, d_w \tan\left(\frac{\phi}{2}\right) \tag{1}$$

where ϕ is the angular resolution of one depth pixel, and d_w is the maximum expected working distance from the camera. We assume the worst-case alignment of the voxel grid, where the camera's projection ray traverses diagonally through the voxels. d_w is determined by the pose of the camera relative to the capture volume as illustrated in figure 4. For a given gapless voxel size and cubic capture volume with edge length V we calculate the minimum octree depth, D, required for an octree to contain the whole volume

$$D = \left\lceil \log_2 \frac{V}{v_g} \right\rceil . \tag{2}$$

Octree leaf nodes correspond to occupied voxels, and each leaf stores the average color of the RGBD points contained in the voxel. In-memory octrees are converted to a serial encoding where internal nodes are represented by one-byte child-occupancy bitfields, and leaf nodes are represented by their color. The serial encoding consists of the breadth-first concatenation of the node bitfields, followed by a contiguous array of leaf data. A voxel frame consists of the serial octree accompanied by the voxel size, frame timestamp, frame sequence number, and a list of 3D poses which are used to dynamically position audio sources in voxelvideo recordings (see Figure 5).

Figures 3 and 4 show a typical VIMR instance: A cubic capture volume of $\approx 2.5 \times 2.5 \times 2.5\,\mathrm{m}^3$ surrounded by three RGBD cameras. The cameras are placed at radial intervals of $\approx 120\,^\circ$ in the horizontal plane and equal distances from the center of the volume. The cameras (RealSense D415[8]) have a depth pixel resolution of $\phi = 0.03935\,^\circ$, and the configuration has a maximum working distance of $d_w \approx 3\,m$, and so the minimum gapless voxel size is $v_g \approx 3\,mm$. To represent the whole volume requires an Octree depth of $D = 11$. Note that v_g is the lower bound and we may use a larger voxel size in order to reduce the voxelvideo stream data rate or file size. We typically use sizes of 4, 6, and 8 mm for those reasons.

[8] https://intelrealsense.com/depth-camera-d415/

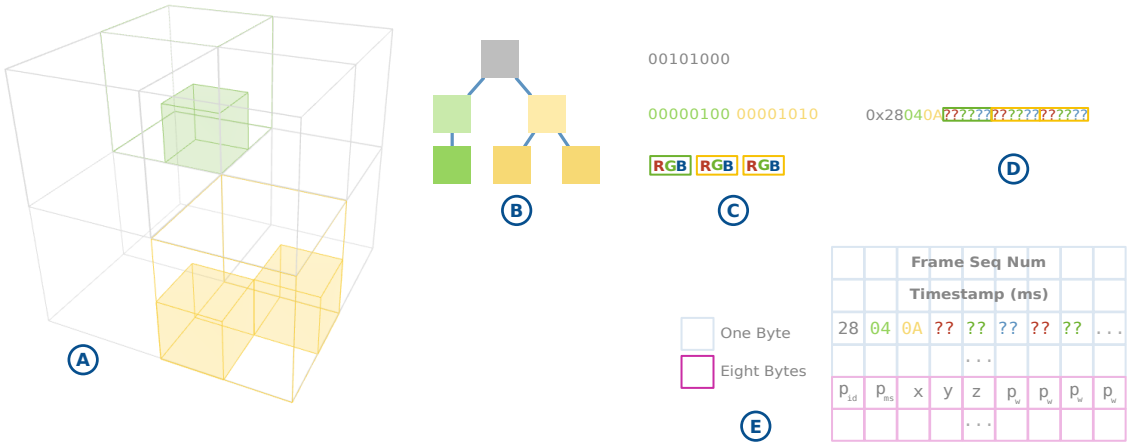

Figure 5: An example of the voxel serial encoding and the voxel frame structure. A voxel grid (A) with three occupied voxels showing a volume hierarchy that corresponds to an octree with three levels (B). Root and internal nodes are represented as child-occupancy bitfields, and leaf nodes store three color bytes in R-G-B order (C). A serial-encoded octree is then the breadth-first concatenation of the root/internal nodes followed by the leaf data (D). Voxel frames (E) consist of a serial octree accompanied by a frame timestamp and a frame sequence number, and an optional list of pose information.

3.3 Camera Registration

For multi-camera configurations, the cameras are registered to a common "world" coordinate space, and their RGBD point clouds are transformed so that they are aligned with one another in that space. The resulting aligned voxel grids can be combined to produce a coherent world voxel grid. Figure 6 illustrates the camera, VR, and world coordinate spaces, and the intermediate spaces used during calibration. Camera registration in VIMR is a rigid transform, $\mathbf{A}_i$, for each camera, i, and is composed of several transforms:

$$\mathbf{A}_i = S\mathbf{A}_{VR}\mathbf{A}_{i,pose}\mathbf{A}_H. \tag{3}$$

Coordinate system handedness is first converted from the right-handed camera coordinate system to a left-handed one used in computer graphics by $\mathbf{A}_H$. All 3D data except for the RGBD point cloud is assumed to be in left-handed space unless otherwise specified. Points are then aligned to the intermediate 'camera-world' space by $\mathbf{A}_{i,pose}$ and finally to world

space by $\mathbf{A}_{VR}$. Scale factor S is then applied to convert from the camera base units to voxels. The composite transform, $\mathbf{A}_i$, is then applied to all RGBD points to transform them from right-handed camera space with base units defined by the camera to a left-handed world-aligned space where the base unit is voxels. S is calculated from v and the base units of the camera, c_b,

$$S = \left(\frac{c_b}{v}\right) \tag{4}$$

We estimate $\mathbf{A}_{i,pose}$ by capturing a checkerboard simultaneously with all cameras, ensuring that it is visible to at least two cameras in each captured image. Camera poses are estimated from the checkerboard images, and bundle-adjustment is applied to optimize the camera pose estimates Triggs et al., 1999. For this approach, the first pose of the checkerboard defines the pose of the camera-world space.

To estimate $\mathbf{A}_{VR}$ we rigidly attach a handheld VR controller to the checkerboard, and first use hand-eye calibration Horaud and Dornaika, 1995 to estimate the, $\mathbf{A}_B$ transform between the controller and checkerboard. The controller and checkerboard are then simultaneously tracked by the VR system and the cameras respectively. A set of point pairs is recorded, where each pair consists of the controller pose in VR space, and the controller pose in camera-world space, estimated from the checkerboard pose, $\mathbf{A}_{i,V}$ and $\mathbf{A}_B$. We then use partial Procrustes analysis (a method that can be used to determine the optimal rotation for an object with respect to another) to estimate the rotation and translation between the camera-world and the world space.

