withyou—An Experimental End-to-End Telepresence System Using Video-Based Reconstruction

David J. Roberts, Allen J. Fairchild, Simon P. Campion, John O'Hare, Carl M. Moore, Rob Aspin, Tobias Duckworth, Paolo Gasparello, and Franco Tecchia

***Abstract*—Supporting a wide set of linked non-verbal resources remains an evergreen challenge for communication technology, limiting effectiveness in many applications. Interpersonal distance, gaze, posture and facial expression, are interpreted together to manage and add meaning to most conversations. Yet today's technologies favor some above others. This induces confusion in conversations, and is believed to limit both feelings of togetherness and trust, and growth of empathy and rapport. Solving this problem will allow technologies to support most rather than a few interactional scenarios. It is likely to benefit teamwork and team cohesion, distributed decision-making and health and wellbeing applications such as tele-therapy, tele-consultation, and isolation. We introduce *withyou*, our telepresence research platform. This paper describes the end-to-end system including the psychology of human interaction and how this drives requirements throughout the design and implementation. Our technology approach is to combine the winning characteristics of video conferencing and immersive collaborative virtual environments. This is to allow, for example, people walking past each other to exchange a glance and smile. A systematic explanation of the theory brings together the linked nature of non-verbal communication and how it is influenced by technology. This leads to functional requirements for telepresence, in terms of the balance of visual, spatial and temporal qualities. The first end-to-end description of *withyou* describes all major processes and the display and capture environment. An unprecedented characterization of our approach is given in terms of the above qualities and what influences them. This leads to non-functional requirements in terms of number and place of cameras and the avoidance of resultant bottlenecks. Proposals are given for improved distribution of processes across networks, computers, and multi-core CPU and GPU. Simple conservative estimation shows that both approaches should meet our requirements. One is implemented and shown to meet minimum and come close to desirable requirements.**

***Index Terms*—Computer supported cooperative working, computer vision, virtual reality.**

Manuscript received April 18, 2014; revised September 30, 2014; accepted January 06, 2015. Date of publication February 11, 2015; date of current version March 18, 2015. This work was supported in part by the Engineering and Physical Research Council (EPSRC) under Grant EP/E007570/1 and three Ph.D. studentships, in part by the EU Framework Program under Grant 607177, in part by VISIONAIR under TNA-131, and in part by the Higher Education Funding Council of England (HEFCE), SRIF II & III. Supporting research material is available from the lead author on researchgate. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Patrick Le Callet.

D. J. Roberts, A. J. Fairchild, S. P. Campion, J. O'Hare, C. M. Moore, R. Aspin, and T. Duckworth are with the University of Salford, Salford, M5 4WT, U.K. (e-mail: d.j.roberts@salford.ac.uk; a.j.fairchild@salford.ac.uk; s.p.campion@salford.ac.uk; j.o'hare@salford.ac.uk; c.m.moore@salford.ac.uk; r.aspin@salford.ac.uk; t.duckworth@salford.ac.uk).

P. Gasparello and F. Tecchia are with the Scuola Superiore Sant'Anna, 33 Pisa PI, Italy (e-mail: p.gasparello@sssup.it; f.tecchia@sssup.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2015.2402635

I. Introduction

CONVERSATIONAL behavior as well as understanding and perception of another's actions and words, are all guided by interpretation of non-verbal behavior. Meaning is derived from sets of audible and visual cues, linked spatially and temporally. Cues often make sense only within both spatial and temporal contexts. Interactional scenarios differ in the subsets of non-verbal communication they rely on.

Today's visual communication technologies favor either visual or spatial qualities and induce temporal disturbance. This has the potential to change the likelihood, flow and meaning of a conversation, how one person perceives the other, and over time, the relationship between them. Combining a display and video-based medium that are both free-viewpoint, offers the potential to balance visual and spatial qualities. However, the computational processing overhead and data transfer is far heavier than with other mediums. Balancing visual, temporal and spatial qualities is thus challenging. Understanding and meeting this challenge is the subject of this article.

II. Background

This section sets the evolving scene for our research in terms of context and our own research that brought us to, and initially developed, our current approach.

A. Context

Our work evolved from a time when the first prototypes of tele-immersion where released [1], [2]. These were already creating 3D computer graphic avatars from live video data, using techniques such as [3]. However, visual and temporal qualities fell well short of what was needed to meaningfully support or study non-verbal interaction.

We were part of another strand of research, bringing pre-authored avatars to life, through movement that reflected gross user interactions with objects and others within a virtual space in which all can move around. This approach was better suited to technology limitations of the time, most notably bandwidth of networks and computation, allowing meaningful behavioral studies to be started. Avatars, initially controlled through a desktop interface, later followed motion tracked user movement. They could move up to, point at, or reach to objects and others. Motion tracking is an integral part of traditional immersive displays, needed to support parallax and useful for natural interaction with objects. Exploiting this to drive an avatar [4] meant that what a person was looking at or interacting with was

faithfully communicated. This was importantly different to the desktop interface where the avatar usually reflected behavior that the user wanted to communicate.

Others combined video based reconstruction with an immersive display [5] demonstrating how spatial and visual qualities could be better balanced. However, visual and temporal qualities where still some way behind what could be achieved with motion tracked avatars. Since then, visual qualities of video based reconstruction have significantly improved [6], [7]. A well-established approach derives shape from silhouette [8], [9]. A good implementation of this is the Exact Polyhedral Visual Hull (EPVH) [10]. The algorithm was implemented in a commercial telepresence system [11]. However, Head Mounted Displays were used, obscuring the view of the face.

The emergence of commodity depth-based cameras (Kinect), using structured infrared light, has introduced new energy into the field of telepresence. Such systems are easier to deploy as background segmentation is taken from depth rather than needing more complex image processing.

Recent [12] and current [13], [14] funded EU research focuses on spatial telepresence. Embodied telepresence using robots is in vogue [13], [15] but less mature.

B. Our Journey to Immersive Video Conferencing

Shared manipulation of virtual objects can be achieved across the network by combining concepts of causal and natural time within consistency control [16]. By adding consistency management to an asymmetrically immersive collaborative virtual environment, we were the first to allow an immersed user to hold, move and fix an object also held and moved by another [17]. In this early work, the other used a desktop display.

Intent and actions of those in more immersive displays was easier to understand. Fully immersed people were perceived as contributing and collaborating more effectively. During such collaboration, non-verbal communication differed greatly between stages, such as planning, moving things, fixing them, and assessing. During each stage, one but not always the same person would do most of the movement. This meant network traffic was usually asymmetrical [18].

Later work connected a variety of immersive displays including single and split screen workbenches, power wall, panorama and up to three cubic. Fig. 1 shows how spatial context is joined using cubic displays. The degree to which remote spaces can be aligned through a 3D medium depends on the dimensionality of the display [19]. Remote spaces can be aligned using "look through" displays, like a power wall, to be adjacent, and through "walk within" displays, like a cubic "CAVE," to coincide. Fragmentation of conversation and workflow are proportional to field of view and was removed by using "walk within" displays [20]. Task performance and feelings associated with collaboration and creativity were proportional to level of immersion [21].

We integrated eye tracking into stereo glasses in order to drive the eyes of remote avatars [22]. We then connected two [23] and three [50] cubic displays to respectively perform perceptual and behavioral experiments. However, while these avatars mimicked the movement of remote users, they usually neither looked like them nor reflected any changes in facial expression [23].

Fig. 1. Cubic immersive display technology used to create a shared spatial context between distal people.

This left us with the following conclusion. While appearance could be faithfully transmitted through video conferencing, and attention through immersive collaborative environments, combining free viewpoint video and display could potentially do both. While such prototypes had been demonstrated, they had not possessed sufficient combined visual, spatial and temporal qualities.

C. From There to Now

We had experience of free-viewpoint displays and computer graphic mediums but not free-viewpoint video. For the latter we decided to adopt the most suitable approach we could find and in parallel build our own. To do both, we chose shape from silhouette as it created a full 3D form that refined with number of cameras. Creating a full 3D form from a set of depth maps has, perhaps, more potential to induce errors.

Survey of the literature revealed a lot more work on performance of reconstruction algorithm and its distribution rather than on that of the distributed system that surrounded it. Reported visual, spatial and temporal qualities of EPVH seemed well balanced. However while frame rate was reported, latency was not. The distribution of the algorithm used to increase frame rate, looked ill suited to maintain low latency. Temporal performance of the acquisition, which includes various imaging processing and transfer across the network, had been largely overlooked. We therefore concentrated on implementing reconstruction solutions that ran on a single commodity computer and on simplifying the distributed system around it to reduce latency.

