William Steptoe*
Department of Computer Science
University College London
London WC1E 6BT UK

Jean-Marie Normand
Event-Lab
University of Barcelona

Oyewole Oyekoya
Fabrizio Pece
Department of Computer Science
University College London
London WC1E 6BT UK

Elias Giannopoulos
Event-Lab
University of Barcelona

Franco Tecchia
PERCRO
Scuola Superiore Sant'Anna

Anthony Steed
Tim Weyrich
Jan Kautz
Department of Computer Science
University College London
London WC1E 6BT UK

Mel Slater
Department of Computer Science
University College London
London WC1E 6BT UK
and
ICREA
University of Barcelona

Acting Rehearsal in Collaborative Multimodal Mixed Reality Environments

Abstract

This paper presents the use of our multimodal mixed reality telecommunication system to support remote acting rehearsal. The rehearsals involved two actors, located in London and Barcelona, and a director in another location in London. This triadic audiovisual telecommunication was performed in a spatial and multimodal collaborative mixed reality environment based on the "destination-visitor" paradigm, which we define and put into use. We detail our heterogeneous system architecture, which spans the three distributed and technologically asymmetric sites, and features a range of capture, display, and transmission technologies. The actors' and director's experience of rehearsing a scene via the system are then discussed, exploring successes and failures of this heterogeneous form of telecollaboration. Overall, the common spatial frame of reference presented by the system to all parties was highly conducive to theatrical acting and directing, allowing blocking, gross gesture, and unambiguous instruction to be issued. The relative inexpressivity of the actors' embodiments was identified as the central limitation of the telecommunication, meaning that moments relying on performing and reacting to consequential facial expression and subtle gesture were less successful.

1 Introduction

The Virtuality Continuum (Milgram & Kishino, 1994) describes virtual reality (VR) -related display technologies in terms of their relative extents of presenting real and virtual stimuli. The spectrum ranges from the display of a real environment (e.g., video) at one end, to a purely synthetic virtual environment (VE) at the other. Mixed reality (MR) occupies the range of the continuum between these extrema, merging both real and virtual objects. MR was originally considered on a per display basis, and later broadened to consider the joining together of distributed physical locations to form MR environments (Benford, Greenhalgh, Reynard, Brown, & Koleva, 1998). When discussing MR displays, it is sufficient to do so with regard to Milgram and Kishino's evolving taxonomy (Milgram & Kishino, 1994) that ranges from handheld devices through to immersive projection technologies (IPTs) such as the CAVE and head-mounted displays (HMDs). However, MR environments as outlined by Benford et al. (1998) bring together a range of technologies, including

Presence, *Vol. 21, No. 4, Fall 2012,* 406–422

*Correspondence to w.steptoe@cs.ucl.ac.uk.

situated and mobile displays and capture devices, with the aim of supporting high-quality spatial telecommunication. The work presented in this paper concerns a system based on an emerging mode of telecommunication that we refer to as the "destination-visitor" paradigm. We assess our distributed heterogeneous MR environment (that also features MR displays) through a case study of remote acting rehearsal.

Telecommunication has long been a driving force behind the development of MR- and VR-related technologies. VR's flexible, immersive, and spatial characteristics provide several opportunities over common modes of audiovisual remote interaction such as video conferencing. In the case of remote acting rehearsal, the use of a nonimmersive VE allowing actors to both control their own avatars and observe others within a shared virtual space was found to be able to form a basis for successful live performance that may not have been achievable using minimal video conferencing (Slater, Howell, Steed, Pertaub, & Garau, 2000). Another study investigating object-focused puzzle solving found that pairs of participants interacting between remote IPTs were able to perform the task almost as well as when they were collocated, and significantly better than when one participant used an IPT while the other used a nonimmersive desktop system (Schroeder et al., 2001). This latter arrangement in Schroeder et al.'s study is an example of asymmetric telecommunication, in the sense that the technological experience of the two interacting participants was quite different: one was immersed in fully-surrounding IPT featuring body tracking and perspective-correct stereoscopic rendering, while the other used standard input devices on a 2D consumer display.

Asymmetry pertains both to the work presented in this paper, and to an ongoing tension between the technological asymmetries often intrinsic to media spaces and our natural desire for social symmetry between participants (A. Voida, S. Voida, Greenberg, & He, 2008). Traditionally, media space research has striven for technological symmetry: an aim that is likely borne from our daily experience of reciprocity in collocated face-to-face interaction, in which the same sensory cues and affordances are generally available to all participants (i.e., social symmetry). Studies in the VE literature illustrate that technological asymmetry affects the social dynamics of virtual interaction. Indeed, participants interacting over a technologically asymmetric system are unable to make discriminating judgments about joint experience (Axelsson et al., 1999). Immersing one participant to a greater degree than their interactional partner makes it significantly more likely for the more-immersed participant to be singled out as the social leader (Slater, Sadagic, Usoh, & Schroeder, 2000), and for the less-immersed participant to be evaluated as contributing less to cooperative tasks (Schroeder et al., 2001). The correlation between technological and social symmetry has significant implications for the conduct of both personal and business communication, and this is likely to account for the traditional desire to design for technological symmetry.

Opposing this tradition, recent innovations in MR telepresence research are embracing technological asymmetry. This trend has arisen naturally, due to increased distribution among teams that are mostly collocated except for one or two members. Acknowledging this geographical distribution in personnel, recent research has focused particularly on how to present remote participants in MR environments. This has been achieved through the use of both situated displays (Venolia et al., 2010) and mobile personal telepresence robots (Tsui, Desai, Yanco, & Uhlik, 2011). Technology used in such MR environments aims to augment a place with virtual content that provides communicational value to the connected group members. Thus, the central concept of this paradigm involves people at a destination (also termed a hub, Venolia et al., 2010; Tsui et al., 2011) that is augmented with technologies aiming to both represent and bestow visitors (also termed satellites, Venolia et al., 2010; or spokes, Tsui et al., 2011) with both physical and social presence to support high-quality remote interaction.

Throughout this paper, we refer to such systems as being examples of the destination-visitor paradigm, and the system that we detail in the following section is an archetype of this paradigm. Studies of systems adopting the destination-visitor paradigm have so far focused on optimizing the social presence of a visitor for the benefit

of those at the destination, but have focused less on the experience of that visitor. Consequently, while the visitor is often represented at the destination with a high degree of physical presence (e.g., as a mobile telepresence robot or a situated display), the destination and its collocated people are not represented to the visitor with an analogous level of fidelity and spatiality. Hence, in systems such as those presented in Venolia et al. (2010) and Tsui et al. (2011), social asymmetry is likely to arise from this unequal consideration of locals' and visitors' sensory experiences that arise due to the differing technological quality over the two sites.