$\mathbf{A}_H$ and S are defined by the cameras and included as part of the reconstruction component. $\mathbf{A}_{VR}$ and $\mathbf{A}_{i,pose}$ are loaded on initialization from files containing a translation vector and a quaternion-encoded orientation. If either $\mathbf{A}_{VR}$ or $\mathbf{A}_{i,pose}$ are missing, $\mathbf{I}$ (the identity transform) will be substituted instead. This allows the use of VIMR with or without a VR system and simplifies camera registration to just estimating $\mathbf{A}_{VR}$ if the system only uses one camera.

Transform, $\mathbf{A}_i$ is applied to RGBD points, p_j, for each frames to convert them to voxel grid indices

$$q_j = \mathbf{A}_i p_j. \tag{5}$$

All q_j are then inserted point-wise into the octree to construct the voxel frame. Since q_j for all cameras are all expressed in the same coordinate space, combining reconstructions from multiple cameras is implemented by simply merging the tree structures and averaging the colors of voxels present in more than one tree. Tree-merge requires that the residual error of data aligned by $\mathbf{A}_i$ is within half of a voxel width, and generally we can allow up to one voxel width before misalignment is obvious. If the residual alignment is too large, the voxel size can be increased to mask the misalignment. In practice, the calibration method produces alignments that are sufficiently accurate to produce reconstructions at 4 mm grid resolutions.

3.4 Voxelvideo Streams

Voxelvideo streams are sent over local network connections (LAN) in real-time for multi-camera reconstructions, to record storytellers and to support embodied experiences. Streams can also be sent over the internet to enable social co-presence between remote users (tele-co-presence). Streams are sent frame-by-frame, and each frame is sent as soon as its serial encoding is complete. Network message sizes are limited either by the UDP maximum datagram size of 65 kB (for LAN connections) or by the de facto standard Ethernet Maximum Transmission Unit (MTU) of 1.5 kB for internet connections. Serial frames are generally larger than either of these limits, and messages are divided into sequentially numbered fragments that fit within those limits. Fragments are sent in order, and if the receiver detects a missing or out-of-order fragment the message is discarded in order to maintain real-time frame rates.

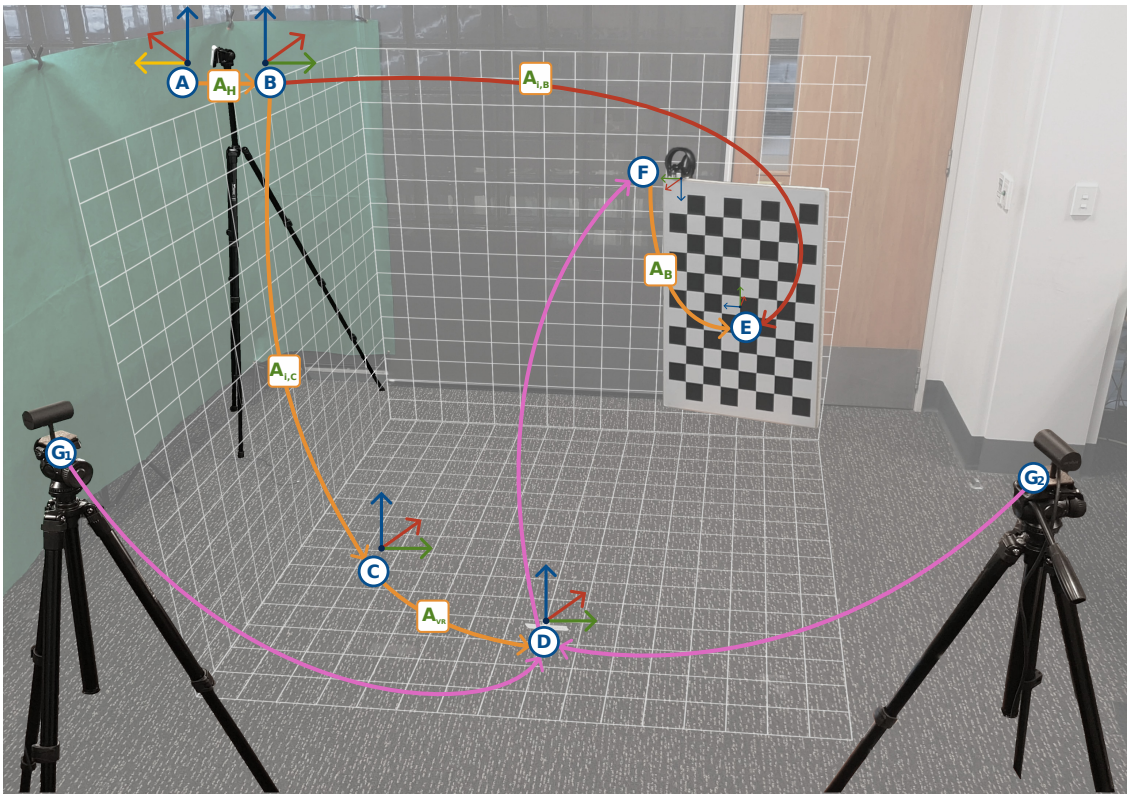

Figure 6: Coordinate spaces and transforms of VIMR illustrated for one of the three cameras in the example capture system. Spaces used for camera registration are the right-handed (A) and left-handed (B) camera coordinate systems, the camera-world coordinate system (C), the VR tracking system (D). The transform between the checkerboard (E) and the controller (F) is also estimated, where the controllers are tracked by the fixed cameras (G_1) and (G_2).

Bandwidth limits over the internet (and sometimes over LAN) require reducing the data rates. This is done by increasing the voxel size, to reduce the number of voxels per frame and thereby reduce the data size of each frame, or by reducing the target frame rate to allow more time to send data. Large messages can also produce a large number of fragments which together with slow internet connections causes a delay as we might also need to wait between fragments. This is to avoid overloading network buffers and causing data loss. Typical configurations for tele-co-presence are to use the maximum frame rate of 30 FPS, and choose a voxel size of $v = 4\,mm$ for LAN connections or $v = 8\,mm$ for internet connections. Figure 7 shows the levels of detail captured for these resolutions, as well as for 2 mm and 1 mm reconstructions which we aim to use in the future.