We demonstrated that neither hardware synchronization of camera captures [25] or the delivery of frames [26], was strictly necessary at real-time frame rates. We increased the parallelization of the EPVH algorithm, thus removing overheads allow it to run in real time on a laptop [27]. We developed our own render-based approach, which ran almost entirely within a single graphics card [28]. We also built simulation tools to allow camera placement to be examined [29]. We demonstrated that eye-gaze could be reliably estimated from our EPVH approach [30]. We are now developing a telepresence system that uses both conventional motion tracked and our EPVH avatars to support distributed space science and operations [14]. Fig. 2 shows our first demonstrator of this application in both "walk within" and "look though" displays within the spatial context of a Mars landing site, and both types of avatars. The difference between the two avatars is that while both can faithfully convey attention, only the latter faithfully conveys

Fig. 2. Telepresence space science and operations demonstrator (right) within a "walk within" display (left) conventional and virtuality (EPVH) avatar seen through a "look through" display.

appearance, doing so through the medium of free viewpoint video.

While we had stated key elements of our theory, we had never explained it as a whole. We had built, and in isolation tested, every component of the end-to-end system. However, these components competed for resources when ran on the same computer and none of our implementations had included all of the components. This paper sets out to address these shortcomings.

III. Grounding Theory—Non-Verbal Communication and How it is Influenced by Technology

The purpose of this section is to provide a systematic explanation of the theory that underpins our work; and from this derive functional requirements. It describes the importance of the relationships between various non-verbal cues and how this relationship is impacted by technology.

A. Natural World

Non-verbal behavior is largely subconscious [31], therefore potentially telling more about feelings than someone might wish to let on. It is hard to manage a conversation or assess or build trust and rapport without it. Non-verbal communication has been described as the transmission of information and influence by an individual's physical and behavioral cues [32]. This transmission is usually via a range of cues that often only retain correct meaning when interpreted together and within context.

There are many situations where knowing both what someone feels, and what or who the feeling is about, is important. Telling either, let alone both, is challenging without interpreting together a variety of linked cues. Interpretation of feeling can be made through observation of visual cues. However, understanding the source of that feeling requires spatial and temporal context.

Micro expressions, are in built human behaviors covering all cultures [33]. Humans cannot control them and they happen in a split second, leaking information about feelings. The ability to accurately represent the muscle groups of the face is significant, as a smile does not only affect the mouth; it also engages muscle groups near the eyes [34].

Eye-gaze is probably the most studied form of non-verbal communication [30]. Yet gaze itself is only one component of non-verbal behavior. A relationship exists between mutual gaze, interpersonal distance and affinity [35]. Equilibrium theory expands this to include the role of a non-verbal cue signaling intimacy [36]. A signal of intimacy might be the orientation of the body to the other, known as body torque. Or it could be a dip of the head, an opening of the mouth or the raise of an eyebrow.

The use of interpersonal space differs depending on familiarity and culture. Identified zones define where intimate, personal, social and public interaction typically takes place [37]. Social space is generally used for acquaintances and extends between 1.2 and 3.5 meters.

Some interactional scenarios rely far more on spatial context than others. For example, an organized meeting relies less on spatial context than an ad-hoc meeting. This is because people have already agreed to take part and thus pay less attention to each other's willingness to do so. In contrast, the likelihood, duration or outcome of an ad-hoc meeting may depend on mutual changes in body orientation, gaze and facial expression.

Timing in a conversation also conveys meaning and can be used to manage turn taking and outcome. Margret Thatcher, the UK's former prime minister, is said to have often defended a contentious argument in the following way. While speaking she would appear to verbally tail off, then look down, and as an opponent prepared to speak, look them straight in the eye and pick up the tone and momentum of speech.

Technology can reproduce aspects of non-verbal behavior across time or between places. However, we argue that there is not yet a single technology that captures most aspects of non-verbal behavior well. In particular, current approaches favor either visual or spatial aspects. Furthermore, those that bring places together in real time, introduce temporal disturbance.

B. Video

Most people in the developed world watch video footage of others on a daily basis. This suggests both that it is effective, and we do not need describe it in detail. We concentrate instead on its limitations and how these are being to be addressed. Video is sufficient for capturing most non-verbal behavior within the filmed spatial/temporal context. However, when what is being responded to is out of view or delayed, the meaning of the response may be lost.

Free-viewpoint video [6] research attempts to seamlessly combine multiple video streams from different viewpoints within a 3D computer graphics medium. This allows the interpretation of non-verbal communication from any direction. However, context is limited within the cross over of numerous camera views. The process of reconstruction impacts on the faithfulness of reproduction, visually, spatially and temporally. It therefore has the potential to hide or distort non-verbal signals.

C. Virtual Reality

Virtual Reality, in its purer form, immerses people in 3D computer graphics; using displays that maintain parallax as viewpoint is changed. Both the medium and the interface are free viewpoint. When this is the case the user can move within the spatial context while interpreting, for example, what non-verbal communication is directed toward her? This is akin to taking the camera off its tripod and entering the proscenium.

Virtual humans can mimic much non-verbal behavior, importantly retaining its spatial context when viewed through an immersive display. There seems to be something compelling about interacting with a life-sized virtual human through an immersive display [38]. While people can see it is not real, they appear to respond and think about it as if it were.

Virtual Reality is useful for understanding non-verbal behavior, and its relationship to cognition. This is partly because it allows complex and reactive simulated behaviors to be controllable and repeatable. It is also because monitoring of human movement is already incorporated and extending this is simpler than facilitating it in the natural world. Movements of the human user can drive non-verbal actions of virtual humans.

In one experiment [39], virtual humans combined body posture, orientation and gaze to mimic interest or disinterest in each participant. This resulted in behavioral and psychophysiological responses akin to the real world. One participant is rumored to have thrown off her stereo glasses and ran from the room saying "they hate me."

Aspects of equilibrium theory were tested in virtual reality [36]. Distance from a virtual human maintained by others was proportional to the extent of non-verbal cues used. For example, were eyes closed and if open, following the observer.

The range of supported non-verbal cues not only determines how a person responds to another but also how they regard them and what they are saying. Fixing eyes to look directly forward from the avatar's face, increased people's tendency to believe what the avatar was saying [24]. This tendency was already higher than in the natural world.

The above virtual reality studies, along with a significant body of similar literature, tells us that:

- Combining 3D medium and display maintains spatial context of non-verbal behavior
- The loss of some non-verbal cues can change impression of and reaction to another

D. Telepresence

We now consider technologies that attempt to reproduce the face-to-face meeting between distal people through some audio/visual medium. They encompass those where the visual medium is 2D or 3D and either video, computer graphic, or the latter derived from the former.

We argue that each of today's approaches favor the reproduction of either visual or spatial aspects of non-verbal behavior. This reduces the interactional scenarios in which each can be trusted to deliver an experience and outcome similar to a physical face-to-face meeting.

Visual aspects of non-verbal behavior, such as appearance, are faithfully transmitted via video conferencing. These aspects can be simulated in collaborative virtual environments. However, the level of faithfulness is dependent on the approach. Traditionally avatars are constructed, and their appearance tailored to an individual, off line. They are brought to life by following the motion-tracked movements of the person they mimic. However, they usually look nothing like them. Detail, such as facial expression, body torque and finger gestures are seldom conveyed in collaborative virtual environments. This is because numerous markers must be painstakingly placed, which is impractical for everyday meetings.

In contrast, video conferencing falls short of faithfully transmitting many spatial qualities of non-verbal communication. This is because each participant is looking into the other's spatial context from a viewpoint which does not move with them. These spaces can be at best aligned to appear adjacent. This illusion is best achieved when the gaze of the observer on each side is in line with the camera viewpoint into the other space [23].

Perception of gaze has been described for 2D and 3D faces in video conferencing [40] and avatars [41]. Mutual gaze can be theoretically achieved by, on both sides, aligning the image of the remote person, with the camera through which he/she sees. When a person moves from this alignment, two problems coincide to prevent mutual gaze [23]. Firstly, their viewpoint into the remote space does not move with them. This results in the Mona Lisa effect [42] where stationary eyes on a flat image appear to follow an observer.

Secondly, the image of the moving person will diverge from the camera through which they see. The result of the two factors is that a moving person will only feel watched when the other is watching the camera and not them. This implies that the relationship between interpersonal distance and gaze would be hard to support.

Another problem is that even when people remain stationary, it is hard to align a camera within the eyes of the image of the person that looks through it. Putting the camera in front obscures the face and putting it behind the screen degrades the quality of both displayed and captured images.

