This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2580425, IEEE Transactions on Circuits and Systems for Video Technology

A Mixed Reality Telepresence System for Collaborative Space Operation

Allen J. Fairchild, Simon P. Campion, Arturo S. García, Robin Wolff, Terrence Fernando and David J. Roberts

***Abstract*—This paper presents a Mixed Reality system that results from the integration of a telepresence system and an application to improve collaborative space exploration. The system combines free viewpoint video with immersive projection technology to support non-verbal communication, including eye gaze, inter-personal distance and facial expression. Importantly, these can be interpreted together as people move around the simulation, maintaining natural social distance. The application is a simulation of Mars, within which the collaborators must come to agreement over, for example, where the Rover should land and go.**

The first contribution is the creation of a Mixed Reality system supporting contextualization of non-verbal communication. Two technological contributions are prototyping a technique to subtract a person from a background that may contain physical objects and/or moving images, and a light weight texturing method for multi-view rendering which provides balance in terms of visual and temporal quality. A practical contribution is the demonstration of pragmatic approaches to sharing space between display systems of distinct levels of immersion. A research tool contribution is a system that allows comparison of conventional authored and video based reconstructed avatars, within an environment that encourages exploration and social interaction. Aspects of system quality, including the communication of facial expression and end-to-end latency are reported.

***Index Terms*—Computer supported collaborative work, mixed reality, telepresence, 3D video based reconstruction, background-foreground segmentation, space science.**

I. Introduction

THIS paper presents the integration of a telepresence system [1] and a Mars simulator [2], in support of a European Union funded CROSS DRIVE project [3]. CROSS DRIVE seeks to improve collaboration between countries across space mission control, science and engineering. The aim of the work is to support most Non-Verbal Communication (NVC) while contextualizing it both within a scientific simulation (of Mars), and a team of people "beamed" into it from different locations.

The motivation behind CROSS DRIVE is to reduce divergence in both planning and science that can creep in between the occasional expensive group visits to another country's simulation facilities. This would be simple if only technology was already available to support across a distance, the quality of dialogue achievable when a team is physically immersed together within a simulation.

Unfortunately, contextualizing a wide range of non-verbal communication within a simulation in which collaborators can move around, is difficult [1]. Approaches tend to favor either the range of non-verbal communication supported (video conference), the level to which its spatial contextualization can be communicated, or freedom of movement within the shared space (collaborative virtual environments). In simple terms, it is surprisingly difficult to communicate both what someone looks like and what or who she is looking at, without constraining movement, e.g. with seats. The problem is that both non-verbal communication and environment based problem solving are inherently spatial.

To understand the relevance of this problem to the CROSS DRIVE project, consider the following scenario. A scientist might point to where she thinks the Mars Rover should be sent. An engineer frowns and points first to the suspension of the Rover and then the terrain it would have to cross. However, seeing that only the mission controller is looking at her, she moves into the scientist's line of sight and throws up her hands. In video conferencing, what people are looking and pointing at would likely be lost and someone cannot walk into the line of sight of a remote user to capture attention. With immersive collaborative virtual environments using conventional motion driven authored avatars, facial expression and often identity would be lost.

The key challenge is to support a wide range of non-verbal communication contextualized within a real simulation and application. This paper describes a set of sub-challenges that were addressed. These include segmentation against backgrounds that may include moving images on a display, extending immersion of a wall display to allow another's space to be entered, real-time texturing of face without overly distorting its appearance, enabling scalability of a streamed 3D video avatar and sharing of spaces without occluding eyes.

Paper submitted: 09/10/2015. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 607177 CrossDrive. The UK's Engineering and Physical Research Council (EPSRC) through DTA and CASE PhD studentships, and through grant EP/E010032/1; and VISIONAIR through TNA-131 also supported this work.

A. J. Fairchild, S. P. Campion, A. S. García, T. Fernando, and D. J. Roberts are with the University of Salford, Salford, M5 4WT, U.K. (e-mail: a.j.fairchild@edu.salford.ac.uk; s.p.campion@salford.ac.uk; a.s.garciajimenez@salford.ac.uk; t.fernando@salford.ac.uk; d.j.roberts@salford.ac.uk).

R. Wolff is with the German Aerospace Center (DLR), Lilienthalplatz 7, D-38108, Braunschweig, Germany. (e-mail: robin.wolff@dlr.de).

Copyright (c) 2016 IEEE. Personal use of this material is permitted.
However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2580425, IEEE Transactions on Circuits and Systems for Video Technology

II. Background

Mixed Reality (MR) merges information from the real and virtual worlds, using mediums and displays. Depending on the amount of virtual and real information, a particular application can fall in different points of the MR continuum described by Milgram and Kishino in [4]. Within this continuum, it is acceptable to place a Head Mounted Display that cannot be seen through in Virtual Reality (VR) and one that can in Augmented Reality. It is also understood that a natural environment overlaid by graphics is Augmented Reality and a virtual avatar abstracted from video of a user is Virtuality. It is less straightforward to place either the Mars simulator or its combination with our telepresence system within these discrete containers. Thus we describe it as a mixed reality system.

A. Reproducing NVC in Telepresence

Telepresence is the feeling of being in a different place derived from a technology. Technologies range from web cameras to embodied humanoid robots. The term is often used to describe systems that attempt to reproduce face-to-face meetings across a distance. Here we focus on such systems that range from video conferencing, through immersive virtual environments, to the combination of video based reconstruction and immersive displays. The focus is the affordances that each give to communicating NVC.

NVC has been described as the transmission of information and influence through an individual's physical and behavioral cues [5]. This transmission is usually via a range of cues that often only retain correct meaning when interpreted together and within context. There are different technologies that can be used in order to capture aspects of NVC for computer-mediated conversation. Video is sufficient for capturing most NVC 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 [1]. At the same time, it is difficult to capture mutual eye gaze, as the camera and display cannot share the same physical position in the space and the user can only look at one of them at a time [6]. On the other hand, VR, in its purer form, immerses people in 3D computer graphics, tracking some of their movements and hence capturing some of the NVC of the user. The use of immersive displays and life size motion tracked avatars make it possible to retain spatial context so the user can be seen by another remote participant sharing the environment [7].

Avatars of varying detail are used to represent users in Immersive Collaborative Virtual Environments (ICVE). The standard approach uses live motion tracking data from the user to mirror her movement through a remote avatar. This varies from simple head and hand tracking to more complex approaches, such as [8], incorporating eye gaze. This method of user representation has proven successful when completing collaborative tasks [7]. Studies have also illustrated how a number of NVC can be successfully portrayed using virtual characters [9]. However, capture and display of facial expressions in real time alongside full body tracking is a much bigger challenge. Affordable commercial software such as Faceshift [10] does allow for real time marker-less capture of facial expressions but relies on a depth camera being so close to the face to capture detail which would be problematic when also capturing the body. Marker based solutions could be used to animate a facially rigged character [11], however, this is not plausible for frequent use because it requires too much time for setup and could also make the wearer feel uncomfortable.