Our approach to the destination-visitor paradigm presented in this paper aims to provide a more consistent degree of social symmetry to all engaged parties interacting via the multimodal and highly technologically-asymmetric system. Hence, a visitor's visual experience of both the destination and of its copresent locals is provided by a range of high-fidelity immersive MR display modes viewed from a first-person perspective. Meanwhile, at the destination site, the visitor is represented using a range of display modes that allow both full-body gesture and movement to be observed, and spatial attention to be naturally perceived. The system aims to provide an audiovisual telecommunication platform featuring both communicational and spatial consistency between all engaged parties, whether they are locals or visitors, thereby fostering social symmetry in an otherwise technologically asymmetric system.

The following section details our system topology, which grants a flexible approach to the destination-visitor paradigm and supports a variety of synchronous networked capture and display technologies. The acting rehearsals are then discussed, exploring the actors' and director's experience of performing together via the system. We decided to study acting, as opposed to pure social interaction and collaboration, which is the principal application of telecommunication systems, for several reasons. The aim of a theatrical rehearsal is to practice and refine the performance of a scene. We chose a scene consisting of varied spatial and interpersonal interplay between two characters. Thus, the actors engage in rich conversation, spatial action, directing attention, and handling objects. Such activity is commonplace in the type of unplanned social interaction and collaborative work that the system aspires to support. Hence, through repeated and evolving run-throughs of the scene with professional actors and a director, this structured activity forms an excellent basis for analytic knowledge regarding the successes and failures of the system. Throughout this paper, we are primarily concerned with the visual mode of telecommunication. While there are several other sensory cues we are actively working on, including spatialized audio and touch, these are scoped for future work.

This work is a continuation of a previous setup and study that investigated acting in an immersive collaborative VE (Normand et al., 2012). The current paper introduces the destination-visitor paradigm, which presents a highly asymmetric multimodal collaborative MR environment.

2 System Architecture

This section presents a high-level overview of our flexible and heterogeneous system architecture based on the destination-visitor paradigm. In Section 3, we investigate how well the system is able to support geographically remote acting rehearsal between two actors and one director. Hence, the setup we present is tailored to suit this spatially and visually dependent application scenario. It should be stated that the technological arrangement is flexible, and different acquisition and display devices, and also visitor sites, may be introduced or removed. Figure 1 illustrates the distinct arrangements at each of the three sites in our particular studied setup: the physical destination, where one actor will be located, is equipped with a range of capture and display technologies; the visitor site, at which the second actor will be located, is composed of an immersive HMD-based VR system with full-body tracking; and the director's setup is an immersive CAVE-like system, although it could be a standard machine located anywhere.

In this triadic interaction, there are two people (the visiting actor and the director) who may be classed, according to the paradigm's specifications, as visitors.

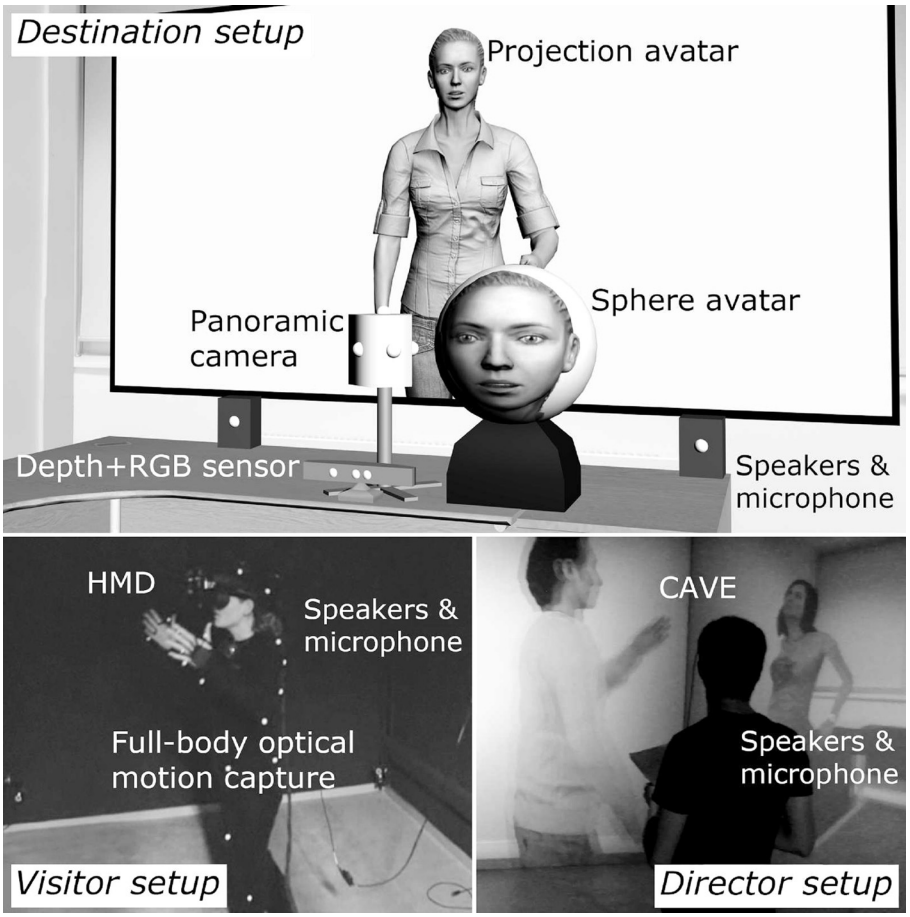

Figure 1. *Asymmetric technical arrangements at the three sites to acquire and display places and people. The destination and visitor sites feature varying acquisition and display technologies, while the director (spectator) site only requires display technology.*

However, the director is more comfortably classified as a spectator, as he is not visually represented to either actor. This enables the director to navigate through the rehearsal space without causing visual distraction or occlusion to the two actors. Acquisition and display technologies at each site are now discussed, followed by an overview of the transmission protocols operating between sites for the various media streams. Acquisition at the destination site dictates the display characteristics at the visitor site and vice versa. We focus solely on the visual components of the system, as aural communication is supported using Skype.[1] We identify spatialized 3D audio as an area of future work.

1. http://www.skype.com

2.1 Destination Site

The destination site is a physical meeting room, approximately $5 \times 3 \times 3$ m, where: the destination actor is physically located, the visiting actor will remotely perceive and be virtually represented, and the director will remotely perceive including virtual representations of both actors. Hence, the destination must be equipped with technology able to acquire both the environment and the copresent actor in order to transmit to the visitor and director sites, and also with display technology to represent the visiting actor. The unique feature of the destination is that it should largely remain a standard meeting room to any collocated people, in the sense that they should not be encumbered by worn devices

Figure 2. *The three display modes available at the visitor site from left to right: spherical video from the Ladybug 3, a hybrid approach featuring embedded 2.5D Kinect video of the destination actor within a VE, and a pure VE featuring Kinect-tracked embodied avatars.*

such as wired sensors or HMDs in order to partake in the interaction.