The compromise is usually to put cameras above the screen. However, the result is to give the impression of the remote person looking at one's body when they are looking at one's eyes. A virtual camera solution for seated participants is presented in [43]. Participants are reconstructed in 3D so that a virtual camera can look at them from the perspective of the image of the remote participant, who then looks through it.

Immersive collaborative virtual environments allow distal people to share the same spatial context, seeing each other as avatars within a shared virtual scene. This allows interpersonal distance and gaze to work together between person and avatar, just as with person and autonomous virtual human. It works not only because of the combination of 3D medium and 3D display but also critically as both viewpoint and avatar follow the user.

Eye tracking has been incorporated to drive the eyes of the avatar thus supporting mutual eye-gaze [22], [23]. This also allows pupil dilation to be communicated [24]. However, faithfulness of spatial representation is at the cost of the visual aspects as described above.

What we call Immersive Virtuality Telepresence, balances visual and spatial qualities of non-verbal communication. The medium is 3D computer graphics, created in real time from video. In many such approaches the viewpoint moves with the observer. In some, one participant can walk around the other, while still seeing a plausible reconstruction. It is those that can do all three which best support spatial qualities of non-verbal communication. Such a system could be said to combine the winning qualities of video conferencing and immersive collaborative virtual environments. The goal of research such as our

own, is to achieve visual, spatial and temporal qualities sufficient to support everyday conversations without changing flow or outcome.

Perhaps the most rigorous such approach in terms of supporting eye gaze is presented in [43]. Unlike shape from silhouette and low-resolution depth based approaches, such as Kinect, this accurately captures the shape of the eye socket. Studies that have provided rigorous empirical evidence of gaze perception include [30], [44]. Interestingly, approaches that do not capture the shape of the eye socket have been shown to be sufficient to determine when being looked at, at typical social distances [30]. This is perhaps as the shape is implied by the texture and it is courser facial form, such as the nose, that play a greater role.

Such empirical evidence of ability to support estimation of gaze [30], [44] strongly indicates that the technology is coming of age. Such systems can be said to be the first in which both appearance and attention are faithfully communicated. Accurate Estimation of gaze, when eyes centered, was demonstrated with the depth-based approach [44]. We demonstrated similar accuracies when eyes centered or turned from shape from silhouette [30]. The latter is known to be significantly harder in the natural world [30].

Communicating through technology will induce perceivable delays over today's Internet. These delays, and their integral jitter, can distort the context of verbal and non-verbal cues, confusing their meaning and the conversation. For example, would Margret Thatcher have been so good at stopping others from talking if she did not hear their intake of breath before what they said after it had began transmission? A detailed breakdown of component performance within what we would call a virtuality (video based reconstruction) tellepresence system is given in [45].

Typically visual information is slower to communicate than audio because of the greater bandwidth required. If the visual channel is delayed for longer than a conversational pause, its influence might disrupt the conversation. The duration of conversational pauses are dependent on many factors. A study across three languages reported a median of medians of 111 ms of which 40% are long enough for someone to react to [46]. Frame rate can also impact. Many micro expressions, such as a blink, can be over in tenth of a second. Thus at below 10 Hz they might be lost.

E. Functional Requirements

The challenge is to balance visual, spatial and temporal qualities of non-verbal communication, so that all fit within the tolerances of meaning. From the above summary of non-verbal communication in the natural world, we set the following requirements:

- Eye-gaze, body-torque and micro-expression of the face must be observable up to 3.5 m between two people potentially moving:
  - — Minimal: past each other
  - — Desirable: around each other
- Latency
  - — Minimal: matching Skype
  - — Desirable: $<$ ITC recommendation for audio [48] $+$ duration of communicational pause $= 400$ ms
- Frame rate
  - — Minimal: sufficient to capture most micro-expressions: 10 Hz
  - — Desirable: matching commercial video-conferencing: 20 Hz

It is important to note that aligning spaces to appear to coincide, today requires the wearing of stereo glasses, through which eyes are hard to discern. We thus set a minimum requirement to avoid this issue by aligning spaces to appear adjacent. A display wall that neither can cross will physically separate each participant from the representation of the other.

IV. End-to-End Description of Our System

The purpose of this section is to provide the first description of our *withyou* telepresence system as a whole. *Withyou* is now described in terms of processes for reconstruction and data acquisition, and the physical display and capture space.

A. Processes

We use two different approaches to shape from silhouette reconstruction: model (polyhedral); and render (voxel) based. The model-based approach [27] splits the reconstruction and rendering into respective processes on machines at respective sites. The render-based approach [28] reconstructs as part of the render process on the graphics card at the site remote to the cameras that feed it.

1) Reconstruction and Approach Dependent Processes:

Model-based Approach

Reconstruction (*polyhedral*) creates and textures the 3D form. This is based on the EPVH algorithm that we have re-implemented and parallelized.

Rendering, including texturing, of the reconstructed model.

Render-based Approach

Reconstruction and Rendering (*Voxel*) creates a visible form from projecting camera images into the reconstruction space and sampling this without forming a 3D model.

Input Data Management weighted organization of images.

2) Acquisition Processing: Both approaches to reconstruction rely on the same input data, namely synchronized frames from multiple cameras and silhouettes taken from each.

Image conversion is typically required before and after video coding/decoding. In various experiments we converted between the three standards of YUV420, YUV422 & RGB32.

Video encoding is used to compress the video streams before they are sent across the network. The streams are then decompressed at the receiving node. While we have used MPEG-4, H.264 performed preferably.

Lens correction is required to remove distortions from the camera lens. This is more important for wide-angle lenses. Retaining a useful working volume without wide-angle lenses is more challenging.

Segmentation of the background is needed to create the silhouettes. Simple pixel color comparison is used.

Fig. 3. Compares the quality of reproduction from our two approaches to shape from silhouette (left) model (polyhedral) based and (right) render (voxel) based. The INRIA dancer data set is used in both, to allow comparison with other's implementations.

Filtering smoothens silhouette edges, greatly reducing the work of subsequent processes, most notably reconstruction. It was found absolutely necessary in order to allow real time performance of mesh compression down the line. The Douglas-Peucker simplification algorithm was used.

Input data management was necessary for the volumetric reconstruction to combine silhouette and texture images into a single, combined image with the silhouette data encoded within the alpha channel.

Mesh compression was used to compress the polygonal model before sending over the Internet by serializing vertices. Two approaches were tested 3Dzip for transmission over TCP and ffmpeg for UDP.

B. Display and Capture Space

Our *octave* display and capture space is an octagonal immersive projection display with cameras mounted around the side and sometimes in the ceiling (Fig. 2 left). It is just over 5 m wide so that cameras placed above its walls can capture a space of 3.5 m around center. This allows a person to move within the extent of social distance to another in the center, be either of them real or virtual.

The octave is highly reconfigurable. The floor comes up and the walls move, allowing the octagon to transform into two cubic displays that can then act as two ends of a telepresence system.

Within the research described in this paper we use the display in either low (single screen) or full (8 walls and floor) immersion mode. Through connection to a similarly configured display, the former allows remote spaces to be aligned to appear adjacent, whereas the latter gives the appearance of them coinciding. Stereo glasses are required for the latter. Eyes are barely discernible through the current glasses.

V. Characterization of Our Approach

The purpose of this section is to provide unprecedented characterization of our technology approach in terms of visual, spatial and temporal qualities, and what influences them; and from this derive non-functional requirements that direct the work in the next section.

We begin by comparing images from our two approaches to reconstruction, Fig. 3. However, we restrict the detailed characterization to that of the model-based approach. This is because the images needed for our characterization take up a lot of space. Thus we would need to cover half the aspects to show both approaches, while camera placement impacts much more.

Fig. 4. Impact that not modeling concavities has on the eye as seen from the side and straight on. The latter is far more likely during a conversation. Source images from a frontal arc of 8 HD cameras, approximately 3.5 m from subject.

We now present new evidence towards the understanding of the visual and spatial qualities of the video-based reconstruction approach, shape from silhouette. We do this to:

- Validate approach—demonstrating that facial features can be reproduced to a useful quality from within a suitable working volume
- Uncover functional requirements—demonstrating implication of number and placement of cameras

A. Validating Approach

With regard to validating the approach, we set out to provide new evidence to aid balancing the views of those for and against it. In our opinion:

- what is most criticized about the approach might have little practical implication
- however, much of the published evidence to support any approach is captured in highly favorable conditions

We begin with investigating what we consider the core criticism leveled against shape from silhouette. This is that fundamentally it cannot model concavities such as eye sockets. This criticism is undoubtedly true in a geometric sense. However, what an observer perceives comes from a combination of information from the geometry, the texture upon it and the mind. The texture of a face contains a pictorial representation of concavities, including eye sockets. However, this is pasted on one or more flat polygons. The brain combines cues and can often see what it expects rather than what is represented by an isolated cue. A full study of this would compare the impression of a large number of participants. We previously undertook a more focused study that showed eye gaze could be reliably determined from our models [30]. Here we provide additional visual evidence, Fig. 4, and give our own interpretation of it in the discussion, leaving readers to make up their own minds.