A contemporary technology approach to telepresence is the use of 3D reconstructed video for communication [12][13]. In such systems, avatars are created in real-time from several video streams. The motivation of such an approach, was our past study comparing the use of gaze enabled ICVE and video conferencing [14]. However, these two systems produce differing levels of NVC, therefore only telling part of the story [1]. In order not to confuse the story, it is essential to faithfully communicate at least eye gaze, interpersonal distance and facial expression [1]. The advantage of combining video based reconstruction and immersive displays is that it attempts faithful transmission of all of these and more [1]. If people are going to act naturally, it helps if they are not encumbered by excessive markers and restrictions on movements within extent of social space [1]. Another advantage of the above approach is that the only thing that needs tracking is viewpoint [1] and this can be done across the extent of social distance by, for example, placing markers on glasses or a hat. However, a challenge is supporting a sufficient balance of visual, spatial and temporal quality [1].

B. Live generation of 3D avatars

Live generation of 3D avatars can be achieved actively or passively. Active methods include time-of-flight devices that project light towards and analyze the time it takes to reach points on an object [15], and structured light devices that analyze disparity in a projected pattern to form a 3D representation [16][17]. The Kinect is an example of a structure light device that has been used in much recent telepresence research. A single Kinect can achieve a partial 3D reconstruction of the subject in the plane it is pointing towards. However, full 3D reconstruction [1] is required to allow people to use movement in space as part of communication, so that a person does not look like an empty shell when viewed from the side. This requires the stitching together of depth maps from multiple Kinects positioned around the subject. Herein lays a problem, because the projected patterns from the individual Kinects interfere with one another, and this causes deterioration in the quality of the depths maps, typically resulting in less faithful shapes with holes in them. Interference between multiple Kinects can be reduced [18] and there are numerous examples of Kinects being used for 3D capture [19][20][21] but, to the authors knowledge, only two produced a 3D avatar that was generated without the surrounding environment [22] [23]. However, there is a bigger problem. The resolution of structured light patterns projected by a Kinect onto a face, is far less than that of pixels capturing a face from the same distance with an RGB camera. This results in poor resolution of shape of face if cameras are far enough away to allow natural interpersonal

Copyright (c) 2016 IEEE. Personal use of this material is permitted.
However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2580425, IEEE Transactions on Circuits and Systems for Video Technology

distance [1].

With passive methods, also known as Image Based Reconstruction (IBR) [24], the 3D model is derived from a set of conventional camera images taken from different angles and positions, capturing light in the visible spectrum. These methods then use this information to generate form (geometry) and appearance (texture). Video Based 3D Reconstruction (VBR) extends this across time to also capture movement from multiple video streams. There are several VBR approaches suitable for reconstructing the 3D form of an entire human that fulfill the requirements of our system. One example is multi-view stereo [25], which is capable of producing high quality, spatially accurate and visually faithful models. Unfortunately, it currently falls short of the temporal requirements of a real time telepresence system. Techniques based on the shape-from-silhouette (SfS) principle [26], which form an approximation to the 3D shape known as the visual hull [27], have demonstrated that they can fulfill this requirement whilst retaining a faithful reconstruction [1]. For that reason, methods to extract silhouette information required for SfS become paramount.

Both active and passive methods have strengths and weaknesses when applied to 3D telepresence avatar generation. Multiple Kinect based approaches are currently of a lower resolution compared to that which can be achieved with SfS using conventional cameras with resolutions typically in excess of 1000x1000 compared to 320x240 pixels depth map resolution. They offer a less faithful reproduction because the holes produced due to pattern interference need to be filled and what fills them may not be a true representation of the real world. Moreover, there is a drop in quality of depth maps over distance [28] and this reduces the potential capture volume, which is not desirable for user movement or interacting with objects. SfS currently offers higher textural resolution and does not suffer from holes thus enabling clearer representation of eye gaze and facial expressions both of which are vital for portraying accurate NVC. Depth based approaches, however, can be deployed within an immersive environment where as, with the exception of [29], SfS requires a sterile background that would prevent the system to fully immerse the user.

1) Texturing

After capturing and reconstructing the user in 3D, texturing is needed to provide a life like representation. A composition of the segmented images of the different cameras is then used to generate the final texture. Our previous approaches to multi-view rendering used these images with no blending [6] and this resulted in a clear eye and face representation, but with undesired visible lines in the border of the different images. The use of image blending [30] improved the quality as there were no visible lines, but this process blurred the eyes, limiting the set of NVC that the reconstructed avatar was able to convey. Floating Textures [31] is a method that provides high quality results, however, it is complicated, and we found no evidence published of the quality of the eye reproduction. This suggests that a simple yet sufficiently effective method can be used.

C. Contextualizing a wide range of NVC by combing immersive displays and live reconstruction

Non-verbal communication is inherently spatial. Retaining this spatial context is highly challenging with today's displays and mediums. In particular much thought has to go into the way in which medium and display are combined. Both the medium and the display impact on the way in which space can be shared. This in turn impacts on the contextualization of NVC. Both [1] and [32] allow users to view each other but not physically walk into each other's space. The former shows the remote user within their space, whereas the latter shows them in a simulated space.

Immersive displays can vary from HMDs to large immersive projection-based displays. Unfortunately, not all of them are appropriate given the requirements of the system proposed. HMD's have been combined with video based reconstruction [33] but this completely hides eye gaze. Immersive projection technology usually uses 3D glasses, which at best make eyes hard to see [32]. In the closest studies identified attempting to support collaborative meetings, two users, captured using a single Kinect [34] and two Kinects [35], were reconstructed in front of a collaborative whiteboard, allowing visual communication of NVC and written notes from a fixed perspective. Our study uses 10 cameras to capture the user, and thus it is possible to walk completely around them while they are contextualized within a much larger synthetic environment, in a similar way to [36].

1) Foreground segmentation from a background containing moving objects or images.

Sharing a completely simulated space requires that people but not their surrounds are transmitted. Different approaches to background-foreground segmentation include simple background subtraction [37], Chroma Keying [38] and more advanced background-foreground detection methods such as [39][40][41]. The choice of background-foreground segmentation method impacts on the faithfulness of the reconstruction.

Our initial approach to combine immersive projection technology with free viewpoint video was inspired by the BBC [42]. This used a retro-reflective material to allow the user but not surrounding cameras to see a projected image. However, this proved to have a number of drawbacks. Firstly, the material does not allow projection from the rear. The projection quality is lower due to material properties. Lastly, the retro-reflective qualities of the material require many projectors to support viewing across a typical display volume. We then developed a solution that segmented a background of unified color [43]. However, this limits the user to looking into rather than sharing other's space, as in [1] and [32]. Another consequence of the need for a unified background color is that the solution is not readily deployable to most simulation facilities. It is desirable to be able to subtract backgrounds comprised of both static objects and moving images. The work toward a solution is described later in this paper.

III. Collaborative Mixed Reality System

The collaborative MR system is realized via combination of

Copyright (c) 2016 IEEE. Personal use of this material is permitted.
However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2580425, IEEE Transactions on Circuits and Systems for Video Technology

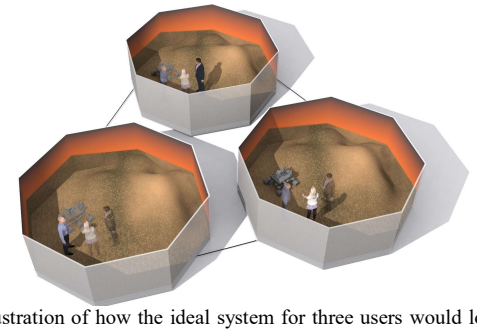

Fig. 1. Illustration of how the ideal system for three users would look like: every user would have one immersive display and all of them would share the same virtual space where the others are faithfully reconstructed in 3D.

enhancements to an existing telepresence system [1] followed by its integration with the Mars simulator. First, this section presents an ideal immersive projection telepresence system, and then the telepresence and Mars simulator systems developed to date, and finally their integration.