2.1.1 Acquisition. We have implemented two methods of visual capture of the local environment, and three methods of visual capture of the local actor. Figure 2 shows how all of these modes appear at the visitor site, and these are discussed in Section 2.2. Spherical ($\sim$288°) video acquisition, which implicitly captures both actor and environment, is achieved with a Point Grey Research Ladybug 3 camera.[2] The camera combines the views acquired by six 2 MP (1600 $\times$ 1200) Sony CCD sensors into a single panoramic image, which has a 12 MP (5400 $\times$ 2700) resolution. The process of de-mosaic, conversion, and stitching together the different views is entirely done in software on the host machine. This solution, while giving control of the acquired image, creates an overhead in the computational time required to acquire a surround video. Hence, the Ladybug 3 is capable of recording surround video at the relatively low frame rate of 15 Hz at full resolution. This results in a potential bandwidth requirement of $\sim$90 MB/s (assuming that each pixel is described as an RGB, 8-bit unsigned character). However, this unmanageable wide-area bandwidth requirement is drastically reduced to $\sim$1 MB/s by reducing the sent resolution to 2048 $\times$ 1024 pixels and using compression algorithms that we cover in Section 2.4. In summary, the Ladybug 3 provides a simple means of visually capturing the destination and the people within. This surrounding acquisition is directly compatible with the immersive display characteristics at the visitor and director sites, both of which feature wide field-of-view displays.

The second method of environment capture is a hand-modeled premade 3D textured scale model of the destination room. The model's wall textures are reprojected from a 50 MP ($10,000 \times 5,000$) panorama, while the furniture textures are extracted from single photographs. During system use, the means of destination acquisition determines key aspects of the visitor's and spectator's experience, including dynamism, spatiality, and fidelity. These issues are outlined in the following sections describing display at the visitor and director sites, and are explored during system use in Section 3 covering the acting rehearsals.

As described above, the first method of real-time dynamic visual capture of the actor at the destination site is achieved with the spherical video acquired by the Ladybug 3. The other two modes of capturing the destination actor make use of the abilities of Microsoft's Kinect (depth plus RGB) sensor.[3] The Kinect features a depth sensor consisting of an infrared laser projector combined with a monochrome CMOS sensor. This allows the camera to estimate 3D information under most ambient lighting conditions. Depth is estimated by projecting a known pattern of speckles into the scene. The way that these speckles are distributed in the scene will allow the sensor to reconstruct a depth map of the environment. The hardware consists of two cameras that output two videos at 30 Hz, with the RGB video

2. http://www.ptgrey.com

3. www.xbox.com/kinect

stream at 8-bit VGA resolution (640 × 480 pixels) and the monochrome video stream used for depth sensing at 11-bit VGA resolution (640 × 480 pixels with 2,048 levels of sensitivity).

The first Kinect-based solution for capturing the destination actor makes use of the OpenNI and NITE middleware libraries.[4] OpenNI (Open Natural Interaction) is a multi-language, cross-platform framework that defines APIs for writing applications utilizing natural interaction. Natural interaction is an HCI concept that refers to the interface between the physical and virtual domains based on detection of human action and behavior, in particular bodily movement and voice. To this end, PrimeSense's NITE (Natural Interaction Technology for End-user) adds skeletal recognition functionality to the OpenNI framework. Making use of the depth-sensing abilities of the Kinect, NITE is able to track an individual's bodily movement, which it represents as a series of joints within a skeletal hierarchy. While the tracking data are not as high-fidelity, high-frequency, or low-latency as a professional motion capture system (as discussed in Section 2.2 detailing the visitor site), NITE has the significant advantages of being markerless, and requiring minimal technical setup and calibration time. The calculated skeletal data are transmitted to the visitor and destination site at 30 Hz, and are used to animate a graphical avatar representing the destination actor. The avatar is positioned correctly within the VE representation mode of the destination, but would not be suited to the 2D spherical Ladybug 3 video mode. This highlights the interplay between capture at one site and display at another.

The final mode of capturing the destination actor, and the second Kinect-based mode, streams a 2.5D point-based textured video representation of the actor, independent of the environment, at 30 Hz. This representation may be considered a 2.5D video avatar, as we only employ one front-facing Kinect to record the actor, and thus do not provide coverage of the rear half of the body. It is likely possible to use several Kinects to ensure that the destination actor is fully captured, as shown by Maimone and Fuchs (2012), but there are potential interference problems and this is an area for future work. In order to segment the destination actor from his environment, we combine knowledge from the NITE-based skeletal tracking in order to only consider pixels within the depth range at which the tracked joints are currently positioned.

4. http://openni.org

2.1.2 Display. Display technology located at the destination is responsible for representing the visiting actress in a manner that fosters a physical presence. We consider mobile telepresence robots (Tsui et al., 2011) and embodied social proxies (Venolia et al., 2010) to bestow the wearer with a high degree of physical presence as they provide a mobile or situated totem by which they may be spatially located, looked toward, and referred to. To this end, our current system architecture offers two solutions: a large high-resolution avatar projection and a novel 360° spherical display showing an avatar head only. While both are not mobile (this is also a limitation of the current destination acquisition technology, and discussed in Section 3), and so restrict the visitor's gross movement around the destination, each display type aims to provide a distinct benefit to the remote interaction.

Firstly, the projection avatar enables life-size and full-body embodiment of the visitor. Due to the corresponding full-body motion capture setup at the visitor site (detailed in the following section), the avatar is puppeteered in real time. Thus, nuances of the visiting actress' body language are represented. Due to the large 3 × 2 m screen size, gross movement on the horizontal and vertical axes, and to a lesser extent on the forward-backward axis, are also supported. So, for instance, the actions of the visiting actress pacing from left to right, sitting down, and standing up, will be shown clearly by the projection avatar. Furthermore, this movement will be consistent both with respect to where the visiting actress perceives herself to be in the destination and where the destination actor perceives her to be. However, forward and backward movement is limited, as the visual result of such movement is that the visitor's representation moves closer or farther away from the virtual camera, resulting either in increased or reduced size rather than being displayed at the physical posi-

tion in the destination. Stereoscopic visualization could bestow the projection avatar with a more convincing sense of the depth at which visitor is positioned, but (unlike a CAVE), the single-walled display featured at the destination would restrict this illusion.

The second mode of visitor display is the use of a Global Imagination Magic Planet.[5] This spherical display aims to enhance the ease of determining the 360° directional attention of the visiting actress, and also to bestow her with a greater sense of physical presence at the destination. The avatar head as displayed on the Magic Planet rotates and animates in real time based on head tracking, eye tracking, and voice-detection data acquired at the visitor site. So, as the visiting actress rotates her head, directs her gaze, and talks, her sphere avatar appears to do the same. Due to the alignment between all participants' perception of the shared space and the people within, the sphere avatar allows for accurate 360° gaze awareness. The nature of the head-only situated sphere avatar implies that it is unable to represent the visiting actress' movement around the destination space. However, due to the system's flexible architecture, the simultaneous dual-representation of both projection and sphere avatars is supported, harnessing the benefits of both display types. Stereo speakers and a microphone support verbal communication. Implementation details, together with a user study, can be found in Oyekoya, Steptoe, and Steed (2012).