Cond	Res (mm)	Vox Count	Frame Size (kB)	Rate MB/s @ 30 FPS
Standing, 3 Cameras	8	25k	80	2.5
	4	75k	255	7
	2	180k	615	20
Sitting, 1 Camera	8	11k	40	1.5
	4	35k	121	3.5
	2	50k	220	6.5

Table 1: Data rates for common configurations. The work presented in this paper mostly refers to standing users captured by three cameras, however, we present for comparison also a single camera capturing a sitting person from the waist up.

3.5 Voxelvideo Recordings

Voxelvideos consist of a recorded voxelstream, accompanied by a human-readable metadata file and optionally one or more audio recordings (cf. figure 8). Metadata files include the title of the recording, duration, frame count, and recording time, as well as a list of audio files to load paired with pose identifiers to position the audio source appropriately (e.g. voice recordings are positioned to coincide with the recorded speaker's mouth). Poses are included with each voxel frame (see section 3.2), and where a frame includes a pose with a matching identifier the pose of the audio source is updated dynamically to match it.
Voice audio poses are estimated either by retrieving the 'head' joint from a Kinect 2 (if used) skeleton tracking system or if the capture system does not include a Kinect then the pose can be estimated in post-processing. Audio pose post-processing requires defining a bounding box for the head in the initial frame and then tracking the head for the rest of the frame sequence. The centroid of the head voxels in each frame defines the voice pose. Note, that neither method accurately estimates the orientation of the storyteller's head, and the direction of the voice is only approximate. Nevertheless, users can still experience spatial audio which differs depending on whether they stand in front of or behind the storyteller.
Audio recordings are captured during recording by the VIMR render & recording component which relies on PortAudio[9]. We use a headset microphone, as shown in figure 3, to record consistent audio even when the storyteller moves their head and looks around. If

[9]https://github.com/PortAudio/portaudio

Figure 7: An illustration of the level of visual detail captured at several different resolutions. 8mm (A) is a common resolution, dictated by the angular resolution of the Kinect depth sensor. 4mm (B) is a second common resolution, typically when using RealSense cameras, for embodiment, and where data is not streamed over the internet, 2mm (C) and 1mm (D) capture very high levels of detail, however, the high data rates (see Table 1) preclude their use in real-time streaming and rendering, and the voxel count limit of our current voxel renderer make capturing this level of detail impractical.

one would use a static microphone placed in the environment, like a shotgun mic, then the to-be-recorded storyteller would have to talk always in one direction to avoid changes in audio recording volume and quality. Synchronization between audio and voxel frame sequence is only approximate after recording, and we use a synchronization event to estimate the offset and ensure synchronous playback. Typically this is a single clap that is recorded before the sequence starts. The precise clap time from the start of both the audio and the voxel frame sequence is then manually identified, and the time offset between the voxel sequence and the audio is estimated. To finalize the voxelvideo we developed a voxelvideo processing utility which we use to insert audio poses, trim and synchronize voxel and audio recordings, and to remove the synchronization clap. While this post-processing

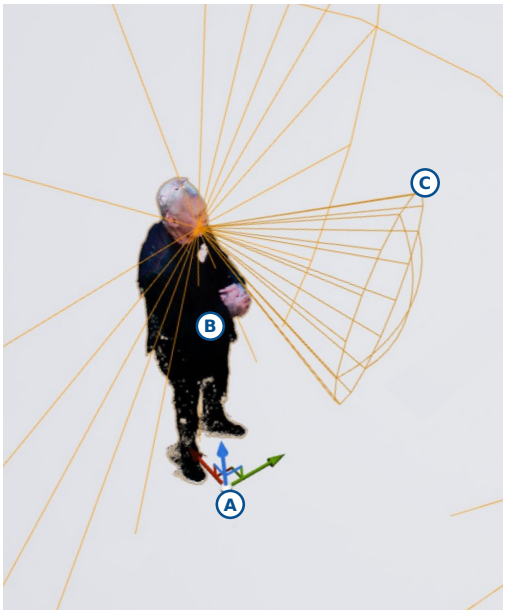

Figure 8: Voxelvideo actors consist of a voxel renderer with a world-space position (A), that determines the pose of the voxel frames (B) and audio source (C). This example (C) illustrates the cone-shaped angular falloff model used for directional audio sources in UE4.

step could be ignored, if one accepts a less accurate synchronization, it does improve the voxelvideo storytelling overall quality.

4 Delivering Story Experiences

We integrated VIMR with Unreal Engine 4[10] (UE4) to deliver an immersive experience. This offers the advantage of a highly optimized scene-graph implementation and comprehensive support for many immersive VR systems, spatial audio rendering, and an online subsystem for networked multi-user experiences. We developed a plugin to interface between VIMR and UE4 as the UE4 plugin system allows plugins to be integrated into arbitrary UE4 projects and simplifies portability between versions. As a result, anyone who is familiar with UE4 can produce immersive experiences and can use VIMR to include live and pre-recorded voxelvideo streams.

In this section we describe how we produce the audio-visual environment in UE4 (Section