A. Ideal immersive projection system

The following system would require that the most advanced display and capture equipment was available at all sites. With hardware extended to support real time segmentation of users against live simulated backgrounds across the entire volume of each display. It would also feature stereo display enabling eye gaze clearly visible at three meters. Finally, a variety of the multidimensional datasets, with different spatial and temporal resolutions, would be available for Mars, including Rover simulation capabilities.

Using this system, each user would be able to move around the Mars simulation within the extent of social space with others, whether in the same or distributed locations. Enabling natural movement around, for example, a Mars Rover, bringing users' attention to attributes of it and the surrounding environment. Fig. 1 depicts this ideal situation.

B. Current systems

This section describes the current state of both the telepresence system and the Mars simulator paying special attention in the updates carried out to meet the requirements of the MR system.

In the context of CROSS DRIVE, there is only one fully immersive display available, the octave [44], the rest are Powerwalls and desktops computers. This results in only one user being able to walk 360 degrees around an object while still retaining eye contact.

Furthermore, the octave is the only one equipped for multi-camera video capture. However, segmentation of the projected simulation is currently not supported across the entire space. This restricts projection of simulation to a single wall but outside of the view of the cameras.

1) Telepresence system

An update on the 3D telepresence research system *withyou* [1] is presented in this paper. First, the end-to-end system architecture is detailed followed by a description of extensions, with justifications, that have been made.

The complete end-to-end system architecture is comprised of multiple network connected components with each contributing to the processing pipeline that is originally

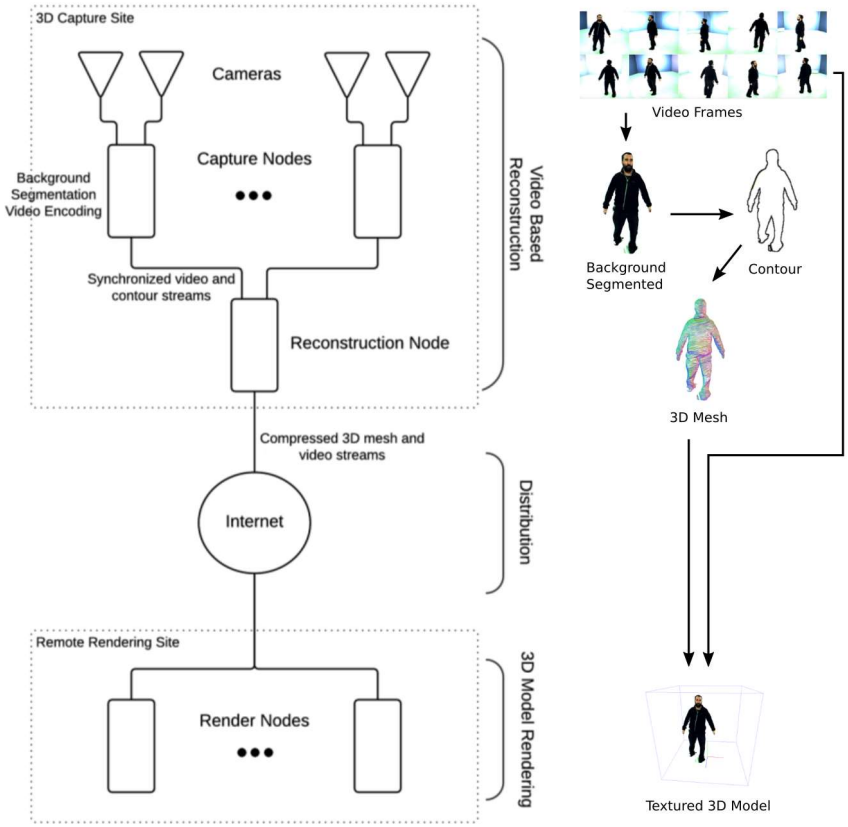

Fig. 2. Generic architecture of our telepresence system.

Copyright (c) 2016 IEEE. Personal use of this material is permitted.
However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2580425, IEEE Transactions on Circuits and Systems for Video Technology

described in depth in [1] and summarized in Fig. 2. This figure depicts the high level visualization of the telepresence system architecture from the cameras that acquire the subject(s) (top) through to rendering on the end nodes, which can be in different and possibly geographically dispersed locations (bottom). The following subsections outline the whole process.

a) Video Based Reconstruction

The outcome of this first stage of the process is the 3D mesh that represents the user together with the video streams that will be used to texture it. This includes the subject acquisition, the background-foreground segmentation and the 3D model generation.

(1) Image Acquisition

The process begins with the acquisition of images of the subject(s) via an array of cameras surrounding them. Cameras are either mounted on tripods or above the displays depending on the display configuration.

(2) Background-foreground Segmentation

A particular challenge of this work is that a user may be stood against a background containing static or moving images. Currently, there are two implementations of segmentation, the first is fully implemented and the second partially.

- Segmentation in the visible light spectrum: In previous publications, the system utilized a GPU Mester based background-foreground segmentation method [45]. However, the new requirement of a more faithful 3D reconstruction posed by the CROSS DRIVE project was not met. Also from a practical perspective, the previous method required domain specific knowledge to configure each time a camera was repositioned so experimentation with different setups [6] was a painstaking process. To improve the faithfulness of the avatar and alleviate the configuration constraint, the system was enhanced with a GPU implementation of Gaussian Mixture-based segmentation [41]. The shadow detection [46] has also assisted especially round the feet of the user. Moreover, in the sterile environment of the octave, it requires no domain specific knowledge to configure, thus enabling researchers to change camera positions without reconfiguring.
- Segmentation in the infrared (IR) light spectrum: Creating a 3D reconstruction of a user while she is immersed in an environment with a moving background is challenging. Regarding the segmentation in the octave, one of the problems with VBR in a fully immersive display system is that the user is surrounded by the display and thus the segmentation method employed needs to extract their silhouette from a moving background. As has already been commented in Section 2, this posed a limitation in the octave resulting in one screen being used for the display, reducing the expressiveness of the 3D avatar (the user was not able to point or look at things out of the screen). In an attempt to solve this limitation, a solution utilizing the IR light spectrum to perform segmentation is being considered. The subject stands in the middle of an immersive display and is illuminated with IR light from strategically positioned surrounding lamps. The user and surrounding background are acquired by cameras that are only receptive to light in the same frequency as the light emitted by the lamps. The cameras only capture the objects illuminated in IR light and nothing projected on the screens, thus, the moving background no longer interferes with the segmentation. The IR camera is physically positioned in close proximity with a visible light camera pair and its pose in relation to it determined using the checkerboard calibration technique.

The results of these two approaches are presented in Section IV where their impact on the quality of the 3D reconstructed avatar and the combination of 3D reconstruction and immersive displays is shown.