2.2 Visitor Site

The technology at the visitor site is responsible for both capture of the visiting actress and the immersive display of the destination and its collocated actor. In our current setup, this comprises a VR facility at which the technologies for acquisition and display are a full-body motion capture system and an immersive HMD, respectively. The physical characteristics of the visitor site are largely inconsequential, as the visitor will be immersed in a virtual representation of the destination, and the acquisition that occurs here is solely that of the visitor herself. We advocate the use of state-of-the-art IPT through the desire to foster social symmetry between all remote interactants. To this end, the HMD stimulates the wearer's near-complete field-of-view, thus displaying a spatial and surrounding representation of the destination and the locals within, while the motion-capture data transmit body movement and nonverbal subtleties from which the visitor's avatar embodiment is animated at the destination site.

2.2.1 Acquisition. Capture of the visitor is achieved using a NaturalPoint Optitrack[6] motion capture system consisting of twelve cameras. The motion capture volume is approximately $3 \times 3 \times 2.5$ m, which allows a one-to-one mapping between the visitor's movements in the perceived virtual destination, and the position of her embodiment at the physical destination. By tracking the positions of reflective markers attached to the fitted motion capture suit, the system calculates near full-body skeletal movement (not fingers or toes) at submillimeter precision at 100 Hz. This skeletal data has higher-fidelity than the equivalent Kinect NITE tracking at the destination, but with the trade-off that the visiting actress must wear a motion capture suit as shown in Figure 1, and the total setup and calibration time is greater ($\sim$20 m as opposed to $<$5 min). The data are converted ready for direct application to the projection and sphere avatar rigs at the destination site, and are then transmitted. More information on transmission protocols is provided in Section 2.4.

2.2.2 Display. Display of the destination and its local actor to the visitor is achieved using an NVis nVisor SX111 HMD.[7] The HMD has a 111°h $\times$ 64°v field of view and a resolution of 1280×1024 displayed at 60 Hz. The visual modes captured and transmitted from the destination have been detailed in the previous section, and are illustrated in Figure 2. These three modes may be dynamically swapped between and include the Ladybug 3's spherical video, hybrid 2.5D Kinect video of the destination actor embedded in a

5. http://www.globalimagination.com

6. http://www.naturalpoint.com/optitrack
7. http://www.nvisinc.com/product.php?id=48

VE, and a pure VE featuring Kinect-tracked and animated avatars. Note that in all three modes, the visitor can look down and see her own virtual body: an important visual cue for spatial reference and presence in a VE (Mohler, Creem-Regehr, Thompson, & Bulthoff, 2010). The three modes are spatially aligned, and provide a surrounding visual environment based on the physical destination room. VRMedia's XVR (Tecchia et al., 2010) software framework is used to render the VE, and the avatars are rendered using the Hardware Accelerated Library for Character Animation (HALCA; Gillies & Spanlang, 2010). Stereo speakers and a microphone are used similarly at the destination to support verbal communication.

2.3 Director Site

Since the director is not represented visually to the actors, the director site does not require any specific acquisition technology. Hence, we mainly focus on the director site's display system. In order to conduct and critique the rehearsal between the two remote actors, the director must have audiovisual and spatial reference to the unfolding interaction. The audiovisual component could be achieved through the use of a standard PC displaying the shared VE, and navigated using standard input devices. However, this would diminish the director's ability to naturally instruct the actors, and observe and refer to locations in the destination. Hence, we position the director in a four-walled CAVE-like system, where a surrounding and perspective-correct stereoscopic view of the destination VE and the two actors is displayed. This enables the director to navigate freely and naturally in the rehearsal space, shifting his viewpoint as desired through head-tracked parallax. Thus, the director views a VE of the destination, populated with the avatar embodiments of the two remote actors.

Due to the differing acquisition and display technologies at the destination and visitor sites, an asymmetry is introduced between the two actors in terms of both fidelity and movement ability. The relative low fidelity of the destination's Kinect NITE-based tracking compared with that of the visitor's Optitrack motion capture has been specified. The implications of this asymmetry is that the destination actor may be perceived to have fewer degrees of freedom (e.g., NITE does not track head, wrist, and ankle orientation) than the visiting actress, and as a result may appear more rigid and, due to the lower capture rate, less dynamic. The restriction upon the visitor's range of movement, due to the situated nature of the displays at the destination, has also been noted. This restriction must simply be acknowledged when planning the artistic direction of the scene. These issues are further expounded in Section 3. Headphones and a microphone complete the verbal communication.

2.4 Transmission

Communication between participants distributed over the three international sites relies on low-latency data transmission. The multimodal nature of the various media streams originating from the range of acquisition devices at the destination and visitor sites means that a monolithic transmission solution is inappropriate, and would likely result in network congestion. Hence, we divide the media into two types by bandwidth requirement.

Low-bandwidth data comprise session management and skeletal motion capture data based on discrete shared objects described by numeric or string-based properties. A client-server replicated shared object database using the cross-platform C++ engine RakNet by Jenkins Software is adopted.[8] Clients create objects and publish updates, which are then replicated across all connected clients. Objects owned by a client include one to describe the client's properties, and then as many as there are joints in the chosen avatar rig. As a tracked actor moves and gestures, the owning client will update the properties (generally positions and rotations) of the corresponding objects. These changes are then serialized and updated on all connected client databases. Different clients may be interested in different data. For instance, the client running the sphere avatar at the destination will update information received from the visiting actress' head node only, while the client

8. http://www.jenkinssoft.com

running the director's CAVE system will update the full-body motion capture data from both actors and animate the avatars accordingly. Objects that are owned by other clients may be queried, retrieving any updates since the last query. Most of the data exchanged via the server arises from skeletal motion capture at the destination and visitor sites. The primary purpose of this data transfer is for the real-time visualization of the actors' avatar embodiments: the visitor actress being displayed at both the destination and director sites, and the destination actor being displayed at the director site, and, depending on visualization mode, at the visitor site. An orthogonal use of the data is session logging for post analysis and replay. This logging is performed on the server, and writes all node updates to a human-readable file that is time-stamped from a central time server. In addition, the server also records an audio file of all participants' talk in OGG-Vorbis format.[9] Log files may be replayed both on nonimmersive and immersive displays including the CAVE and HMDs.