[10]https://unrealengine.com

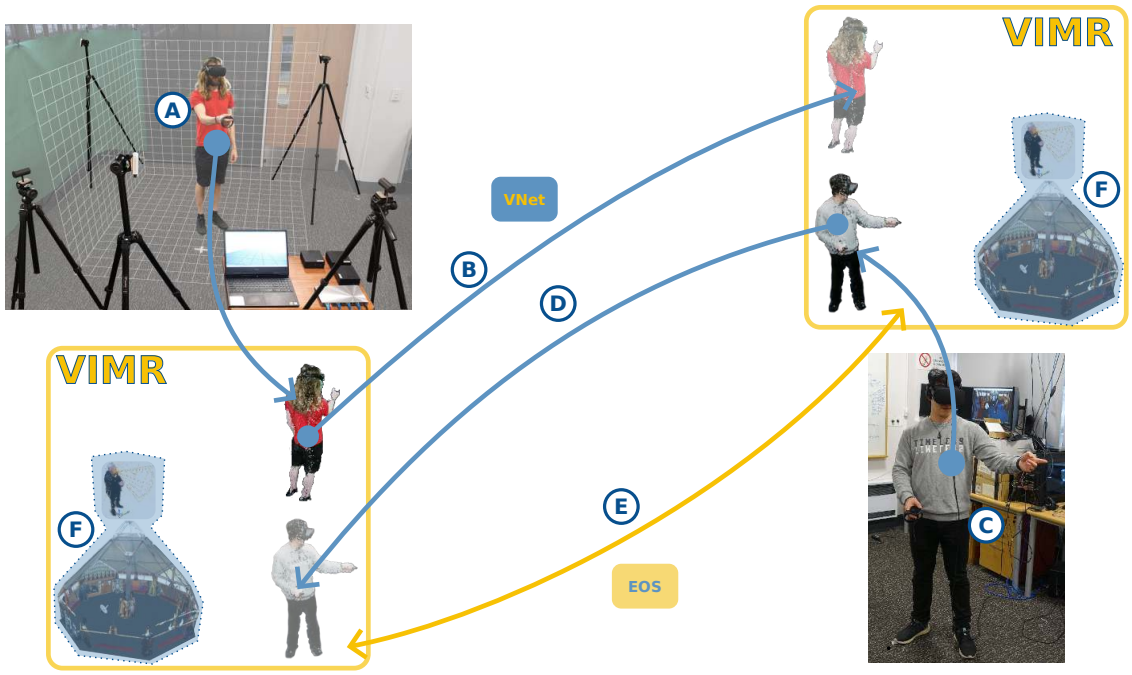

Figure 9: Two-user tele-co-presence with VIMR. Local (A) and remote (C) users have pre-distributed copies of the environment and voxelvideo storyteller (F). VNet is used to negotiate P2P connections between the instances to stream the local (B) and remote (D) users' live voxelvideo streams. Voice audio and environment state are synchronized between the two users' instances by a UE4 online subsystem connection (E), which is facilitated by the EOS server.

4.1), and then how voxelvideo rendering is implemented (Section 4.2)—namely how we work around the plugin's limited access to the rendering pipeline. We also describe how live voxelvideo is used to support embodied and tele-co-present experiences (Section 4.3). Figure 9 illustrates the components arranged for an immersive storytelling experience that is shared by two users in separate locations.

4.1 The Environment

The virtual environment, illustrated in Figure 10, consists of a 3D textured mesh with a matching audio model and an ambient audio source. We used RealityCapture[11] to produce a 3D model from photographs, using $\approx$ 3000 images to reconstruct the inside of the wharenui for our example. RealityCapture initially produces a dense mesh model with

[11]https://capturingreality.com/

millions of vertices, which is iteratively simplified while manually removing spurious reconstruction artifacts. Once the mesh is sufficiently simplified, to the order of 100 k triangles, RealityCapture is used to produce a photo-realistic texture from the photographs. The simplified mesh and matching texture are exported, and can then be imported into a UE4 project. To ensure a 1:1 scale in VR experiences, at least one measurement of an object in the reconstruction is required. The reconstruction scale can either be adjusted manually in UE4, or control points can be used to create a distance constraint in RealityCapture to correct the scale before exporting.

We use a mesh model to represent the environment, and in some cases other rigid and non-deformable props, because it allows us to capture a higher level of detail than if we used our voxelvideo representation. In our Māori storytelling example this allows us to capture the intricate visual detail of the carvings and artwork inside the wharenui.

The spatial audio model should match the audio characteristics of the real place, or if the environment is synthetic it should match the visual appearance of the space. In addition to correctly positioning the audio sources, reverberation and attenuation must also be modeled. Attenuation models are applied to each source and include parameters such as the falloff rate over distance and the attenuation projection model. Figure 8 illustrates the audio source positioning and the cone-shaped directional attenuation projection model which we use to render the voices of storytellers and of users. The audio submix is used to apply the audio model to a spatial volume as illustrated in figure 10. Audio characteristics can be modeled approximately using a reverberation effect, however for a more accurate model a loudspeaker and a microphone can be used to capture a room impulse response (RIR). To produce the RIR we used the Convolutional Reverb tool[12] to capture and process sweep responses and produce a wave-encoded RIR which can be used with UE4. The RIR is then applied via the submix graph.

Occlusions are not modeled by the RIR, and rather rely on physics collision modeling to

[12]Apple Impulse Response Utility: This has now become part of the Logic Pro studio suite by Apple.

Figure 10: The virtual environment consists of a textured mesh (A) which should be an accurate visual representation of the real place, an ambient audio source (B), and an audio post-processing volume (C) which applies an audio rendering model which should match the expected audio characteristics of the visual model.

attenuate audio when a collision mesh is between the user and the audio source. We approximate occlusion modeling of the environment by placing an invisible collision cylinder over the central pillar of the room. More sophisticated occlusion attenuation is possible with the use of third-party plugins for UE4.

Ambient audio was recorded at eight different positions distributed throughout the wharenui. Recordings were then combined to produce a monophonic looping audio track which is played back as an ambient audio source (where the spatial audio rendering is not used).

4.2 Rendering Voxels

Our voxel encoding is not natively supported by rendering hardware or software. Thus the rendering pipeline must be adapted to include support for voxel frames. However, the plugin framework provides only limited access to the rendering pipeline. For example, we cannot use geometry shaders to dynamically spawn voxel cubes. We work around those limitations by re-purposing three texture buffers and using a set of pre-generated mesh cubes; two RGB8 buffers store the coarse and fine voxel grid positions (t_{coarse} and t_{fine}), and the third stores the color. Position textures are used with the voxel size, v, to recover the position of the voxel relative to the origin of the voxel rendering actor.

$$p_w = v \times (256 \times t_{\text{coarse}} + t_{\text{fine}}) \,. \tag{6}$$

The texture buffers are updated for each reconstructed frame, and a vertex shader is used to recover p_w for each voxel and then shift, scale, and color each cube appropriately. Due to texture size limitations and the fixed number of pre-generated cubes, we dynamically adjust the number of renderer objects if the voxel count is higher than a 196k voxel limit for a single voxel renderer object. In practice, for gapless reconstructions, the voxel count is limited by the angular resolution of the depth sensor and the minimum camera working distance required for human reconstruction at full height. Typical voxel counts per frame set for systems with 1 to 4 cameras consist of 20k-80k voxels with voxel sizes between 8mm and 4mm.