We now look at the ability of the approach to reflect microexpressions of the face. This is done through comparison with a

Fig. 5. Compares footage from video conferencing (Skype) with that from our telepresence system. The former has far less errors and thus gives a more humanlike impression. Two errors from camera calibration are evident in the latter. Firstly, one camera has been jogged, resulting in a slicing across the top of the head and the rising of one shoulder. Secondly the skin is red. Both images demonstrate similar micro-expressions, including tilt of the head and raise of eyebrows. However, those of the lower face, including chin raise, seem less pronounced in the reconstructed avatar.

widely used video conferencing tool, Skype, Fig. 5. Whereas the above used favorable conditions, the following does not. The factors impacting on favorability are discussed later. For those conversant with the Facial Action Coding System [47] we have attempted to code the pair of images in Fig. 4. We found it easier to code the image from the video conferencing, which we believe to have a Facial Action Coding score of 1D, 2C, 17C, 55B. This means that we consider both inner and outer eye browse to be raised, with inner raised more than outer, head to be tilted and chin raised. With the virtuality avatar, we were confident of the first two (1D & 2C). However, we differed in opinion of clarity of head tilt and agreed that clarity of chin raise was poor.

B. Influences on Visual and Spatial Quality

We now look at the impact of number, placement and calibration of cameras. This is relevant as firstly it impacts greatly on quality and secondly gives an idea of the number and resolution of cameras needed to capture from throughout our required volume. The requirements for this volume come from the extent of social gaze, around 3.5 m. We want a person to be able to move up to another from this distance, preferably from any direction.

We begin by looking at the impact of the extent to which cameras surround a person on the horizontal plane. Consideration of the vertical is given later. Shape from silhouette constructs a full 3D form, viewable from any angle. However, as we now demonstrate, the extent to which cameras surround the subject, impacts on its correctness. This has been reported before. However, what we concentrate on here is how the form might look correct from one viewpoint and very wrong from another.

Combining a free viewpoint medium, such as shape from silhouette, with a free viewpoint display, allows one participant to walk around the other, inspecting them from all sides. Examples of such displays are immersive projection technology or head mounted displays.

However, the placement of cameras determines the correctness of the form as seen from different viewpoints. In order to join remote spaces as if adjacent, it would only be necessary for the reconstruction to look correct from within a frontal arc. Therefore, we now look at the impact of breadth of a frontal arc of cameras and the number of cameras. Fig. 6 demonstrates that while a reconstruction can look correct when viewed from within the capture arc, this can hide a hideous deformity viewable from outside the arc. We conclude from this that a frontal capture arc, as wide as the viewing screen, is sufficient to reproduce a human form that looks correct as the observer walks past the screen. However, the form itself is incorrect unless the capture arc reaches 180 degrees. While this might impact on interpretation of eye-gaze, our previous tests [30] suggest that it does not. The number of cameras within the arc impacts subtly on the form from within the arc.

A reconstruction viewed from all sides is shown in Fig. 7. Such would allow two remote spaces to be apparently aligned so that they coincide. For this we have used eight cameras, positioned approximately symmetrically around the entire circle.

We next check the correctness of the reconstruction and in particular, the placement of texture upon this. To do this we use one of the captured images simply for comparison, not using it to reconstruct or texture the model.

We had previously reported [30] an apparent droop in the face that appears when texturing camera are above head height. However, our evidence had focused on the effect on placement of the eyes. A far greater effect is noticeable on the lower face. We now examine the effect across the face, Fig. 8. In our opinion, the subject appears to age when geometry is textured from a camera with steep elevation from the face. Such an elevation is likely when cameras are placed above screens large enough to display a life size standing person. A person walking toward such a screen would likely appear to age and loose health and wellbeing. With reference to Fig. 5 right, we suspect that a subtle droop effect, combined with reproduction of the beard and raise of the chin, is making the lower face harder to interpret.

C. Influences on Temporal Quality

We now describe the major influences on temporal quality for both approaches.

The temporal performance of both approaches is related to the number of cameras. Frame rates within our desirable 20 Hz are obtainable with both, using 12 cameras.

The temporal performance of the model-based approach is also relative to the complexity of silhouettes [27]. The complexity of silhouettes is dependent on a person's hair and clothing and how clearly the background can be discerned, dependent in simple techniques on color and lighting. Loosely cropped hair results in higher frame rates than dreadlocks, for example. The extent to which these factors impacts depend greatly on the filtering used to simplify silhouettes. Silhouette complexity also changes as a person moves. For example, there will be more edges when legs or fingers are opened. In our experience this varies frame rate by around 25%.

The temporal performance of the render base approach is also dependent on capture volume [28]. We can achieve interactive frame rates with 8 cameras (graphics hardware limited) when

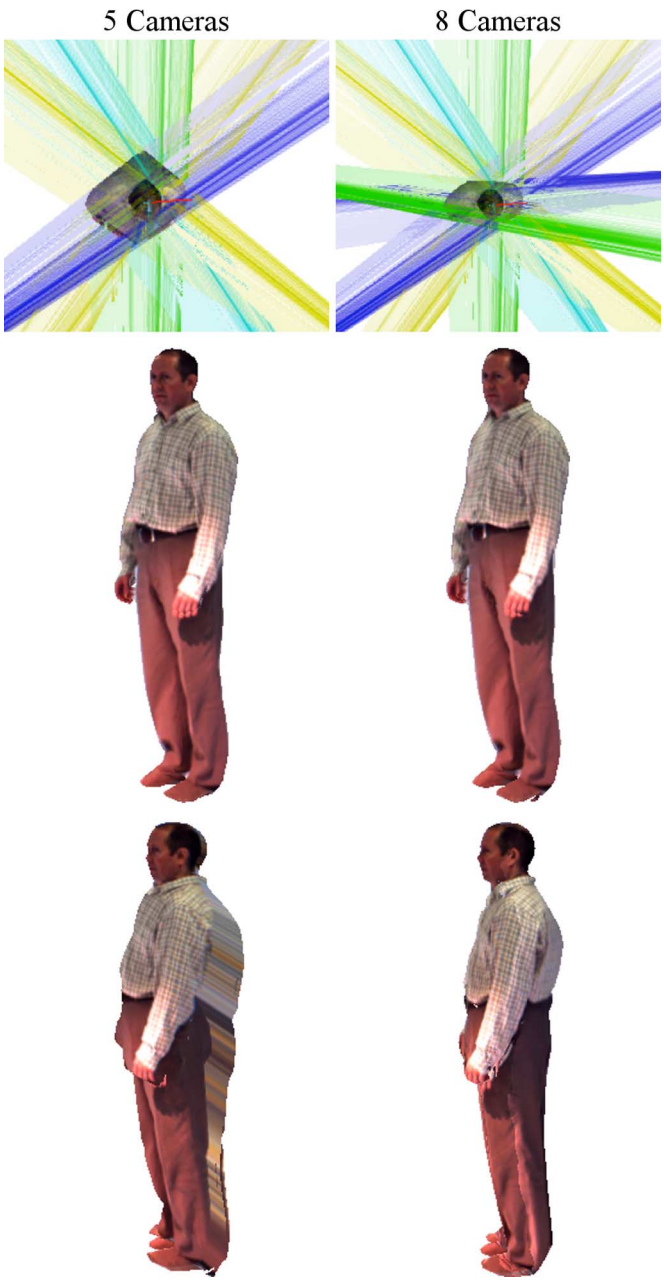


Fig. 6. The impact on reconstruction of size of camera arc. Reconstruction looks correct provided it is viewed from within the capture arc (middle row), this can hide a hideous deformity viewable from outside the arc (bottom left). Close inspection of the images in the middle row, shows that fewer cameras give a subtle effect of a heavier build to the body. The top row shows, from above, viewing lines of the cameras and the form reconstructed with their intersection.

the capture volume is large enough for dancing on the spot with arms outstretched.

Impact of Component Parallelisation Across CPU or GPU: We now present a comparison of process performance when parallelized across CPUs or GPUs. These are detailed in Table I. Results for the reconstruction algorithms have been previously published: model-based [27]; render-based [28]. All others have not. The purpose of the render-based approach was to tune the process of shape from silhouette to execution on a single graphics pipeline. We thus do not test its performance on CPUs.