High-bandwidth data are composed of video acquired from the Ladybug 3 and Kinect cameras at the destination site and transmitted to the visitor site for display in the HMD. Several solutions to encode, transmit, and decode video streams, including the transmission of color-plus-depth data, were investigated. Regarding color-plus-depth, it may seem that one should be able to stream these depth videos using standard video codecs, such as Google's VP8[10] or H.264.[11] However, the quality degrades considerably, because the compression algorithms are geared toward standard three-channel (8-bit) color video, whereas depth videos are single-channel but have a higher bit depth (e.g., Kinect uses 11-bit depth). To this end, we have developed a novel encoding scheme that efficiently converts the single-channel depth images to standard 8-bit three-channel images, which can then be streamed using standard codecs. Our encoding scheme ensures that the sent depth values are received and decoded with a high degree of accuracy and is detailed in Pece, Kautz, and Weyrich (2011).

Our current end-to-end video transmission solution implements Google's VP8 encoding with RakNet streaming. The process is as follows: Microsoft's DirectShow[12] is used to initially acquire images from a camera source. The raw RGB images are converted to YUV space, after which the YUV image is compressed using the libvpx VP8 codec library.[13] The compressed frames are then sent as a RakNet bitstream to a server process (in our case located in Pisa, Italy), which simply relays the stream to other connected peers such as the visitor site running the HMD. (RakNet's bitstream class is a mechanism to compress and transmit generic raw data.) Upon being received, the compressed VP8 frames are decompressed to YUV and then converted back to RGB space. The OpenGL-based XVR renderer then transfers the RGB buffer to textures for final display. With the Ladybug 3, our implementation achieves frame rates of ~13 Hz (from the original 15 Hz) in the visitor's HMD, while the end-display frame rate of the Kinect 2.5D video runs at ~20 Hz (from the original 30 Hz). End-to-end latency of transmitted frames is <200 ms for both cameras.

2.5 Summary

Figure 3 presents a simplified view of the system's major components and links between processes. The three sites are represented by the large boxes, while the the inner boxes within each site each represent a machine that is responsible for at least one element of the overall system. Hardware is signified by a preceding "–" symbol, while software processes are signified by a "+." The lines between processes indicate wide-area network transmission, and the mode of data transmission between sites varies based on media. High-bandwidth video streams are shown as solid lines, while low-bandwidth data, including session management and skeletal motion capture data sent between clients

9. http://www.vorbis.com
10. http://www.webmproject.org/tools/vp8-sdk
11. http://www.itu.int/rec/T-REC-H.264
12. http://msdn.microsoft.com/en-us/library/ms783323.aspx
13. http://www.webmproject.org/code

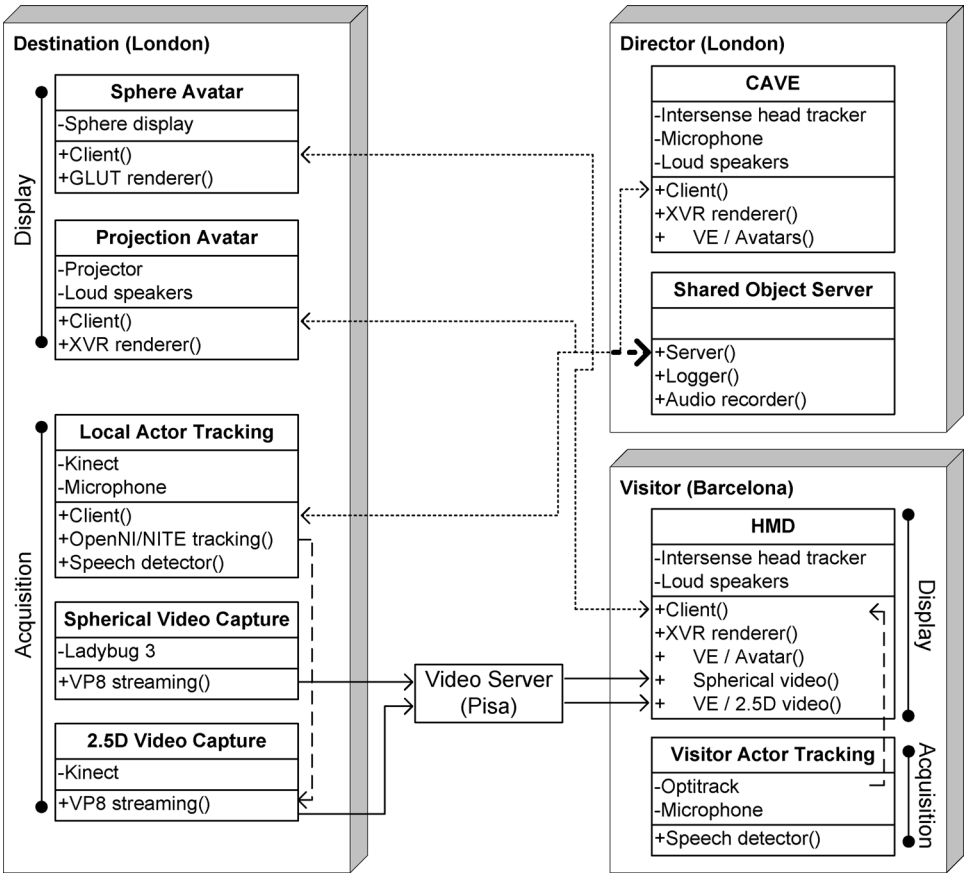

Figure 3. *Simplified system architecture over the three sites. The inner boxes within each site each represent a machine that is responsible for at least one element of the overall system. Hardware is signified by a preceding "–" symbol, while software processes are signified by a "+". Lines between processes indicate wide-area network transmission: high-bandwidth video streams are shown as solid lines, while low-bandwidth session management and skeletal tracking data are indicated by dotted lines.*

connected to the shared object server, are indicated by dotted lines.

According to Figure 3, the destination site features five machines responsible for: rendering the sphere avatar, rendering the projection avatar, tracking the local actor's body movement, capturing and streaming spherical video, and capturing and streaming 2.5D video. The machines performing the avatar rendering and body tracking are connected via the Internet to the RakNet-based shared object server at the director site in order to send and receive updates to and from other connected clients. The machines responsible for the video streaming do not require access to the shared object database, so instead compress and stream their processed video via VP8 and RakNet to a server in Pisa, which is then relayed to the HMD machine at the visitor site for decompression and display. Note that the 2.5D capture using the Kinect also features a local-area network link to the skeletal tracking process (performed by another Kinect) in order to segment the depth image based on the position of the local actor.

Figure 3 shows the visitor site, consisting of two machines. The machine responsible for acquisition runs an Optitrack motion capture system, which streams data to the display machine's RakNet client process, connected on a local-area network. The display machine updates the network with the visitor actor's tracking data, and also displays the destination environment in

one of three modes as illustrated in Figure 2. The Intersense IS-900 head tracker provides XVR with robust positional data in order for perspective-correct rendering in the HMD.[14] This acquisition-display arrangement is tightly coupled due to the tracking data informing the HMD rendering, and, thus, there is logically only one shared-object client required at the visitor site. In contrast, the local actor tracking that occurs at the destination is independent from the sphere and projection avatar displays, and so three shared-object clients are required at the destination: one to update the destination actor's skeletal tracking, and one each to read and display updates of the visitor actress' avatar on the sphere and projection displays.