Rudimentary voxel transforms (figure 11) are implemented by transforming the voxel positions by the world-space transform of the voxel rendering actor, $\mathbf{A}_a$, and then rounding the result to the nearest integer voxel size.

$$p_v = v \times \left\lfloor \frac{\mathbf{A}_a p_w}{v} \right\rceil \,. \tag{7}$$

Since this operation does not resample the transformed voxel grid, the result preserves

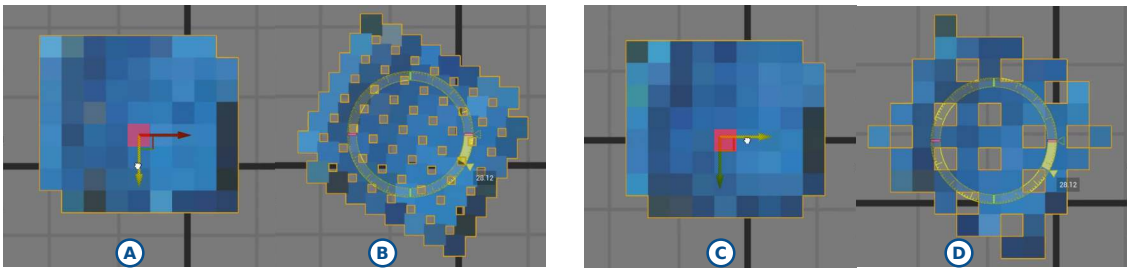

Figure 11: Rudimentary voxel translation (C) and rotation (D) operations are implemented by simply quantizing the results of the continuous translation (A) and rotation (B). The result is that voxel grid spacing is preserved. Because no resampling is done the rotation operation will reduce the voxel count for some rotation angles.

voxel alignment with the grid but may loose some data depending on the rotation angle of $\mathbf{A}_a$ when transformed voxel positions are rounded to the same grid indices.

4.3 Voxelstream Visualization

In contrast to other volumetric video recording and playback approaches, we follow a strictly real-time approach, i.e. we avoid post-production altogether and we are minimizing latency during capture and experience. This real-time approach allows us to (1) playback voxelvideos in a simple-to-integrate and operate way, (2) provide embodied experiences for our users by visualizing their own bodies within the virtual environment, and (3) optionally offer a social tele-co-presence experience for non co-located (remote) users.

Voxelvideo Playback: Voxel frame sequences and their associated audio components should be played back in synchrony. Since audio is played back at real-time playback speed, voxel frame sequences should also be played back at their capture frame rate. Voxelvideo actors include a frame buffer, which is asynchronously populated by loading frames from the voxelvideo file, and a second thread updates the renderer buffers at the correct rate. Frame buffer updates are triggered when the cumulative time from the start of playback matches the cumulative time from the start of the frame sequence (corrected for pause-time). Playback is therefore decoupled from file read variability, is not susceptible to cumulative errors of inter-frame delay errors, and will play back in real-time.

Voxelvideo playback can be configured to automatically play once the frame buffer is full and to loop when playback reaches the end of the video. Users of the framework can manually play, pause, resume, and re-start video playback.

Embodied Experiences: With our system, users can see themselves in the environment. Such an embodied representation relies on precisely aligning live voxelvideo streams with reality so that when the user is presented with the live voxelvideo via an immersive head-mounted display, they see their voxel reconstruction from a first-person perspective. Camera registration (Section 3.3) aligns reconstructions in the VR coordinate system such that this requirement is satisfied. Where the user can teleport to explore the scene, the position of their voxelvideo stream must also be updated to maintain the first-person view. Temporal delays between the user's initiated action and visual feedback are known to negatively impact the sense of agency Sato and Yasuda, 2005. In VIMR, temporal delays between action and visual feedback are typically between 100-120ms. This is similar to average self-perceived latency measured below 135ms in a previous study conducted with VIMR, which also confirmed the sense of embodiment N. J. Park, Regenbrecht, et al., 2019. Additionally, on average participants reported that the embodiment experience became disruptive at 255ms, and unbearable at 476ms.

Tele-co-presence: For a coherent experience of multiple users meeting as voxelstreams in one mutually shared virtual environment, all users require an instance of VIMR with at least one camera, and a copy of a UE4 project with the environment and voxelvideo recordings of the storyteller. Typically a standalone packaged copy of the UE4 project is distributed beforehand and configured to use the user's VIMR instance configuration in order to register the capture system to the VR system, and to receive live voxelvideo of the user. The instance configuration also lists the internet-accessible VNet server to use for negotiating P2P connections between users.

Synchronizing voxelvideo playback state, user teleportation, and supporting spatially

rendered voice communication relies on the Epic Online Subsystem[13] (EOS). This currently requires an epic games developer account[14] in order to connect to the EOS server and interact with EOS session management.

This general approach, and our approach to streaming voxels described in section 3.4, has been successfully used to demonstrate tele-co-presence between a number of locations. Notable tests include running the system between locations in New Zealand (Dunedin-Bluff, Dunedin-Christchurch), and between Dunedin, New Zealand and a collaborator in London, United Kingdom.

5 Lessons Learned & Future Work

We have assessed our system for its capacity to elicit presence, embodiment, and co-presence finding that our system does indeed support these experiential factors. Results from this work are published elsewhere J.-W. N. Park et al., 2019; N. Park et al., 2022; Regenbrecht et al., 2021; Regenbrecht et al., 2022, here we rather reflect on our collaborative work and experience engaging with potential users outside of the research lab. That is, we describe our ongoing work to examine how presence relates to tūrangawaewae and how co-presence with recorded storytellers and remote users can connect people to their community.

The overarching theme of the Ātea project is cultural and community impact. To that end, we have attended various community events hosted by Te Rau Aroha marae and the wider Māori community. Our attendance at these events was with the explicit goal of 'giving back' to the community and to our collaborators. With this goal in mind, and when attending Māori-hosted events in general, it is not always appropriate (culturally, or ethically) to conduct formal data collection. Furthermore, these events are attended by hundreds to thousands of people in a few days, and in order to engage with as many people

[13] https://docs.unrealengine.com/4.27/en-US/ProgrammingAndScripting/Online/EOS/
[14] https://dev.epicgames.com/

as possible, we cannot include extra steps before or after the experience in particular as we want users to spend time in the actual experience listening to the story.