(3) 3D Model Generation

Upon receiving and decoding the video streams and contour data from the capture nodes, the reconstruction node generates a 3D model avatar via a parallelized "Exact Polyhedral Visual Hulls" (EPVH [27]) implementation [47]. To generate a 3D model the system requires knowledge of the cameras image planes in relation to real word 3D coordinates [48].

b) Distribution

In the *withyou* system, both model generation and rendering were executed on the 3D Reconstruction node and this limited the practical usability of the system. To overcome this restriction, a new method of distributed rendering was proposed and implemented. The rendering process was detached from the 3D reconstruction component and placed in its own self-contained client. To allow for multiple remote rendering sites, the new renderer is network enabled. After the 3D model generation, the reconstruction node prepares the 3D mesh and video data for broadcasting to connected remote rendering sites, packaging all relevant data into a network message. A message contains vertex positions, triangle indices, a video frame per camera, as well as frame number and timestamp. In order to reduce the amount of data sent across the network, the 3D mesh is compressed using the LZMA algorithm [49] after serialization, and this results in between 67% and 75% reduction in size. The h.264 encoded video is taken directly from the input of the capture nodes to avoid decompression and recompression by the reconstruction component. Synchronization of the video and mesh data is handled by placing the data together in the same network message.

c) 3D Model Rendering

A new texturing method is implemented with the aim of removing visible lines at polygon joins without confusing the image through blending. Another goal was to test if this could be achieved with an approach simpler than [31].

The render node decompresses the incoming geometry mesh data and computes vertex normals via the weighted average of the angle between connected triangle edges. It then pushes the vertex positions, normals and triangle indices into OpenGL buffers on the GPU. The compressed video frames are decoded and pushed directly onto texture buffers on the GPU.

Copyright (c) 2016 IEEE. Personal use of this material is permitted.
However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2580425, IEEE Transactions on Circuits and Systems for Video Technology

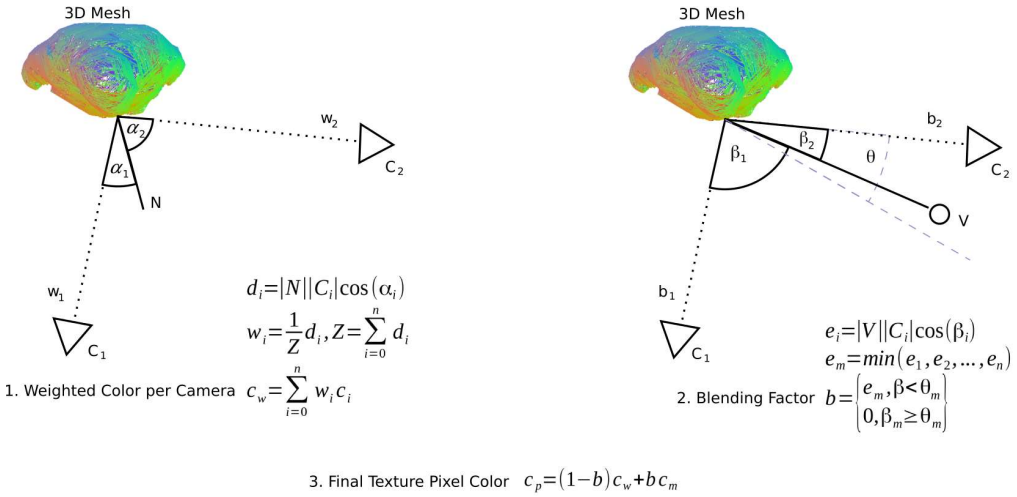

Fig. 3. Computing the final pixel color of the texture by combining weight based blending with viewpoint dependent blending of projected camera images. 1.) Weighted blending based on the angle between surface normal N and camera vector C_i, where w_i is the weight of a camera, c_i is the projected pixel color of a camera, and n is the number of cameras. 2.) Viewpoint-dependent blending based on the angle between the viewpoint vector V and vector of the closest camera, where e_m is the smallest angle and θ is a threshold. 3.) The final color is the combination of the weighted color and the color of the closest camera to the viewpoint blended-in, where c_p is the resulting pixel color, c_w is the computed weighted color of a camera, c_m is the projected pixel color of the camera closest to the viewpoint.

Texturing is realized in a pixel shader program in which each texture is projected onto the mesh from the corresponding camera perspective. The algorithm computes the color of a pixel based on a weighted blending of projected pixels from the camera images. The blending weights w are determined by computing the dot product between the fragment normal N and the direction from the fragment to a camera C, so that $w = 1$ with $\alpha = 0°$ and $w = 0$ with $\alpha \geq 90°$. The weights are then normalized so that their sum equals one and applied when adding the projected pixel colors of the respective camera images, see Fig. 3, 1.

The weighted blending method provides that surfaces facing closer toward a specific camera receive a higher contribution to the final pixel color from this camera's image than from others. The result is a smooth blending of the projected textures. While this method is simple and does not require on a specific camera arrangement, it can cause distortions in areas without a dominant camera and where cameras have similar weights. Furthermore, it does not take occluded areas into account.

In order to further improve visual quality, the texture mapping algorithm has been extended with a viewpoint-based blending method, where a camera image that was captured from a direction close to the current viewing direction of the user has higher influence than the color determined via the surface normals as described above. The algorithm starts with finding the closest camera by comparing the angles between the camera directions (vector from surface to camera) and the direction to the current viewpoint (vector from surface to viewer), see Fig. 3, 2. If the smallest angle is below a threshold, then the image of this camera is blended over the previously computed texture, see Fig. 3, 3. The blending factor is inversely proportional to the angle between the closest camera and viewer direction and ranges from zero to one. Smaller angles produce a higher blend-in factor and an angle of zero results in fully displaying the pixel of the closest camera.

A suitable choice for the threshold is influenced by the arrangement of capture cameras and the preference of blending behavior. A narrow threshold causes the texture to fade-in only when the viewer is very close to a camera view, whereas a large threshold causes the texture to fade-in from a larger distance. In our setup, a threshold of 12 degrees was chosen.

The result is shown in Fig. 4. The combination of both texture mapping methods provides the best compromise of

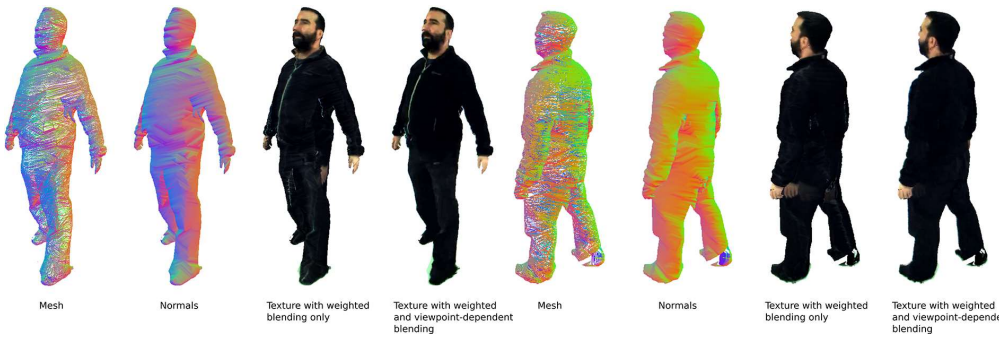

Fig. 4. Stages of 3D model rendering: incoming mesh; normal generation; texture generation via weighted blending of projected camera images; and blending of the image of the camera that is close to the user's viewpoint.