Finally, the director site consists of two machines. The CAVE (actually consisting of one master and four slave machines) has a client connected to the shared object server, and displays the destination VE and both actors as avatars. Intersense IS-900 head tracking enables perspective-correct rendering. The shared object server is located (arbitrarily) at the destination site, and provides a range of ports to which incoming clients may connect. The logging process also occurs on the server machine, recording both updates to the shared object database and talk between all participants.

3 Acting Rehearsal

Three experienced theater actors/directors were paid £60 each to take part in the rehearsal, which took place over a period of 4 hr in a single afternoon. Prior to the rehearsal, the actors had, separately and apart, learned the "spider in the bathroom" scene from Woody Allen's 1977 movie *Annie Hall*.[15] The scene begins when Alvy, played by Woody Allen, receives an emergency phone call (actually a false, manufactured crisis) to come to Annie's (played by Diane Keaton) apartment in the middle of the night. He arrives and an hysterical Annie wants to be rescued from a big spider in her bathroom. Initially disgusted ("Don't you have a can of Raid in the house? I told you a thousand times. You should always keep a lot of insect spray. You never know who's gonna crawl over."), Alvy skirts around the issue for 2–3 min; firstly by discussing a rock concert program on Annie's bureau, and then a *National Review* magazine that he finds on her coffee table. An arachnophobe himself, Alvy eventually goes on to thrash around in the bathroom, using Annie's tennis racket as a swatter, in an attempt to kill the spider: "Don't worry!" he calls from the bathroom, amidst the clatter of articles being knocked off from a shelf.

This scene was chosen as it consists of varied spatial and interpersonal interplay between the two characters. Thus, the actors engage in intense talk on varied subjects, spatial action (particularly Alvy's character), and directing attention toward and manipulation of objects in the environment. The scene's duration is 3 min in the original movie. This short length allows for multiple run-throughs over the 4-hr rehearsal period, and encourages the director and actors to experiment with new ideas and methods toward the final performance. Alvy was portrayed by a male actor at the destination site at University College London (UCL), and Annie was played by an actress at the visitor site at University of Barcelona (UB). The male director was located in UCL's CAVE facility, separate from the destination room. Drawing from post-rehearsal discussion with the actors and director, together with a group of three experienced theater artists and academics who were spectators at the two UCL sites, this section discusses the successes and failures of the rehearsals in terms of the central elements of spatiality and embodiment. Figure 4 illustrates the visual experience of the rehearsals over the three sites.

3.1 Spatiality

The common spatial frame of reference experienced by all parties was highly conducive to the nature of theatrical acting and directing. The artists were able to perform blocking, referring to the precise movement and positioning of actors in the space, with relative ease. This was demonstrated through the director issuing both absolute and relative instructions interchange-

14. http://www.intersense.com/pages/20/14
15. http://www.youtube.com/watch?v=XQMjrGnGHDY

Figure 4. *The acting rehearsal in progress at each of the three sites. Left: The destination site at UCL featuring the copresent destination actor, and the visiting actress' representations. Center: The visual stimuli (running in VE/Avatar display mode) of the destination site and actor displayed in the HMD worn by the visiting actress (not pictured) at UB. Right: The director located in UCL's CAVE displaying the destination site and avatars representing the two actors.*

Figure 5. *Screen captures from the rehearsal taken with our virtual replay tool, from left to right: picking up a leaflet from the bookcase, "That's what you got me here for at 3 am in the morning, 'cause there's a spider in the bathroom?" passing a glass of chocolate milk, pleading him to kill the spider "without squishing it," and attempting to swat the spider: "don't worry!".*

ably. For instance, asking Alvy either to pick up the magazine "on the table" or "to Annie's right" were both unambiguous to all parties due to the aligned visual environment. The director was able to issue such blocking instructions on both macro and micro scales, ranging from general positioning and Alvy's point of entrance into the scene, down to the technical aspects of movement on a per-line basis.

The artists considered the asymmetry in allowed range of physical movement between the two actors, based on their status as a local or a visitor, as a limitation. The destination actor was free to move around the entire rehearsal space, and would be observed by the visiting actress and director as doing so. However, the visiting actress' allowed movement was limited, particularly forward and backward, due to her situated representation at the destination. The two displays at the destination differ in terms of how well they accurately represent the position of the visiting actress. The projection avatar display, which covers the whole rear wall of the destination site, is able to represent horizontal and vertical movement well. Depth cues, however, are less easily perceived, and a forward movement performed by the visiting actress results in the projection avatar getting larger, due to being closer to the virtual camera, rather than having physical presence at a location further into the destination room. The situated sphere display only shows an avatar head representation, and so cannot express bodily movement at all. The implications of this differing movement allowance between the two actors resulted in some frustrations for the director, who reacted by issuing more gross blocking instructions to Alvy, while focusing more on instructing Annie's expressive gestures. This situation, however, matched the scene's dynamics, in which Alvy is the more physically active of the two characters. Figure 5 illustrates some key moments during the scene, captured with our virtual replay tool.

The solution to this issue of the visitor being spatially restricted at the destination is the use of mobile displays.

The use of personal telepresence robots is likely to solve one set of issues, to the detriment of others. On one hand, they would provide a physical entity with which the destination actor could "play distance," and focus actions and emotions toward. However, the predominant design of such devices is a face-only LCD display, recorded from webcam video. Thus, the fuller body language and gestural ability provided by the avatar representation of the visitor would be missing, which is a critical cue used throughout the processes of acting and direction. In addition, if we are to grant an immersive experience to the visitor, then the ability to capture unobscured facial video is problematic in both HMD- and CAVE-based (Gross et al., 2003) VR systems due to worn devices. This is a rich area for future work, and we are developing more sophisticated articulated teleoperated robots such as those detailed in Peer, Pongrac, and Buss (2010). Finally, it should be noted that, particularly in film and television work, actors are highly skilled at working with props: either when there is nothing to play against as in green-screen work, or inanimate objects. Hence, the notion of an actor imbuing their imagination onto an empty space or object is a natural part of their process. This is a different situation from the practice of general interpersonal interaction, in which nonverbal cues provide information regarding the beliefs, desires, and intentions of an individual, and also provide indicators regarding various psychological states, including cognitive, emotional, physical, intentional, attentional, perceptual, interactional, and social (Duck, 1986).