We consequently decided to not conduct formal data collection during these events, but rather engage directly with the attendees and make informal observations about how people receive our work and how they react to the experiences. Over the last three years, we have attended several notable events with this goal in mind: an event in 2019 where local school students were invited to the marae to experience our system, a big annual event of the tribe we are working with (Ngāi Tahu Waitangi Day, Figure 12), a multi-day youth outreach event and community day in 2022, and an annual gathering and exhibition event (Hui-a-Iwi 2022) attracting thousands of attendees (Figure 12). Each of these events hosted hundreds to thousands of people with some affiliation to the marae or the Māori community, and we typically engaged with 100 or more people at each of these events. In precis, several hundred users have experienced our work and many of these have been members of the Māori community with a direct ancestral link to Te Rau Aroha marae.

Based on our informal observations during our attendance of these events, we can categorize users into three groups: 1) Novelty-seekers (e.g. children who want to play, or those intrigued by the novelty of VR but not necessarily the stories), 2) Indifferent users and those who don't even want to try, and 3) users who derive a meaningful experience from the virtual 'visit' to the marae and/or from the stories. For the purpose of this discussion, we ignore novelty-seekers in part because these are overwhelmingly children. We can mostly ignore the indifferent users as well since this was the smallest group by far, and note (encouragingly) that the worst-case response is (a non-negative) indifference to the experience. Our observations indicate that, however, both of these categories do 'get' the experience, and in some cases respond in accordance with tikanga. We then consider the third group, mainly to begin to understand what it is that they value about the experience and whether we can say that the experience conveys aspects of tūrangawaewae.

There are some general effects—we have observed some users expressing concern that they

wore shoes (in the real world) during the experience, where wearing shoes inside the wharenui would break tikanga. Similarly, users have asked about the concept of a virtual powhiri (a ceremony to formally welcome visitors to a marae, without which visitors are not allowed to enter the wharenui) in conjunction with visiting the virtual model of the wharenui. Users react to the storytellers, and in some cases feel as if their personal space is being invaded and attempt to move away from the storyteller if they're too close. Many users whakapapa (have ancestral links) to the marae through ancestors depicted as tall figures in each corner of the marae (Tipuna). Many of these users who have not yet visited the marae seek out their Tipuna as a first action when entering the virtual space. Likewise, many users express the desire to experience more of the stories if time allows for it. (There are eight stories in total, and each is about three minutes long. Listening to all of them during busy events is impractical); one user even responded to the karakia which preceded the story as if the storyteller was delivering it in person. While these informal observations do not demonstrate conclusively that tūrangawaewae with respect to a real place can be conveyed by way of virtual reality, it is encouraging that users often behave as if they are in a real marae, and listening to a real storyteller.

In addition, we have conducted two more formal events with selected members of the Māori community chosen for their experience or position in the tribe (iwi). One of these was a demonstration to academic staff at the University of Waikato, and another one was a study investigating the tele-co-presence experience between users in Christchurch and Dunedin. We have also demonstrated our work at the International Science Festival in Dunedin in 2021 (and an earlier version in 2018), with permission granted by Te Rau Aroha marae, where members of the general public have two days to visit the university to experience our work among other university projects. Another specific user group of interest is those with real-life experience of tikanga, including kaumātua—some of whom are experts in Māori language and tikanga. Two specific issues were raised by this group, regarding the appropriate representation of the storyteller. We presented an early 'talking

head' representation similar to the example in figure 7, and discovered that such representations are inappropriate because the head is *tapu* (sacred) and the representation disconnected from the rest of the body may imply decapitation. Similarly, the 'ghostly' appearance of voxelvideo recordings with sparse or missing parts (particularly missing parts of the head) was deemed inappropriate representations of living people. In general, it seems that some missing parts of the virtual body are acceptable—lower extremities, for example. It is not clear how this might change for voxelvideos representing the deceased, or whether the 'ghostly' appearance would then be appropriate. In general, the positivity with which our work has been received (and the absence of negative responses) suggests that our representation is appropriate.

It is clear that both the expert group and the broader group including non-experts with whakapapa to Te Rau Aroha marae, contribute critical feedback to our work. Particularly in engagement with the latter case, where we wish to find as many in group 3 as possible, the portability and encumbrance-free nature of our volumetric capture system have been critical in allowing many users to have these experiences. If the experience required reconstructing each user beforehand, or even simply instrumenting the user for motion capture, the number of people we could engage with at busy events would have been significantly reduced.

In addition to the cultural considerations, our system has some technical limitations; the coarse resolution of the voxel grid and the HMD obscuring the user's face are the two most common limitations pointed out by users. Wearing an HMD is necessary for the immersive experience, so we cannot directly capture the whole user's face in real-time using our current approach. Some HMDs now include cameras for eye-tracking, and these might be used to capture the eyes and/or to infer gaze. Voxel resolution is limited by the spatial resolution of the cameras, and by the bandwidth of the network connections used for streaming. We propose using a smaller number of cameras to reconstruct smaller sub-volumes of the overall capture volume in order to work camera resolution limits and to

Figure 12: Demonstrations of our system at Hui-a-Iwi 2022 in Temuka / Aotearoa (left) and Waitangi Day 2021 in Bluff / Aotearoa. Approximately 200 people used the system at Hui-a-Iwi and 100 at the Waitangi event. Users were able to explore the virtual wharenui and play recorded volumetric video stories, as well as meet people in the virtual space to experience the story together. For the Hui-a-Iwi the remote party (two of the authors) was located at the University of Otago's HCI Lab, and for the Waitangi event two portable instances of VIMR were installed in adjacent rooms, and users in both rooms would share the same virtual storytelling experiences.

add color compression to our voxelvideo frame encoding. Figure 7 shows the results of using a smaller capture volume which has allowed us to produce reconstructions with resolutions of 1 mm. At this resolution, the accuracy requirements of the calibration method must also be improved, and a more sophisticated method of fusing data from multiple cameras would be required to correct depth estimation errors. Voxel colors are currently represented by three bytes, however, we can use a lookup table to compress colors to one byte and reduce the total required bandwidth by one-third.