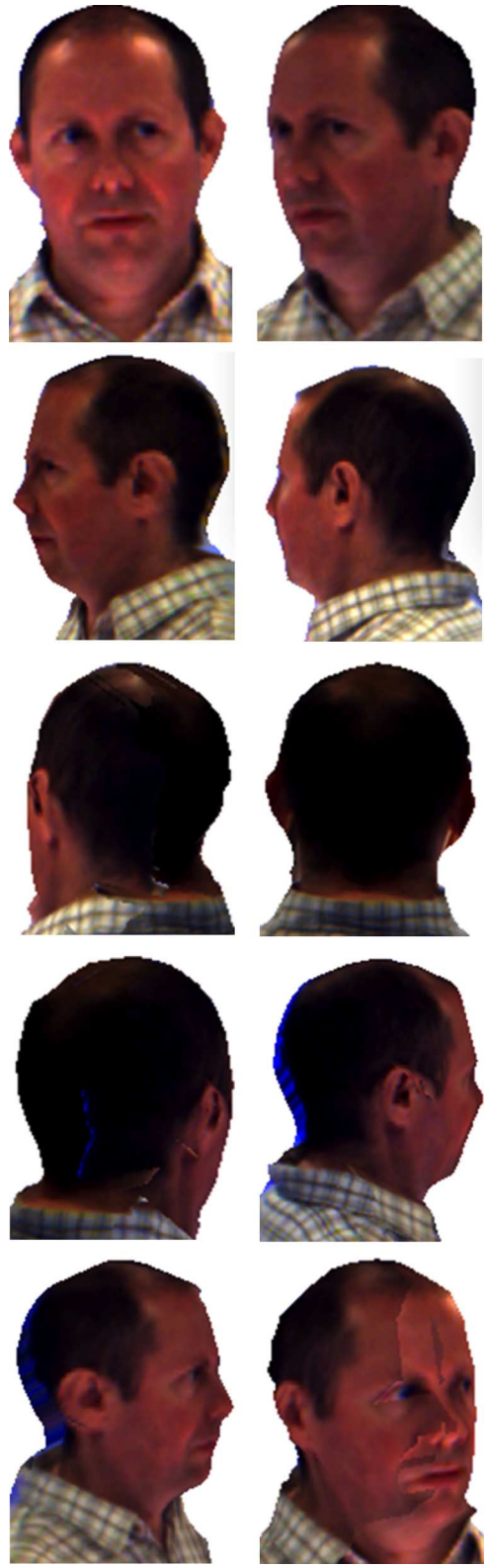


Fig. 7. Views of model, reconstructed from 8 surrounding cameras, viewed from all directions in 36° internals. Subject stood 1 m back from center with eyes turned. The three colors of the eye are visible when both eyes are in view, and the whites of the eyes from the side. The corner of the eye is distinguishable from the iris when the eye is turned to it. A polygonal effect is noticeable both from the profile and from poor color calibration across the cameras.

Some processes run much faster across CPUs and others across GPU. Polyhedral (model-based) reconstruction is ten

Fig. 8. Two reconstructions giving different impressions of the subject from data captured at the same time. The person on the right seems older, and less healthy and happy than the one on the left. The only difference in input data is the elevation of the camera from which the form is textured. That on right is steeper. This shows the danger of placing cameras above high screens immediately in front of the subject.

TABLE I
PERFORMANCE OF PROCESSES ON CPU AND GPU

Process	CPU	GPU	Unit tested
Reconstruction (polygonal)	24.5ms (i7 4 core)	242ms (30 core)	
Reconstruction and Rendering (Voxel)	N/A	41.66ms	Model from frames from 8 cameras
Mesh compression	3.5ms	---	
Mesh de-compression	2ms	---	
Video encoding (H.264@ 1Mbit/s)	3.82ms (i5-650)	5.4ms (560Ti)	
Video-decoding (H.264@ 1Mbit/s)	1.1ms (i5-650)	4.4ms – (560Ti)	
Lens correction	34.19ms (i5-650)	0.11ms (560Ti)	Single frame
Segmentation	6.15ms (i5-650)	0.18ms (560Ti)	
Filtering	---	0.29ms (560Ti)	

TABLE II
CONSERVATIVE ESTIMATION OF PERFORMANCE OF PROPOSED DISTRIBUTIONS (8 CAMERAS)

Component group	Frame rate (Hz)
Model-based approach	18
Render-based approach	23
Capture node	65

times faster on the CPU. Lens correction is 300 times slower. Other differences are more modest.

The reader should take care when interpreting Table I, as image processing timings are given per frame, while that for other processes are given for a model, derived from 8 frames. A conservative estimate can be derived through multiplication of times for each frame. However, this will not capture the advantage of parallelization.

A conservative estimate of the frame time to run all processes on single machine can be achieved by adding the timings of model processes with a sum of the eight fold of all image processes. For the model-based approach the model processes are reconstruction and mesh compression. For the render-based approach it is the combined process of voxel reconstruction and rendering. We leave such estimation and its implication to the discussion.

D. Discussion

In our opinion, the lack of concavities only effects how eyes are perceived from the side (Fig. 4 left) and not from the front (Fig. 4 right). The appearance of an eye socket is captured by the texture image. When viewed face on, this and the mind seem to work together to give the impression of an eye being in a socket. However, the texture is not altered by parallax, but given the relatively shallow depth of the concavity the impact of this is minimal. If this causes a noticeable effect during conversations, we have yet to see it. Based on the evidence we have seen, of which some is presented here, we are of the opinion that a lack of geometric concavity in the reproduced eye socket is unlikely to be noticed during normal conversational interaction. We therefore argue, that what is most criticized about shape from silhouette actually has little impact within the application of telepresence.

What has far more of an impact is the number, placement and calibration of cameras. A reconstruction that looks plausible when viewed from within the arc it was captured, can look grotesque when viewed from outside of the capture arc. To avoid this, we advise that cameras are placed to each side of every possible viewing angle. We previously showed that eye-gaze could be accurately estimated from a textured form reconstructed from cameras up to around 30 degrees apart [30]. Thus, putting the two together, to align two spaces so that they appeared adjacent, we would need 7 cameras ($1 + 180/30$). This would allow gaze to be maintained as one person walked past the other. To align remote spaces so that they appeared to completely coincide would require 12 cameras (360/30). This would allow gaze to be maintained as one person moved around another.

Placing cameras above eye high can create different, yet still plausible, appearance in people (Fig. 8). This might affect their apparent age and wellbeing and even cause both to change as they move towards the observer or as the observer takes a seat. We advise that cameras are placed as close to eye level as possible without obscuring the image of another person on the screen. We further advise that seats at only one side of a telepresent link are avoided.

Using timings gained for each process on various parallel CPUs and GPUs, we made a rough estimate of what could be achieved on a single machine. This suggested that the CPU/GPU computational resource of single commodity machine should be sufficient to support 10 cameras at 10 Hz, thus exceeding our minimal requirements.

However, we argue that distribution of acquisition stage was still necessary when using only commodity computers. This is because of the practicalities of moving the data from the cameras onto the central computer. Cameras need to be in an arc or preferably circle of radius 3.5 m. Therefore, degradation of image is likely across the wire. Furthermore, while commodity computers can often run up to 4 cameras from 4 USB ports, the bandwidth of USB busses and controllers would be a bottleneck when scaling up.

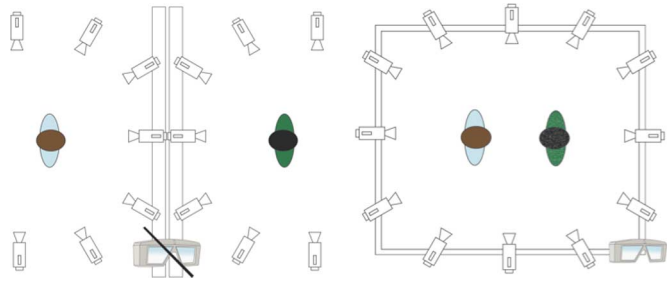


Fig. 9. Display and capture configurations for minimal (left) and desirable (right) spatial requirements. The former aligns spaces so that they appear adjacent, allowing people to walk past or up to each other. The latter allows them to also walk around each other.

This suggests using either Ethernet cameras or physically distributed capture node computers driving cameras from, say, USB. In either case, 10 HD cameras would generate 10 gigabits/sec. This would require an expensive network infrastructure. Such expense can be avoided by using commodity computers as capture nodes and having them compress the streams before sending. The question then becomes how to best split processing between central machine and capture node. Doing most of acquisition processing on capture nodes increases scalability through reducing the bottleneck of the central machine. However, reducing it allows each camera node to handle more cameras, thus reducing cost.