Some general observations on the benefit of the common frame of reference were also made. For instance, our senior guest academic, the Artistic Director of the Royal Academy of Dramatic Art, discussed the way that many actors are able to learn their lines more quickly by physically being in the rehearsal room or theatrical set as opposed to being in a neutral location such as their own home. In particular, some older actors can only learn lines once they have established the blocking of a scene. Hence, the interactive and visual nature of the system was considered highly beneficial to the process of learning lines and planning movements, even in a solo rehearsal setting.

3.2 Embodiment

The interactions were significantly influenced by modes of embodiment and display at each site. Firstly, it should be noted that throughout the rehearsal period, no critical failures in communication occurred. While we have not formally measured the end-to-end latency of all modes of capture, transmission, and display, this suggests that it is acceptable to support both the verbal and nonverbal triadic interaction. The initial period of acclimatization to the interaction paradigm resulted in some confusion between the three participants due to the evident asymmetry between them. Each party was unclear about the nature of the visual stimuli the others were perceiving. Once some initial descriptions were provided by each party (the destination actor only needed to provide minimal information as he was physically present in the place where the others were virtually present), the group became confident about the unified space they were all perceptually sharing, together with their displayed embodiments.

The initial period of the rehearsal was used to determine each participant's local display preferences. At the destination site, the projection avatar was preferred over the sphere avatar or a dual-representation of the visiting actress. The destination actor considered the projection avatar to provide more useful information through the display of full-body language as opposed to the attentional cues that the head-only sphere display provided. Simultaneous use of the projection and sphere avatars was disliked as it resulted in confusion due to division of attention between two locations. The projection avatar bestowed Annie with a higher degree of physical presence for Alvy to play against and observe (which enhanced the physicality of the performance from Alvy's perspective). During the post-rehearsal discussion, Alvy recalled his excitement when Annie took a step toward him, and an impression of their close proximity was provided by the depth-cue of Annie's avatar increasing in size on the projection display: "When you do go close to the screen; when there are situations where you're flirting, when she's supposed to touch my chest and so on, that is really interesting because she's in Barcelona and I'm here, but there's still some part of you that tries to reach out and touch her hand on the screen. And

when she reacts; for instance when I start smacking the floor looking for that spider, she automatically did that [gestures to cover his head] sort of thing. There was interplay between us; a natural reaction to what I was doing. That was exciting and when the project shined the most, in my eyes."

The visiting actress wearing the HMD decided to observe the destination using the spherical video mode as captured by the Ladybug 3. This mode preserved the actual appearance of both the destination and Alvy, with the trade-off being a decreased perception of depth due to the monoscopic video. This mode was preferred over both the VE/Avatar and VE/2.5D video display modes, due to the improved dynamism of the video compared to the "stiff" avatar embodiment that did not feature emotional facial expression, and the clearer image of Alvy due to the higher resolution camera.

Central to the system's asymmetry are the physical abilities of the two actors depending on at which site they are located. There are several moments during the scene when the actors are required to interact with each other and their environment. This includes knocking on a door; looking at, picking up, and passing objects; and hitting an imaginary spider. When performing such actions during the rehearsal, Alvy has a tangible sense of doing so due to the physicality of his local environment. So, when he knocks on the door or picks up a magazine, he is doing just that, and these actions (and sounds) are observed by both Annie and the director, albeit in varying visual forms. However, this ability does not extend to Annie, as, regardless of display mode, the visitor is only able to mime interaction with perceived objects that are, in reality, located at the destination. Fortunately, most of these moments in the scene belong to Alvy rather than Annie, so this issue did not result in critical failures. However, we identify the support for shared objects as an area for future investigation.

The director in the CAVE viewed the rehearsals as a VE populated with the two actors' virtual avatars. Due to Annie's actual appearance not being captured by video cameras, an avatar is her only available mode of representation at the other sites. While both video and avatar representations of Alvy are available in the CAVE, the director preferred visual consistency, and preferred the avatar representations in the VE. The VE was also considered to provide an excellent spatial representation of the destination that was free of the details and clutter that existed in the real-world location, thus providing a cleaner space in which to conduct the rehearsal. The director considered his ability to freely move and observe the actors within the rehearsal space as a powerful feature of the system. He was able to observe the scene from any viewpoint, which allowed him to move up close to the actors to instruct the expressive dynamics of their relationship, or stand back and observe their positions in the scene as a whole. The fact that the actors were represented as life-sized avatars aided direction by enhancing the interpersonal realism of the rehearsal. Both actors noted our decision to not visually represent the director. Although the benefit of the director's unobstructive movements was universally acknowledged, the inability of the director to use nonverbal gesture, particularly pointing, was considered a hindrance to the rehearsal process. Allowing the director to make his representation visible or invisible to the actors is a potentially interesting avenue of investigation that may have implications for general telecommunication in such systems.

The overall impression of the abilities afforded by the actors' embodiments over the three sites was that movement and general intent was communicated well, but details of expressive behavior were lacking. Facial expression, gaze, and finger movement were highlighted as the key missing features. (Our system is able to track, transmit, and represent gaze and finger movement with high fidelity, and some facial expression is supported; however, these cues require participants to wear encumbering devices, and so were decided against for this rehearsal application.) As a result, moments in the scene that have intended emotional prescience, such as those featuring flirting, fear, and touch, appeared flat. In an attempt to counter these limitations of expressive ability, the actors noted that their natural (and at times subconscious) reaction was to over-act in order to elicit a response from their partner. Correspondingly, the director found himself requesting the actors to perform exaggerated gestures and movements that he would not have done if the finer facets of facial expression were available.

3.3 Discussion

Depending on the characteristics of the play or production, the artists speculated that rehearsing using the multimodal MR system could reduce the subsequent required collocated rehearsal time by up to 25%. The primary benefit to the rehearsal would be blocking the scene, planning actors' major bodily gestures, and, in the case of television and film work, planning camera shots and movement. In television and film work, the artists noted that rehearsal is often minimal or nonexistent due to time and travel constraints. The system provides a potentially cheaper and less time-consuming mode of being able to rehearse with remote colleagues. This benefit would likely extend to technical operators and set designers, who would be able to familiarize themselves with the space in order to identify locations for technical equipment, and optimize lighting and prop-placement. The heterogeneity and multimodal nature of the system was also suggested as a novel paradigm for live performance in its own right, including the potential for art and science exhibitions, and even reality television.

Blocking and spatial dimensions are paramount to a theatrical scene, and determining these aspects is frequently divisive between international performers. Such disputes may be reduced or settled early by allowing all parties to virtually observe and experience the rehearsal or set layout prior to a collocated performance. Both actors and the director advocated the system as a means of overcoming the initial apprehension and nervousness of working with one another, and suggested that they would be more immediately comfortable when the time came for a subsequent collocated meeting. Solo performance and reviewing prior run-throughs was discussed as a potentially useful mode of system operation. To this end, the virtual replay abilities of the system allow for random-access and time-dilation of previous sessions. We are currently implementing an autonomous proxy agent representing an absent human actor that is able to respond in real time to the tracked nonverbal and verbal cues. This may prove useful for solo rehearsal, presenting a responsive humanoid entity to play against.