And finally, we note that the computational inefficiency of the voxel renderer requires relatively powerful graphics hardware to render in real-time. This requires the use of a gaming computer to experience stories, whereas many recent and more accessible VR systems are less powerful standalone systems such as the Oculus Quest and HTC Vive Go systems. In part, this is due to the limited access to the UE4 rendering pipeline, and we may benefit from choosing an alternative visual representation such as a splat-based method Chen et al., 2012.

We have outlined the potential value of immersive virtual storytelling for connecting people to real places and communities without having them visit those people/places in person. While our technical system captures relatively low levels of visual detail and has some significant limitations (e.g. not being able to see users' facial expressions), the accessibility of the system has been instrumental in engaging with enough people to find those users who can derive meaning from the experience. Those who have family relationships with Te Rau Aroha often have strong positive and emotional reactions to seeing places they have a strong personal connection to.

6 Conclusion

Despite the limitations of our system, our decision to trade level of detail for immediacy (real-time capture and visualization), portability, and accessibility has been critical in allowing us to engage with a broader audience—particularly with members of the community of the Te Rau Aroha marae and with the broader Ngāi Tahu iwi of whom many would not have the opportunity to experience our work if it remained constrained to our lab. Moreover, we have demonstrated that our system is capable of supporting a sense of presence in the model of the wharenui, co-presence with the storyteller, and social co-presence with remote users sharing the experience. The feedback we have received during our various engagements with these communities suggests that the sense of presence in this experience is relating to the concept of tūrangawaewae, as indicated by emotional responses to the experiences, and by unprompted suggestions that the users assume that they should follow the same tikanga as if they were visiting the marae in person. The effect is particularly strong for users who have either visited the real marae, or who know that their whakapapa (genealogy) connects them to the physical marae and community. Our focus is to present not only the technical system but to reflect on the experience of our collaboration with our Māori partners and the value of informal engagement with a large

group of non-academic users. We take it as a sign of success that Te Rau Aroha marae has decided to dedicate a room in their marae to volumetric video capture and to provide their community with the opportunity to experience immersive storytelling, and to virtually meet their whānau (family/extended family) who are unable to visit in person. Our hope is that our approach, and perhaps even our system, is useful to other people interested in conducting similar research.

Acknowledgments

We would like to thank the many people who have supported our work: Rosa Lutz has contributed considerably by helping to produce the content for the immersive experience, and to host exposition events. Members of the Otago University Human-Computer Interaction Lab have all contributed in one way or another. We thank Te Rau Aroha Marae in Bluff for the opportunity to work together—in particular Bubba and Gail Thompson, and Dean and Hēmi Whaanga. This work was partially funded by the National Science Challenges, Science for Technological Innovation, Ātea Spearhead project.

Māori Terms[15]

Ātea: Be clear, free from obstruction.

Kaumātua: Adult, elder, elderly man, elderly woman, old man - a person of status within the *whānau* (extended family).

Marae: Courtyard - the open area in front of the *wharenui*, where formal greetings and discussions take place. Often also used to include the complex of buildings around the *marae*.

[15]Translations are based on the Te Aka Māori Dictionary: https://maoridictionary.co.nz

Mātauranga: Knowledge, wisdom, understanding, skill - sometimes used in the plural. Also education as an extension of the original meaning and commonly used in modern Māori with this meaning.

Tapu: Be sacred, prohibited, restricted, set apart, forbidden, under *atua* (God) protection

Tikanga: Correct procedure, custom, habit, lore, method, manner, rule, way, code, meaning, plan, practice, convention, protocol - the customary system of values and practices that have developed over time and are deeply embedded in the social context.

Tūrangawaewae: Domicile, standing, place where one has the right to stand - place where one has rights of residence and belonging through kinship and *whakapapa* (genealogy, descent).

Whakapapa: Genealogy, genealogical table, lineage, descent—reciting whakapapa was, and is, an important skill and reflected the importance of genealogies in Māori society in terms of leadership, land and fishing rights, kinship and status. It is central to all Māori institutions.

Whānau: extended family, family group, a familiar term of address to a number of people - the primary economic unit of traditional Māori society. In the modern context, the term is sometimes used to include friends who may not have any kinship ties to other members.

Wharenui: Meeting house, large house - main building of a *marae* where guests are accommodated. Traditionally the *wharenui* belonged to a *hapū* (kinship group, tribe) or *whānau* (extended family) but some modern meeting houses, especially in large urban areas, have been built for non-tribal groups, including schools and tertiary institutions. Many are decorated with carvings, rafter paintings and *tukutuku* (ornamental lattice-work) panels.

References

Chen, M., Kaufman, A. E., & Yagel, R. (2012). *Volume graphics*. Springer Science & Business Media.

Collins, J., Regenbrecht, H., & Langlotz, T. (2017). Visual Coherence in Mixed Reality: A Systematic Enquiry. *Presence: Teleoperators and Virtual Environments*, *26*(1), 16–41. https://doi.org/10.1162/PRES_a_00284

Curless, B., & Levoy, M. (1996). A volumetric method for building complex models from range images. *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '96*, 303–312. https://doi.org/10.1145/237170.237269

Gamelin, G., Chellali, A., Cheikh, S., Ricca, A., Dumas, C., & Otmane, S. (2021). Point-cloud avatars to improve spatial communication in immersive collaborative virtual environments. *Personal and Ubiquitous Computing*, *25*(3), 467–484.

Garau, M., Friedman, D., Widenfeld, H. R., Antley, A., Brogni, A., & Slater, M. (2008). Temporal and spatial variations in presence: Qualitative analysis of interviews from an experiment on breaks in presence. *Presence: Teleoperators and Virtual Environments*, *17*(3), 293–309.

Gunkel, S. N., Hindriks, R., Assal, K. M. E., Stokking, H. M., Dijkstra-Soudarissanane, S., Haar, F. t., & Niamut, O. (2021). Vrcomm: An end-to-end web system for real-time photorealistic social vr communication. *Proceedings of the 12th ACM Multimedia Systems Conference*, 65–79.

Guven, S., Podlaseck, M., & Pingali, G. (2009). Exploring co-presence for next generation technical support. *2009 IEEE Virtual Reality Conference*, 103–106.

Hauber, J., Regenbrecht, H., Hills, A., Cockburn, A., & Billinghurst, M. (2005). Social presence in two-and three-dimensional videoconferencing.