Copyright (c) 2016 IEEE. Personal use of this material is permitted.
However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2580425, IEEE Transactions on Circuits and Systems for Video Technology

computation effort and visual quality for our system. Although, the viewpoint-based blending technique is only effective when the viewer looks at the reconstructed mesh from near a camera view, it significantly improves the visual quality at these occasions. For example, the collar is correctly colored in the front view and the ear is rendered with a shadow, see Fig. 4. As our texture mapping technique does not test for visibility of surfaces to cameras, the viewpoint-based blending method has a further advantage, as it hides wrongly applied pixels to occluded areas. For example, with the simple weighted blending method based on surface normals, the pixels of the hand captured by the camera to the left of the subject are mapped onto an area of the reconstructed mesh (near the hip), see Fig. 4. By applying the viewpoint-based texture mapping method, the image of the front camera is blended over the weights-based texture and the projected hand on the hip disappears.

2) Mars simulator

At this stage of the CROSS DRIVE project, only geology datasets have been integrated into the Mars simulator, making it possible to study the surface of Mars with different digital terrain model (DTM) resolutions and even using subsurface data obtained from subsurface sounding radar.

Therefore, the simulator is based on a VR cartography system designed for interactive exploration and analysis of a planet's surface within immersive virtual environments [2]. The system is capable of visualizing very large DTM datasets at interactive frame rates while assuring that the best available resolution is always shown. The renderer creates DTMs from geo-referenced raster data and provides interactive tools commonly found in Geo-Information Systems (GIS).

C. Integration

The Mars simulator provides interactive geology tools and supports collaboration between remote people within the immersive virtual environment using traditional authored avatars. In this paper, it is used to explore the surface of Mars. Each participating site runs an instance of the simulator with the Mars data stored locally. User interactions are then synchronized by a collaboration manager created within the CROSS DRIVE project. In its default configuration, the instances use a traditional CGI character as an avatar to represent remote users in the virtual environment. The Mars simulator has then been extended for supporting 3D video avatars. The extensions include a communication module for receiving the 3D mesh and video stream, as well as a rendering module for visualizing the 3D reconstructed avatar within the Mars terrain renderer.

When a participant joins a collaborative 3D Mars exploration session from inside a 3D capture space (octave), then this user will be represented by a 3D reconstructed avatar instead of the traditional CGI avatar. The system communicates the server address of the reconstruction node to the other participants, which open a connection to the reconstruction node directly for receiving the 3D mesh and video stream. The server starts streaming as soon as a remote client is connected.

The communication module runs decoupled of the rendering in a separate process, whereas the rendering module is triggered each rendering frame within the Mars simulator. If new data is available on receiving sites, the data is copied from the network message buffer; the mesh is uncompressed and the video decoded; and the rendering process, as described in Section B.1)c), is initiated.

The position of the local user watching the 3D reconstructed avatar, needed for the viewpoint-based texture mapping, is provided by the head tracking system of the immersive virtual environment.

Additionally, a transformation matrix has been added that, firstly, aligns the origin of the capture space, and thus the origin of the 3D reconstructed avatar, with the origin of the virtual world in the 3D Mars simulator, and secondly, scales the units in the capture space to match the units in the renderer space. This way, when moving around and pointing at references within the 3D Mars simulation in the capture space, the reconstructed avatar appears in life-size and at the corresponding position and orientation within the simulation in the remote virtual environments. With our current camera configuration we can capture and reconstruct people so that their gaze and facial expression are clear while they occupy any position within 1.5m radii from the center of the octave.

IV. Results

This section overviews the qualities of the system. It provides evidence toward validating the approach, although neither a perceptual or behavioral study is provided. However, we hope it provides sufficient evidence that such in-depth studies would now be achievable. We argue that our balanced approach to supporting interpersonal movement, gaze, facial expression and the integration of an application to encourage their use, opens the door to such experiments.

A. Visual quality and communication of NVC

Firstly, previous and new methods for segmenting in the visible light spectrum are compared.

Fig. 5 shows the finer granularity achieved with the new approach used. Notice less jaggedness and closer match to the actual form of the face (indentation at bridge of nose, mouth, hairline and chin) and better representation of the digits of the hand without webbing effect. This finer granularity results in the generation of more faithful avatars.

With this improvement, we have been able to demonstrate the systems capability to show gross NVC such as waving (Fig. 8), pointing and interpersonal distance (Fig. 6). In addition, the system is capable of capturing and displaying subtler NVC such as eye gaze and facial expressions. Fig. 7 shows the quality and clarity of facial expressions achievable with a good camera calibration. It illustrates the seven universal emotions described in [50]. Highlighting that the reconstruction quality is high enough to achieve this and in addition it shows quality of eye gaze captured. It should be noted that camera rig height can have an impact on reconstruction quality inducing a droop effect that can make a user appear sad, aged or unwell, which becomes worse as the

Copyright (c) 2016 IEEE. Personal use of this material is permitted.
However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2580425, IEEE Transactions on Circuits and Systems for Video Technology

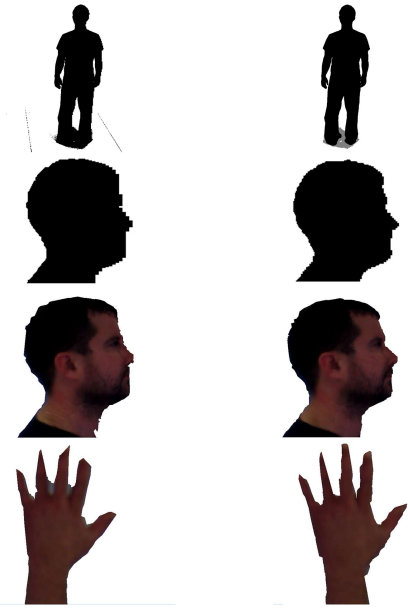

Fig. 5. Segmentation comparison of previous MESTER (left) and current Gaussian Mixture-based (right) implementations. Notice incorrectly classified regions and jagged edges that were present using the previous method (left) compared to the finer granularity, including shadow detection highlighted in grey, of the current method (right).

user approaches the outer limits of the capture space [1]. The camera setup for Fig. 7 appears to be set just right for the user being captured yet the cameras are all above the screens.

Apart from the quality of the reconstructed 3D model, the fact that users may have different hardware available to them can raise new issues. Fig. 9 shows three images of a 3D reconstructed avatar of a user. This highlights the issues of different display technologies when used with 3D reconstruction. If shutter glasses are used to enable stereo, then eye gaze is not captured and the facial muscles around the eyes are partially occluded. If a full HMD is used, then very little of the face is visible. The result is a complete lack of facial expressions. The system described in this paper is capable of utilizing both as we recognize that not every user will have a 3D capture system or immersive display.

B. *Segmentation from static or moving background*

A specific goal was to segment against background including moving images. This section shows the preliminary results of a new approach in IR light spectrum to allow segmentation across immersive displays and static backgrounds. For this proof of concept, we used Kinect as an IR camera, but it will be replaced for higher resolution cameras in further tests.