The relative inexpressivity of the actors' embodiments (due either to the stiff avatar representations or the suboptimal resolution of the video streams) implies that scenes relying on performing and reacting to consequential facial expression and subtle gesture would not benefit significantly from rehearing via the system in its current form. To this end, opera rehearsal was suggested by the observers as a potentially compatible domain. Opera has a number of characteristics that suggest its compatibility with conducting rehearsals via the system. Opera performance relies heavily on gross gesticulation, and while facial expression and minute movements are certainly important, they are perhaps less salient to the overall performance than they are in dramatic theatrical and filmed work. This is illustrated further by the reduced amount of interplay between opera performers, as they are often projecting to an audience. A key strength of the system is its ability for remote participants to move within and observe a perceptually unified space. Therefore, operatic performers, who, due to tight international schedules, can often dedicate only minimal collocated rehearsal time, may find the system useful for familiarizing themselves with the stage space and planning action. Additionally, the progression and timing of an operatic performance is dictated by the musical score, thus leaving minimal room for improvisation. This strictly-structured performance could serve to offset some of the expressive failings of the current system, as behavioral options at a given point in time are limited, thereby potentially reducing instructional complexity.

4 Conclusions

This paper has presented our experience to date of setting up a distributed multimodal MR system to support remote acting rehearsals between two remote actors and a director. Using a range of capture and display devices, our heterogeneous architecture connects the people located in three technologically distinct locations via a spatial audiovisual telecommunications medium based on the destination-visitor paradigm. The implications and characteristics of the paradigm have been explored; the paradigm is emerging due to the geographical distribution of teams at technologi-

cally asymmetric sites. Aiming to support distributed rehearsal between two actors and a director, the technological setup at each site has been detailed, focusing on acquisition, display, and data transmission. The rehearsals were then covered, exploring the successes and failures of the system in terms of the central aspects of spatiality and embodiment. Overall, the common spatial frame of reference presented by the system to all parties was highly conducive to theatrical acting and directing, allowing blocking, gross gesture, and unambiguous instruction to be issued. The relative inexpressivity of the actors' embodiments was identified as the central limitation of the telecommunication, meaning that moments relying on performing and reacting to consequential facial expression and subtle gesture were less successful. We have highlighted spatialized audio, haptics, mobile embodiments, surround acquisition of depth images, facial expression capture, and supporting shared objects as key areas for future work toward advanced iterations of this heterogeneous multimodal form of MR telecommunication.

Acknowledgments

We would like to thank our actors, Jannik Kuczynski and Jasmina Zuazaga; our director, Morgan Rhys; and the Artistic Director of the Royal Academy of Dramatic Art, Edward Kemp, and his colleagues. This work is part of the BEAMING project supported by the European Commission under the EU FP7 ICT Work Programme.

References

Axelsson, A., Abelin, Å., Heldal, I., Nilsson, A., Schroeder, R., & Widestrom, J. (1999). Collaboration and communication in multi-user virtual environments: A comparison of desktop and immersive virtual reality systems for molecular visualization. In *Proceedings of the Sixth UKVRSIG Conference,* 107–117.

Benford, S., Greenhalgh, C., Reynard, G., Brown, C., & Koleva, B. (1998). Understanding and constructing shared spaces with mixed-reality boundaries. *ACM Transactions on Computer–Human Interaction (TOCHI)*, *5*(3), 185–223.

Duck, S. (1986). *Human relationships: An introduction to social psychology*. Thousand Oaks, CA: Sage.

Gillies, M., & Spanlang, B. (2010). Comparing and evaluating real time character engines for virtual environments. *Presence: Teleoperators and Virtual Environments*, *19*(2), 95–117.

Gross, M., Wurmlin, S., Naef, M., Lamboray, E., Spagno, C., Kunz, A., et al. (2003). Blue-C: A spatially immersive display and 3D video portal for telepresence. *ACM Transactions on Graphics*, *22*(3), 819–827.

Maimone, A., Bidwell, J., Peng, K., & Fuchs, H. (2012). Enhanced personal autostereoscopic telepresence system using commodity depth cameras. *Computers & Graphics*, *36*(7), 791–807.

Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems E, Series D*, *77*, 1321.

Mohler, B., Creem-Regehr, S., Thompson, W., & Bulthoff, H. (2010). The effect of viewing a self-avatar on distance judgments in an HMD-based virtual environment. *Presence: Teleoperators and Virtual Environments*, *19*(3), 230–242.

Normand, J., Spanlang, B., Tecchia, F., Carrozzino, M., Swapp, D., & Slater, M. (2012). Full body acting rehearsal in a networked virtual environment—A case study. *Presence: Teleoperators and Virtual Environments*, *21*(2), 229–243.

Oyekoya, O., Steptoe, W., & Steed, A. (2012). Sphereavatar: A situated display to represent a remote collaborator. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.*

Pece, F., Kautz, J., & Weyrich, T. (2011). Adapting standard video codecs for depth streaming. In *Proceedings of the Joint Virtual Reality Conference of EuroVR (JVRC),* 59–66.

Peer, A., Pongrac, H., & Buss, M. (2010). Influence of varied human movement control on task performance and feelings of telepresence. *Presence: Teleoperators and Virtual Environments*, *19*(5), 463–481.

Schroeder, R., Steed, A., Axelsson, A., Heldal, I., Abelin, Å., Wideström, J., et al. (2001). Collaborating in networked immersive spaces: As good as being there together? *Computers & Graphics*, *25*(5), 781–788.

Slater, M., Howell, J., Steed, A., Pertaub, D., & Garau, M. (2000). Acting in virtual reality. In *Proceedings of the Third International Conference on Collaborative Virtual Environments*, 103–110.

Slater, M., Sadagic, A., Usoh, M., & Schroeder, R. (2000). Small-group behavior in a virtual and real environment: A comparative study. *Presence: Teleoperators and Virtual Environments*, *9*(1), 37–51.

Tecchia, F., Carrozzino, M., Bacinelli, S., Rossi, F., Vercelli, D., Marino, G., et al. (2010). A flexible framework for wide-spectrum VR development. *Presence: Teleoperators and Virtual Environments*, *19*(4), 302–312.

Tsui, K., Desai, M., Yanco, H., & Uhlik, C. (2011). Exploring use cases for telepresence robots. In *Proceedings of the 6th International Conference on Human–Robot Interaction,* 11–18.

Venolia, G., Tang, J., Cervantes, R., Bly, S., Robertson, G., Lee, B., et al. (2010). Embodied social proxy: Mediating interpersonal connection in hub-and-satellite teams. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems,* 1049–1058.

Voida, A., Voida, S., Greenberg, S., & He, H. (2008). Asymmetry in media spaces. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work,* 313–322.