Hoetzlein, R. K. (2016). GVDB: Raytracing Sparse Voxel Database Structures on the GPU. *Proceedings of High Performance Graphics*, 109–117. https://doi.org/10.2312/hpg.20161197

Horaud, R., & Dornaika, F. (1995). Hand-eye calibration. *The international journal of robotics research*, *14*(3), 195–210.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, *34*(5), 827–828. https://doi.org/10.1107/S0567739478001680

Kammerl, J., Blodow, N., Rusu, R. B., Gedikli, S., Beetz, M., & Steinbach, E. (2012). Real-time compression of point cloud streams. *2012 IEEE International Conference on Robotics and Automation*, 778–785.

Kilteni, K., Groten, R., & Slater, M. (2012). The Sense of Embodiment in Virtual Reality. *Presence: Teleoperators and Virtual Environments*, *21*(4), 373–387. https://doi.org/10.1162/PRES_a_00124

Lindlbauer, D., & Wilson, A. D. (2018). Remixed Reality: Manipulating Space and Time in Augmented Reality. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3173703

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, *34*(6), 248:1–248:16. https://doi.org/10.1145/2816795.2818013

Meng, K. (2016). *Mixed Reality Embodiment Platform – Recording and replay of volumetric characters* (Masters). University of Koblenz and University of Otago.

Minsky, M. (1980). Telepresence. *OMNI Magazine*.

Niamut, O., D'Acunto, L., & Havinga, P. (n.d.). SOCIAL XR: BARRIER-BREAKING TECHNOLOGY FOR SMART SOCIETIES.

Park, J.-W. N., Mills, S., Whaanga, H., Mato, P., Lindeman, R. W., & Regenbrecht, H. (2019). Towards a Māori Telepresence System. *2019 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 1–6. https://doi.org/10.1109/IVCNZ48456.2019.8961016

Park, N., Regenbrecht, H., Duncan, S., Mills, S., Lindeman, R. W., Pantidi, N., & Whaanga, H. (2022). Mixed Reality Co-Design for Indigenous Culture Preservation amp; Continuation. *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 149–157. https://doi.org/10.1109/VR51125.2022.00033

Park, N. J., Regenbrecht, H. et al. (2019). Resolutions and network latencies concerning a voxel telepresence experience. *Journal of Software Engineering and Applications*, *12*(05), 171.

Regenbrecht, H., Meng, K., Reepen, A., Beck, S., & Langlotz, T. (2017). Mixed Voxel Reality: Presence and Embodiment in Low Fidelity, Visually Coherent, Mixed Reality Environments. *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 90–99. https://doi.org/10.1109/ISMAR.2017.26

Regenbrecht, H., Ott, C., Park, N., Duncan, S., & Collins, J. (2021). Voxelvideos for Entertainment, Education, and Training. *IEEE Access*, *9*, 68185–68196. https://doi.org/10.1109/ACCESS.2021.3076488

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

Regenbrecht, H., Park, N., Duncan, S., Mills, S., Lutz, R., Lloyd-Jones, L., Ott, C., Thompson, B., Whaanga, D., Lindeman, R. W., Tong, K., Clifford, R., Jones, N., Mato, P., Keegan, T. T., & Whaanga, H. (2022). Ātea Presence— Enabling Virtual Storytelling, Presence, and Tele-Co-Presence in an Indigenous Setting. *IEEE Technology and Society Magazine*, *41*(1), 32–42. https://doi.org/10.1109/MTS.2022.3147525

Sato, A., & Yasuda, A. (2005). Illusion of sense of self-agency: discrepancy between the predicted and actual sensory consequences of actions modulates the sense of self-agency, but not the sense of self-ownership. *Cognition*, *94*(3), 241–255. https://doi.org/10.1016/J.COGNITION.2004.04.003

Schreer, O., Feldmann, I., Ebner, T., Renault, S., Weissig, C., Tatzelt, D., & Kauff, P. (2019). Advanced Volumetric Capture and Processing. *SMPTE Motion Imaging Journal*, *128*(5), 18–24. https://doi.org/10.5594/JMI.2019.2906835

Schreer, O., Feldmann, I., Renault, S., Zepp, M., Worchel, M., Eisert, P., & Kauff, P. (2019). Capture and 3D Video Processing of Volumetric Video. *2019 IEEE International Conference on Image Processing (ICIP)*, 4310–4314. https://doi.org/10.1109/ICIP.2019.8803576

Schubert, T., Friedmann, F., & Regenbrecht, H. (1999). Embodied Presence in Virtual Environments. In R. Paton & I. Neilson (Eds.), *Visual Representations and Interpretations* (pp. 269–278). Springer London.

Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The Experience of Presence: Factor Analytic Insights. *Presence: Teleoperators and Virtual Environments*, *10*(3), 266–281. https://doi.org/10.1162/105474601300343603

Triggs, B., McLauchlan, P. F., Hartley, R. I., & Fitzgibbon, A. W. (1999). Bundle adjustment—a modern synthesis. *International workshop on vision algorithms*, 298–372.

Tsai, R., & Lenz, R. (1989). A new technique for fully autonomous and efficient 3D robotics hand/eye calibration. *IEEE Transactions on Robotics and Automation*, *5*(3), 345–358. https://doi.org/10.1109/70.34770

Valenzise, G., Alain, M., Zerman, E., & Ozcinar, C. (Eds.). (2023). *Immersive Video Technologies*. Academic Press, an imprint of Elsevier
OCLC: 1346536558.

Waltemate, T., Gall, D., Roth, D., Botsch, M., & Latoschik, M. E. (2018). The Impact of Avatar Personalization and Immersion on Virtual Body Ownership, Presence, and Emotional Response. *IEEE Transactions on Visualization and Computer Graphics*, *24*(4), 1643–1652. https://doi.org/10.1109/TVCG.2018.2794629

Young, G., O'Dwyer, N., & Smolic, A. (2023). Volumetric video as a novelmedium for creative storytelling. In G. Valenzise, M. Alain, E. Zerman, & C. Ozcinar (Eds.), *Immersive video technologies* (pp. 591–608). Academic Press, an imprint of Elsevier.

Zerman, E., Ozcinar, C., Gao, P., & Smolic, A. (2020). Textured Mesh vs Coloured Point Cloud: A Subjective Study for Volumetric Video Compression. *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 1–6. https://doi.org/10.1109/QoMEX48832.2020.9123137