E. Non-Functional Requirements

We now summarize our non-functional requirements. These are derived from the above evidence of impact of camera placement on visual and/or spatial characteristics of shape from silhouette. Non-functional requirements, Fig. 9:

- Minimal (adjacent): frontal semicircle of 7 HD cameras.
- Desirable (coincident): Circle of 12 HD cameras.

VI. A Proposal for Improved Distribution of Processes

The purpose of this section is to propose improved distribution of processes across networks, computers, and multi-core CPU and GPUs, towards meeting our requirements. We propose two models of distribution across CPU and GPU resources on distributed nodes. One for model-based and the other, render-based reconstruction. Distribution of the model and render-based approaches is shown in Fig. 10. There are three differences in the distribution of the remaining processes.

Firstly is the location of the reconstruction algorithm. In the model-based approach, it is executed on the CPU and in the render-based approach, on the GPU. This is because the former contains much branching, whereas, the latter undertakes the same parallel instructions on different data.

The second difference is the rendering. With the model-based approach, the rendering is run on a separate machine. This is to remove the effects of the wide area network on correctness of the textured model, as described above. With the render-based approach this is not practical, as reconstruction and rendering are tightly linked within the graphics pipeline. The render process thus shares the graphics card with the majority of the image processing.

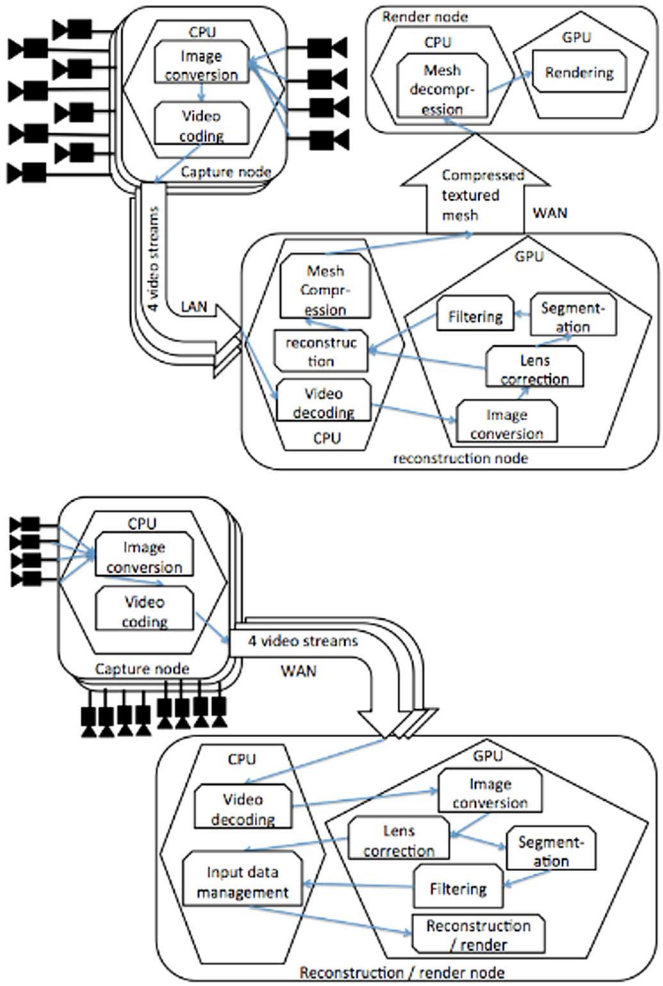


Fig. 10. Proposed distribution for (top) the model-based and (bottom) render-based approaches.

The final difference is in additional processes. Each has an different additional process that runs on the CPU. The model-based approach can include a mesh compression process. The render-based approach requires input data management.

The capture nodes, in both cases, simply capture and encode (compress) the videos, converting format between the two processes. All of this is done on the CPUs. The image processing on the reconstruction node is split between CPU and GPU. This split is identical for both reconstruction approaches. In both, all image processing, apart from video decoding, is done on the GPUs.

We now give conservative estimations of performance for eight cameras. Eight is just above our minimal requirement and most component testing used that number. By conservative we mean that processes on a given machine are assumed to run sequentially. In practice much runs in parallel.

We implemented the model-based approach (save mesh compression) and obtained a frame rate between 15 and 20 fps with 10 cameras. However, this was only possible when each capture node grabbed from no more than a pair of cameras.

VII. Discussion

Theoretically both approaches should meet our minimal temporal requirements and approach the desirable ones. This is even

when the parallelization across processes on a given machine is not considered. We have only implemented an end-to-end system using the model-based approach, however, without mesh compression. Mesh compression should take just over 3.5 ms for 10 cameras and thus should make negligible difference.

The implementation always well exceeded our minimal requirement for frame rate of 10 Hz, and sometimes the desired 20 Hz. The variation is almost certainly due to human movement impacting on complexity of silhouette. The implementation has not allowed parallelization across processes. Thus it is no surprise that it is close to our conservative estimate. Addressing this should increase the performance to reliably meet the desirable requirements regardless of human movement. It fell well short of our desirable requirement for latency. However, its latency across a local area network is not discernible from that of Skype, which is almost certainly the most widely used video-conferencing tool. This is evident from comparison of facial actions in Fig. 5. The reading of videos into any given computer still remains the bottleneck. Currently we can only service two to four cameras from each capture node. However, the capture nodes used are five years old.

We are performing very basic segmentation, impractical in everyday settings. In general, segmentation is a task suited for GPU parallelization as it typically uses little branching. The model-based approach makes less use of the GPUs than the render-based approach. It might make more sense for a better segmentation approach to be run on the reconstruction node in the model-based approach and on the capture client in the render approach.

Our approach performs segmentation after video encoding/decoding. An experiment showing the validity of this is being written up separately. Had we not implemented it, we may have needed additional time management to synchronize delivery of pictures and silhouettes. Furthermore, there would have been a network bandwidth overhead. However, a black and white silhouette image uses far less bandwidth than the color image it was taken from.

VIII. Conclusion

Video conferencing has improved teleconferencing yet it does not support much of the non-verbal interactions used in physical face-to-face meetings. Crucially it does not support the important relationship between interpersonal distance, gaze, and the wide range of non-verbal signals of intimacy. Feelings of ambient togetherness, trust, empathy and rapport, all require a wide range of non-verbal cues to be linkable. Technology disturbance to these links can also change the meaning and course of conversations. Communication technologies that allow people to move around faithful live reconstructions of each other have the potential to support a wide range of linkable non-verbal cues. Yet after 15 years, prototypes are just leaving infancy. Achieving sufficiently balanced visual, spatial and temporal qualities remains challenging. This article described this challenge, setting out both requirements and impacting factors, and evaluated our approach to tackle it.

We derived basic functional requirements for general telepresence. From key knowledge of non-verbal interaction in the natural world and the influence of technology, we stipulated the following. Most non-verbal cues must be visible between people moving within social space (3.5 m). These include eye-gaze, interpersonal distance, body torque and facial expressions. At minimum this is for people walking in front of each other and desirably for those moving around each other. Adding the length of a conversational pause long enough to respond to, to the ITU recommended delay for audio latency, derives our target for delay of 400 ms. Frame rate must be sufficient to capture micro expressions (10 Hz). However, a frame rate comparable to today's common video conferencing systems is desirable (20 Hz).

Combing a 3D medium and interface allows spatial context to be shared. Deriving the content of the medium from live video, captures general appearance and most non-verbal cues. However, creating and delivering the live full 3D video in real time is challenging. A balance must be sought between visual, spatial and temporal qualities.

Through describing our own approach and its visual, spatial and temporal characteristics, we demonstrated the implications of such a balance. Specifically we demonstrated that while the technique of shape from silhouette could provide such a balance, care must be taken in choosing the number and placement of cameras.

While shape from silhouette cannot capture concavities and thus eye sockets, this is not apparent when viewed from angles typical in conversations. We had previously shown that it allows estimation of eye gaze to accuracies underpinning social interaction in the natural world [30]. 8 cameras were sufficient to create a 3D model of a person that looked plausible from all sides. When cameras are placed in an arc, rather than a circle, the reconstruction looks plausible when viewed from within the arc but hideous deformities are seen from outside it. When cameras are placed above a screen high enough to display a life size person, someone moving close to the screen might appear to age and look less healthy and happy when seen at remote sites. Now understood, these issues can be easily avoided by setup. This setup places certain requirements on the underlying system, which is now described.