We have confirmed that both projection systems in the octave and one of the Powerwall are not emitting IR light at a frequency that interferes with the Kinect's IR camera Thus, it is possible to segment the user against a moving background with it. Although not tested, we have no reason to assume that the result in the other Powerwall would be different. Fig. 10 shows the preliminary results of experimenting with different IR emitters and lamp positions and the effect on segmentation results due to differences in scene illumination. It is clear that the result is best when two IR lamps are active (as shown in the bottom-left part of Fig. 10). Further experimentation is now required to determine if the addition of more IR lamps and perhaps cameras with greater resolution than the Kinect could improve the result.

C. *Temporal quality*

A quantitative temporal evaluation of the existing *withyou* platform is presented in [1]. As a summary, the time taken to acquire a sequence of frames from 10 cameras then segment, encode video and reconstruct the 3D mesh is 79.32ms. This section presents some results of the new distribution and rendering subsystem (Table I) followed by the end-to-end latencies observed during the linkups.

The first stage is mesh compression, which is currently achieved using the LZMA algorithm. This is followed by serializing the compressed mesh, timestamp and encoded video frames. Table I presents times for processes that are required to distribute and render the participants at the remote sites.

The packet format and sizes are shown in Table II. The packets are then distributed to the render clients via network connection. Upon receiving the packets, the video frames are decoded and the mesh decompressed.

Whilst conducting the linkup, the authors carried out some latency tests for the system and repeated with Skype for comparison. The results are shown below in Table III.

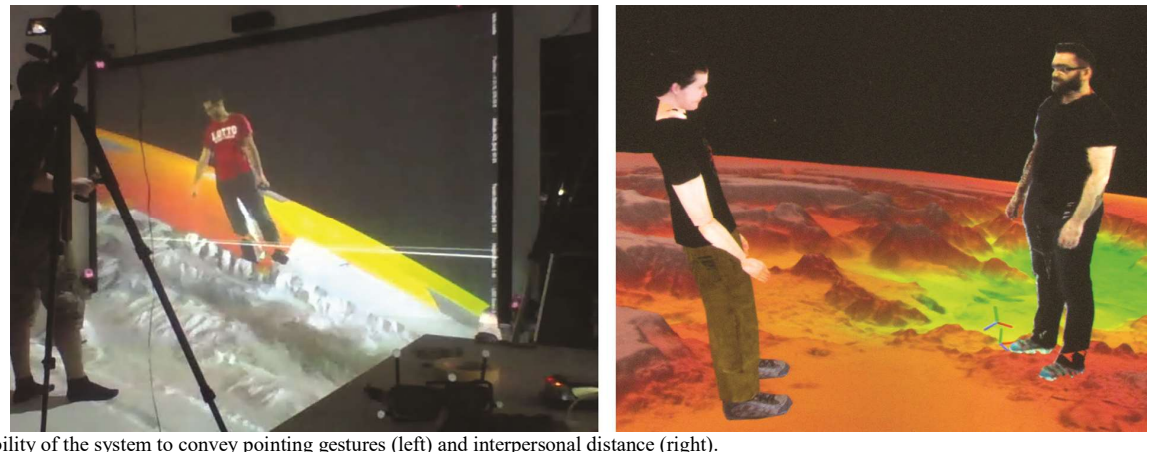

Fig. 6. Ability of the system to convey pointing gestures (left) and interpersonal distance (right).

Copyright (c) 2016 IEEE. Personal use of this material is permitted.
However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2580425, IEEE Transactions on Circuits and Systems for Video Technology

Fig. 7. Universal facial expressions of emotion and eye gaze. This figure shows a 3D reconstruction of an author attempting to display the seven universal emotions and quality of eye gaze in three directions.

TABLE I
DISTRIBUTION AND RENDERING SUBSYSTEM TIMINGS

Process	Time (ms)
Mesh compression	118.5
Mesh decompression	23.4
Video decompression (10 frames)	16.1
Upload textures to GPU	105.2
Upload mesh to GPU	27.3
Render	0.05

The update to the segmentation procedure described in Section III.B resulted in an improvement in the accuracy and thus in the 3D form of the reconstructed avatar. However, it is currently hindered by our aging hardware. As a consequence, a reduction in the temporal quality of the system has been experienced. With newer hardware the temporal issue could be resolved thus enabling overall improvements. Table III demonstrates this by presenting the mean time taken to segment a sequence of 50 images on a number of different GPUs. The GeForce GT 730 is the card currently deployed in the reconstruction system.

Another aspect influencing the temporal qualities of the system is the streaming of the reconstructed avatar due to the high requirements on bandwidth. This is due to the fact that the 3D geometry and several HD video frames are streamed across the network. Although simple mesh compression and fast h.264 video are used, relatively low framerates are currently achieved. In order to avoid sending large amounts of data to several remote users, a proxy server was set up at the site in Germany. This reduced the traffic to a single stream from the UK to Germany and allowed us to distribute the 3D video stream to users inside the LAN and to the cluster nodes of our multi-pipe visualization system. However, this leads to increased delay. More advanced compression and data reduction methods are necessary to reach high framerates.

TABLE II
TYPICAL PACKET COMPONENT SIZES

Item	Size (bytes)
Calibration data	740
Encoded video frames	9258
(10 cameras)	(average of 100 frames)
Compressed vertices	84745
(average number of vertices 9214)	(average of 100 frames)
Compressed triangles	47266
(average number of triangles 52607)	(average of 100 frames)

TABLE III
LATENCY OBSERVED DURING THE LINKUPS

	Local (Octave Salford)	THINKlab (Salford)	DLR (Germany)
3D Reconstruction	1.06s	1.12s	1.5s
Skype	-	0.103s	-

D. *Initial linkup test*

This section summarizes the initial connection tests of the system across Europe.

In this linkup, the surface of Mars is explored using elevation and imagery data from NASA' Mars Reconnaissance Orbiter (MOLA data, 500m/pixel) and ESA's Mars Express (HRSC data, 12m/pixel) with a data volume of more than 600GB.

In contrast to the ideal immersive projection system depicted in Fig. 1, the current system deployed for this initial test does not feature immersive systems that surround the users. Fig. 11 illustrate this current configuration that shows three users, only one of them is 3D reconstructed (the person in the octave) while the others are represented by traditional motion-driven avatars.

Fig. 12 shows the components and interconnections of the three sites that were linked in the test, including the configuration of the capture system at the octave. Four different users took part, over three sites, two connecting from each country (see Fig. 8). Since only one site is currently able to generate and stream 3D video avatars, the rest of them were represented by traditional authored avatars. The second site, also in Salford, was the ThinkLab, using a stereo Powerwall and an optical tracking system. The third site, DLR (Germany), had one user connected using a stereo Powerwall with floor extension and optical tracking system, and the other using a desktop system.