Required bandwidth and thus the latency are dependent on number of cameras, which is dependent on placement. We reported the timings of component processes across multiple core CPUs and GPUs. Results suggested that all the processing necessary to meet our requirements could theoretically be supported on a single commodity computer. However, the bandwidth of commodity interfaces to the cameras were insufficient, be they via network or USB. Furthermore, cable lengths could induce noise.

Two distribution models were proposed, each fitting a different approach to video-based reconstruction. Worst-case estimates were made from the above component timings to show that theoretically both should meet our minimal requirements for frame rate and visual/spatial quality. One was mostly implemented. It met all minimal requirements when ran across a local area network. However, it fell well short of the desirable latency but met the minimal requirement of matching that of Skype.

The current state of the system is as follows. All components are implemented but in distinct subsets in various versions. No version has both mesh compression and display parallax. Supported operating systems are Microsoft, OS and Linux. Limita-

tions include the following. Background segmentation against a moving background has only been tested against a single projection wall. While we have demonstrated our virtuality avatar within a space science and operations application, in a link between sites UK and Germany, video input was prerecorded.

The foci of our current work is completion of integration with the Mars simulator and convergence of functionality of versions. Priorities for future work are: increasing parallelization across processes through less conservative handshaking; selective composition of texture maps to complement mesh compression; and subjective quality testing, such as [48], following ITU-T Recommendation P.1301 [49].

ACKNOWLEDGMENT

The authors would like to especially thank Robin Wolff from DLR Germany for his many contributions to porting of our system to the space science application.

REFERENCES

[1] A. Sadagic, H. Towles, L. Holden, K. Daniilidis, and B. Zeleznik, "Tele-immersion portal: Towards an ultimate synthesis of computer graphics and computer vision systems," in *Proc. IWP*, Philadelphia, PA, USA.

[2] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs, "The office of the future: A unified approach to image-based modeling and spatially immersive displays," *ACM SIGGRAPH*, vol. 32, no. Annual Conference Series, pp. 179–188, 1998.

[3] G. Meenakshisundaram, S. Krishnan, and C. T. Silva, "Surface reconstruction based on lower dimensional localized delaunay triangulation," *Comput. Graphics Forum*, vol. 19, no. 3, pp. 467–478, Dec. 1991.

[4] R. Schroeder, A. Steed, A.-S. Axelsson, I. Heldal, Å. Abelin, J. Wideström, A. Nilsson, and M. Slater, "Collaborating in networked immersive spaces: As good as being there together?," *Comput. Graphics*, vol. 25, pp. 781–788, Oct. 2001.

[5] M. Gross, S. Würmlin, M. Naef, E. Lamboray, C. Spagno, A. Kunz, E. Koller-Meier, T. Svoboda, L. Van Gool, S. Lang, K. Strehlke, A. V. Moere, and O. Staadt, "Blue-c: A spatially immersive display and 3D video portal for telepresence," *ACM Trans. Graphics*, vol. 22, no. 3, pp. 819–827, 2003.

[6] O. Grau, A. Hilton, J. Kilner, G. Miller, T. Sargeant, and J. Starck, "A free-viewpoint video system for visualisation of sport scenes," *SMPTE Moti. Imag.*, vol. 116, no. 5–6, pp. 213–219, May 2007.

[7] W. Waizenegger, I. Feldmann, and O. Schreer, "Real-time patch sweeping for high-quality depth estimation in 3D videoconferencing applications," in *Proc. SPIE 7871 RTIVP*, San Francisco, CA, USA, Feb. 2011.

[8] B. Baumgart, "A polyhedron representation for computer vision," in *Proc. IFIPS*, May 1975, vol. 44, pp. 589–596.

[9] A. Laurentini, "The visual hull concept for silhouette-based image understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 2, pp. 150–162, Feb. 1994.

[10] J. Franco and E. Boyer, "Efficient polyhedral modeling from silhouettes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 414–427, Mar. 2009.

[11] J. Allard, J. Franco, C. Menier, E. Boyer, and B. Raffin, "The GrImage platform: A mixed reality environment for interactions," in *Proc. IEEE Int. Conf. Comput. Vis. Syst.*, 2006, pp. 46–46.

[12] O. D. Escoda, J. Civit, F. Zuo, H. Belt, I. Feldmann, O. Schreer, E. Yellin, W. Ijsselsteijn, R. van Eijk, D. Espinola, P. Hagendorf, W. Waizennegger, and R. Braspenning, "Towards 3D-aware telepresence: Working on technologies behind the scene," in *Proc. ACM CSCW*, Savannah, GA, USA, Feb. 06–10, 2010.

[13] A. Steed, W. Steptoe, W. Oyekoya, F. Pece, T. Weyrich, J. Kautz, D. Friedman, A. Peer, M. Solazzi, F. Tecchia, M. Bergamasco, and M. Slater, "Beaming: An asymmetric telepresence system," *IEEE Comput. Graphics Appl.*, vol. 32, no. 6, pp. 10–17, Dec. 2012.

[14] D. J. Roberts, A. S. Garcia, J. Dodiya, R. Wolff, A. J. Fairchild, and T. Fernando, "Collaborative telepresence workspaces for space operation and science," in *Proc. IEEE VR*, 2015.

[15] P. Lincoln, G. Welch, A. Nashel, A. State, A. Ilie, and H. Fuchs, "Animatronic shader lamps avatars," *Virtual Real.*, vol. 15, no. 2–3, pp. 225–238, Jun. 2011.

[16] D. J. Roberts and P. M. Sharkey, "Maximising concurrency and scalability in a consistent, causal, distributed virtual reality system, whilst minimising the effect of network delays," in *Proc. IEEE WETICE*, Boston, MA, USA, pp. 161–166.

[17] D. J. Roberts, R. Wolff, and O. Otto, "Constructing a Gazebo: Supporting team work in a tightly coupled, distributed task in virtual reality," *MIT Presence: Teleoper. Virtual Environ*, vol. 12, no. 6, pp. 644–667, Dec. 2003.

[18] R. Wolff, D. J. Roberts, and O. Otto, "A study of event traffic during the shared manipulation of objects within collaborative virtual environments," *MIT Presence: Teleoper. Virtual Environ.*, vol. 13, no. 3, pp. 251–262, Jun. 2004.

[19] R. Wolff, D. J. Roberts, A. Steed, and O. Otto, "A review of tele-collaboration technologies with respect to closely coupled collaboration," *Int. J. Comput. Applicat. Technol.*, vol. 29, no. 1, pp. 11–26, Jul. 2007.

[20] D. J. Roberts, M. Ai-Liabi, O. Otto, R. Wolff, and A. Al-Khalifah, "Reducing fragmentation in telecollaboration by using IPT interfaces," in *Proc. Eurographics ACM SIGGRAPH EGVE*, 2005, pp. 211–216.

[21] D. J. Roberts, I. Heldal, O. Otto, and R. Wolff, "Factors influencing flow of object focused collaboration in collaborative virtual environments," *Virtual Real.*, vol. 10, no. 2, pp. 119–133, Sep. 2006.

[22] R. Wolff, D. J. Roberts, A. Murgia, N. Murray, J. Rae, W. Steptoe, A. Steed, and P. M. Sharkey, "Communicating eye gaze across a distance without rooting participants to the spot," in *Proc. IEEE ACM DSRT*, Vancouver, BC, Canada, Sep. 2008, pp. 111–118.

[23] D. J. Roberts, R. Wolff, J. Rae, A. Steed, R. Aspin, M. Mcintyre, A. Pena, O. Oyekoya, and W. Steptoe, "Communicating eye-gaze across a distance: Comparing an eye-gaze enabled immersive collaborative virtual environment, aligned video conferencing, and being together," in *Proc. IEEE Virtual Reality Conf.*, Lafayette, LA, USA, Mar. 2009, pp. 135–142.

[24] W. Steptoe, A. Steed, A. Rovira, and J. Rae, "Lie tracking: Social presence, truth and deception in avatar-mediated telecommunication," in *Proc. ACM CHI*, Apr. 2010, pp. 1039–1048.

[25] C. Moore, T. Duckworth, and D. J. Roberts, "Synchronization of images from multiple cameras to reconstruct a moving human," in *Proc. IEEE/ACM DSRT*, Fairfax, VA, USA, Oct. 2010, pp. 53–60.

[26] T. Duckworth and D. J. Roberts, "Camera image synchronization in multiple camera real-time 3D reconstruction of moving humans," in *Proc. IEEE/ACM DSRT*, Salford, U.K., Oct. 2011, pp. 138–144.