TABLE IV
COMPARISON OF SEGMENTATION TIMES WHEN PROCESSING TWO STREAMS SIMULTANEOUSLY USING VARIOUS GRAPHICS CARDS

GeForce GT 730	GeForce GTX 660	Quadro K5000	GeForce GTX 970
27.17ms	5.79ms	5.31ms	2.82ms

Copyright (c) 2016 IEEE. Personal use of this material is permitted.
However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2580425, IEEE Transactions on Circuits and Systems for Video Technology

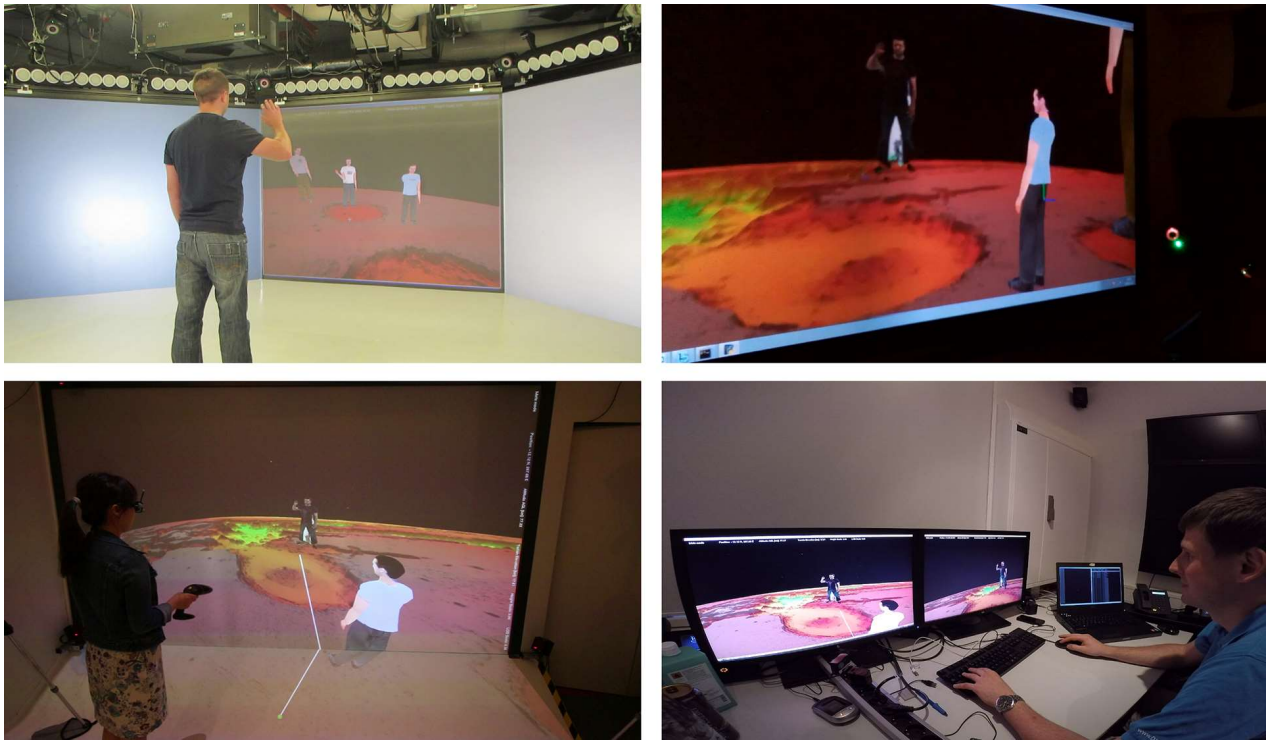

Fig. 8. Users greeting each other from Mars. This figure shows a user waving at three other users and his 3D reconstructed avatar from the different sites that took part in the experiment. The top left picture shows a user in the octave waving at three traditional avatars. The top right picture shows the 3D reconstructed avatar waving in the ThinkLab´s Powerwall (UK). The bottom left and right show this action as it was viewed by the Powerwall and the desktop system in DLR (Germany). The stereo view was removed from the two Powerwalls to take the pictures.

Fig. 8 shows four pictures showing users greeting each other from Mars. The first picture shows a user waving at the rest from the octave, whereas the other three show this action performed by the 3D reconstructed avatar through the viewpoint of each of the other users. The participants could freely navigate and talk to each other. Fig. 13depicts how the 3D reconstructed user is viewed from different viewpoints as one user moves around him. This free viewpoint navigation offers the user interactive functionality over and above that offered by traditional video conferencing, allowing users to view NVC from any angle.

The 3D avatar is streamed from the octave, but instead of sending it directly to each user, a proxy was used in DLR to send it to the desktop client, reducing the bandwidth needed in the octave.

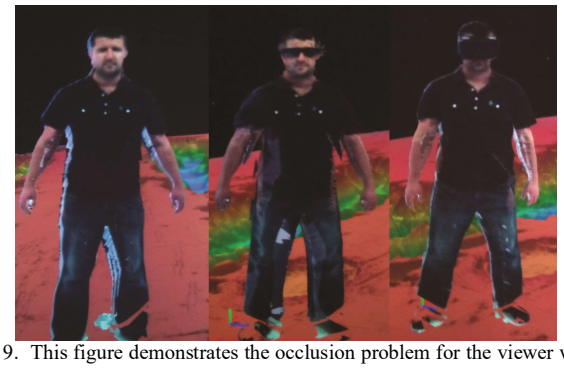

Fig. 9. This figure demonstrates the occlusion problem for the viewer when using various display devices. Left no stereo (full facial expressions), center stereo glasses (eye gaze and some facial features obscured) and right HMD (Most of the face obscured, identification of facial expressions not possible).

V. Discussion

This paper presented the integration of the *withyou* telepresence system and a Mars simulator that will allow scientists in remote locations to collaborate whenever necessary (saving travel time and money), while simultaneously exploring data sets. The former is designed to facilitate natural interplay between interpersonal distance, interactional eye gaze and a representative range of non-verbal signals (including facial expressions) associated with emotion, familiarity and trust. The latter provides a shared context and application where people in remote spaces can come to joint understanding and decisions concerning an environment. Both

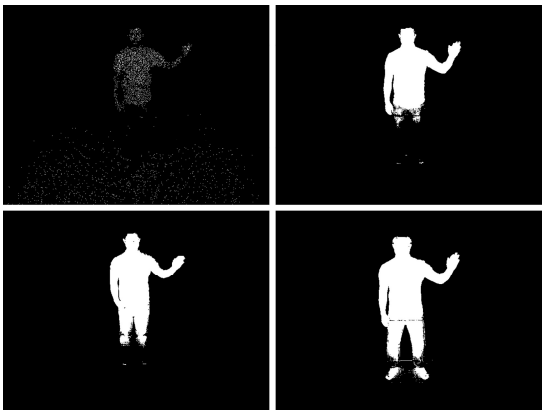

Fig. 10. Preliminary IR segmentation results. Top-left: Kinect IR projector, top-right: IR lamp mounted aligned with Kinect, bottom-left: IR lamp positioned on floor angled upwards and bottom-right: both IR lamps.

Copyright (c) 2016 IEEE. Personal use of this material is permitted.
However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2580425, IEEE Transactions on Circuits and Systems for Video Technology

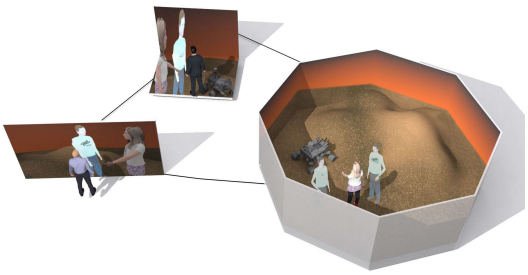

Fig. 11. Illustration of the current system: only one user has a fully immersive display and she is also the only one faithfully reconstructed in 3D. Note: The system tested utilized only one screen of the octave however the diagram reflects what is achievable with the new segmentation method.