[27] T. Duckworth and D. J. Roberts, "Parallel processing for real-time 3D reconstruction from video streams," *Real-Time Image Proc.*, vol. 9, no. 3, pp. 427–445, Dec. 2012.

[28] R. Aspin and D. J. Roberts, "A GPU based, projective multi-texturing approach to reconstructing the 3D human form for application in tele-presence," in *Proc. ACM CSCW*, NY, New York, USA, 2011, pp. 105–112.

[29] T. Duckworth and D. J. Roberts, "3DRecon A utility for 3D reconstruction from video," in *EEG, Eurographics, ICAT, Proc. JVRC*, Madrid, Spain, Oct. 2012.

[30] D. J. Roberts, J. Rae, T. Duckworth, C. Moore, and R. Aspin, "Estimating gaze of a virtuality human," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 4, pp. 681–690, Apr. 2013.

[31] R. B. Zajonc, "Feeling and thinking: Preferences need no inferences," *Amer. Psychol.*, vol. 35, no. 2, pp. 151–175, Feb. 1980.

[32] M. L. Patterson, "An arousal model of interpersonal intimacy," *Psychol. Rev.*, vol. 83, no. 3, pp. 235–245, May 1976.

[33] P. Ekman, "Facial expression and emotion," *Amer. Psychol.*, vol. 48, no. 4, pp. 376–379, Apr. 1993.

[34] S. Porter and L. ten Brinke, "Reading between the lies: Identifying concealed and falsified emotions in universal facial expressior," *Psychol. Sci.*, vol. 19, no. 5, pp. 508–514, May 2008.

[35] M. Argyle and J. Dean, "Eye-contact, distance and affiliation," *Sociometry*, vol. 28, no. 3, pp. 289–304, 1965.

[36] J. N. Bailenson, J. Blascovich, A. C. Beall, and J. M. Loomis, "Equilibrium theory revisited: Mutual gaze and personal space in virtual environments," *Presence: Teleoper. Virtual Environ.*, vol. 10, no. 6, pp. 583–598, Dec. 2001.

[37] E. Hall, *The Hidden Dimension*. New York, NY, USA: Anchor Books, 1966, 0-385-08476-5.

[38] M. Slater, in *Presentat. Construct. Sci. of Social Interact. Workshop*, London, U.K., Sep. 2014, UCL.

[39] D. P. Pertaub, M. Slater, and C. Barker, "An experiment on public speaking anxiety in response to three different types of virtual audience," *Presence: Teleoper. Virtual Environ.*, vol. 11, no. 1, pp. 68–78, Feb. 2002.
[40] R. van Eijk, A. Kuijsters, K. Dijkstra, and W. A. IJsselsteijn, "Human sensitivity to eye contact in 2D and 3D videoconferencing," in *Proc. QoMEX*, Jun. 2010, pp. 76–81.
[41] N. Murray, D. J. Roberts, A. Steed, P. M. Sharkey, P. Dickerson, J. Rae, and R. Wolff, "Eye gaze in virtual environments: Evaluating the need and initial work on implementation," *Concurrency and Comput.: Practice and Experience*, vol. 22, no. 11, pp. 1437–1449, Feb. 2009.
[42] S. M. Anstis, J. W. Mayhew, and T. Morley, "The perception of where a face or television 'portrait' is looking," *Amer. J. Psychol.*, vol. 82, no. 4, pp. 474–489, 1969.
[43] W. Waizenegger, N. Atzpadin, O. Schreer, I. Feldmann, and P. Eisert, "Model based 3D gaze estimation for provision of virtual eye contact," in *Proc. IEEE ICIP*, Orlando, FL, USA, Sep. 2012, pp. 1973–1976.
[44] K. Kim, J. Bolton, A. Girouard, J. Cooperstock, and R. Vertegaal, "Telehuman: Effects of 3D perspective on gaze and pose estimation with a life-size cylindrical telepresence pod," in *Proc. CHI*, New York, NY, USA, 2012, pp. 2531–2540.
[45] H. H. Baker, N. Bhatti, D. Tanguay, I. Sobel, D. Gelb, M. E. Goss, W. B. Culbertson, and T. Malzbender, "Understanding performance issues in coliseum, an immersive video conferencing," *ACM Trans. Multimedia Comput., Commun. Apps*, vol. 1, no. 2, pp. 190–210, May 2005.
[46] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Phonetics*, vol. 38, pp. 555–568, Jan. 2010.
[47] P. Ekman and W. V. Friesen, *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Oxford, U.K.: Psychologists Press, 1978.
[48] *One Way Transmission Time (05/03) International Telecommunication Union. In Force 30 September 2014*, ITU Recommendation G.114.
[49] G. Berndtsson, M. Folkesson, and V. Kulyk, "Subjective quality assessment of video conferences and telemeetings," in *Proc. PV*, Munich, Germany, May 2012, pp. 25–30.
[50] W. Steptoe, O. Oyekoya, A. Murgia, R. B. Wolff, J. P. Rae, E. Guimaraes, D. J. Roberts, and A. J. Steed, "Eye tracking for avatar eye gaze control during object-focused multiparty interaction in immersive collaborative virtual environments," in *Proc. IEEE Virtual Reality Conf.*, Lafayette, LA, USA, Mar. 2009, pp. 83–90.

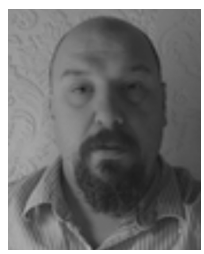
David J. Roberts is a Professor of Telepresence at the University of Salford. He has over 100 publications mostly in telepresence. David is a PI of the EU CROSS DRIVE project, applying telepresence to support distributed space science and mission operation and planning, and recently led the EPSRC Eye catching project which developed a telepresence approach to support mutual eye gaze between moving people. He both led Salford's Centre for Virtual Environments and chaired IEEE/ACM DSRT for 6 years.

Allen J. Fairchild is a Ph.D. student at the University of Salford, under the supervision of Prof. David Roberts. His Ph.D. seeks to refine an experimental 3D reconstruction telepresence platform to sufficient visual, spatial and temporal quality, to allow observers to link non-verbal cues across resources. He previously worked in PRimA developing pattern recognition software.

Simon P. Campion is a Project Manager in the THINKLab of University of Salford, where he manages a portfolio of Virtual Reality research projects. He is also undertaking a part time Ph.D. under the supervision of David Roberts. His Ph.D. is assessing the impact of medium and interface on non-verbal communication.

John O'Hare is Technical Director of the octave visualization facility at the University of Salford and is undertaking a part time Ph.D. supervised by Prof. David Roberts. His Ph.D. studies telepresence mediated togetherness, focusing on gaze-enabled representation through furniture. He previously worked as a Virtual Environments consultant at the University of Salford.

Carl M. Moore is now a research and development Software Engineer at 4Sight Imaging Ltd. He completed his Ph.D., supervised by Prof. David Roberts, at the University of Salford in 2013. The focus of his Ph.D. was distribution and processing of video for real-time 3D telepresence. He is now applying this in industry through developing and optimizing computer vision and machine learning systems.



Rob Aspin is a Senior Lecturer at the University of Salford, where he heads Computer Science. He completed his Ph.D. at the University of Salford in 2014, supervised by Prof. David Roberts. The focus of his Ph.D. was volumetric reconstruction of humans for telepresence. His core research is computer graphics, including volumetric rendering of internal and external human form.

Tobias Duckworth is Director of Orogenic Developments Ltd. He completed his Ph.D. at the University of Salford in 2013, supervised by Prof. David Roberts, focusing on improving the performance of 3D reconstruction algorithms through a novel approach to parallel processing. He is now applying his research expertise in industry, acting as a consultant to consumer electronics companies.

Paolo Gasparello completed a Ph.D. at the Scuola Superiore Sant'Anna in 2012 where he now works as a Researcher on Virtual Reality projects. His Ph.D. was supervised by Prof. Franco Tecchia, with a placement year, spent at the University of Salford, supervised by Prof. David Roberts. The focus of the Ph.D. was 3D mesh representations for multimodal distributed virtual environments.

Franco Tecchia is Assistant Professor at Scuola Superiore Sant'Anna with primary specialization on Software Engineering and Computer Graphics. He obtained a Master degree from the University of Pisa and Ph.D. from UCL under supervision of Mel Slater. He heads the *Computer Graphics and Virtual Environment Areat* at the TeCIP Institute. Projects include VIRTUAL, PURE FORM, CREATE, PRESENCCIA, SKILLS, BEAMING, VERE and the NoE ENACTIVE and INTUITION. He is an associate editor of journal *Presence: Teleoperators and Virtual Environments* (MIT PRESS).