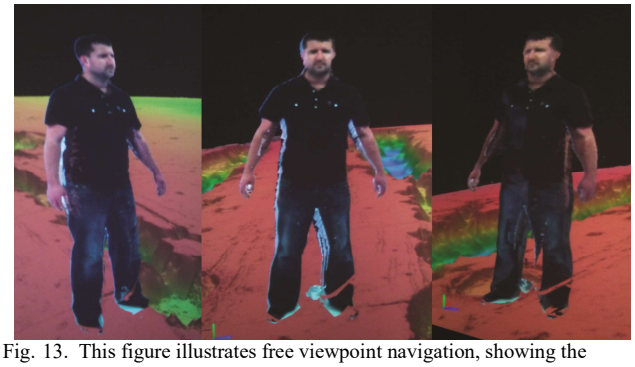

Fig. 13. This figure illustrates free viewpoint navigation, showing the capture subject as viewed from the left, centre and right of the display.

together form a testbed that could facilitate experimentation around social interaction and also provide a demonstration of this functionality in a remote collaborative simulation.

Most NVCs, such as interpersonal distance, interactional gaze and gestures of familiarity, are linked in social interaction. People obey these basic social rules when sharing a virtual environment with even a very simple virtual human [51]. However, this has not been tested with an avatar representation or with faithful reconstruction of identity and facial expression. This is not surprising as technology to support it has not been readily available. *Withyou* may be the first system able to support it. In addition eye gaze can be estimated to within the tolerances of social interaction from a video reconstructed avatar captured by cameras outside immediate social space [6] [47]. However, while support for each of these components had been tested, their linkage had not.

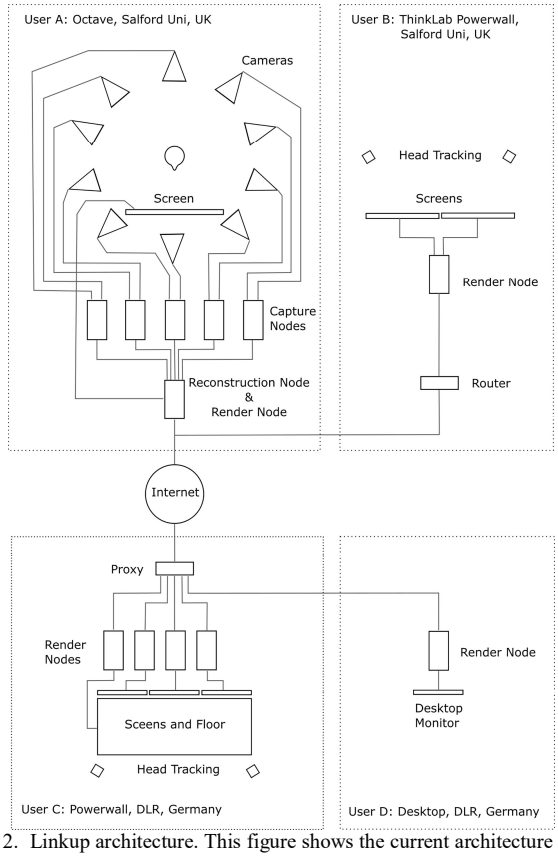

Fig. 12. Linkup architecture. This figure shows the current architecture of the initial test carried out.

This linkage between these non-verbal behaviors is associated with familiarity and trust and is used to mediate interactions, deciding for instance if a conversation should start and when it should end. Because of this today's communication technology is generally less effective in anything but round the table meetings, where people's movement is constrained to a seat. This kind of meeting has been conducted using telepresence [23], however, bringing someone's attention to something in the surroundings is hard when people cannot move around each other within a shared context. We argue that confusing or restricting spatiality of gaze and interpersonal distance holds back three applications of communication technology: the ad-hoc meeting; building of trust and rapport; and shared exploration of an environment, be it real, virtual or mixed. Supporting part of the CROSS DRIVE application by the integration of the Mars simulator has provided a case study and platform to test if these linkages are supported and ultimately if this makes a difference to task performance and experience. Fig. 13, demonstrates that a remote person can be viewed from any side without degradation of quality. This is unlike most many approaches that capture a person from predominantly one side [34][35].

Unfortunately, NVCs may lose their meaning if the shared space is not correctly orientated between participants. The linkup test showed a pragmatic approach to share the space between legacy display systems of distinct levels of immersion. In simple terms, displays that surround and immerse a user can be linked so that people can walk around each other. However, if any display does not surround the user, people can at most walk up to each other's avatars [47]. The legacy displays of the CROSS DRIVE partners were of both kinds. The pragmatic solution of projecting onto the floor in front of a wall display was used to provide a compromise. In this compromise, people can comfortably move within each other's social space without the need for those by a wall display to stand right up against it. Fig. 8, demonstrates this configuration, with one avatar just within and one outside natural social space. A seemingly "catch 22" problem is that 3D is needed to allow mutual eye gaze between moving people, yet 3D glasses or HMD occlude the eyes, Fig. 9. Our pragmatic solution to this is not to use stereo but rely instead

Copyright (c) 2016 IEEE. Personal use of this material is permitted.
However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2580425, IEEE Transactions on Circuits and Systems for Video Technology

on parallax to support mutual gaze. We had already shown that gaze could be accurately determined from a 3D model without stereo glasses [6] by participants rotating reconstructed humans until they appeared to look at them. Here we have tried it out for real, all be it not in a rigorous experiment.

Our approach provides a compromise between the qualities of video and VR. One of our goals is to balance visual, spatial and temporal qualities to the point where they can support the linkage between interpersonal distance, eye gaze and facial expression. The other goal is to provide an application that encourages people to move around and discus. Visual quality appears sufficient to communicate through the face: identity; gaze; and emotion, Fig. 7. The granularity of other non-verbal communication scales to clear finger gestures. In terms of spatial quality, our capture and reconstruction technique scales to full 360 degrees around the subject. While one of our display systems matches this, the others do not. Latency is much lower than what has been demonstrated by other full free viewpoint systems. However, it is still 1.5 seconds. This may impact on mutual eye gaze behaviour and role of non-verbal communication in conversational turn taking. Further improvements are being carried out by using a less conservative time management approach and we believe this will make possible to achieve latencies of around half this.

VI. Conclusions

We have presented the combination of the "withyou" telepresence system and a Mars simulator across visualization facilities in Germany and the UK. This pilot begins to demonstrate how space scientists, engineers and mission controllers could discuss Mars missions without the need to travel to single simulation facility in one of the participating countries. The contribution of this work is the combination of a telepresence system that allows spatial contextualization of a large range of Non-Verbal Communication (NVC), with a simulation that provides context and an application that demonstrates utility, while addressing a set of key challenges.

Currently communication technologies do not well support the linkage between interpersonal distance, mutual gaze and facial expression. A key problem is that NVC is inherently spatial yet retaining spatial context across these non-verbal resources is hard without diminishing the quality of some. With our approach, gaze, interpersonal distance, facial expression and other NVC can be communicated as people move together around a place of interest, such as a landing site. The ultimate aim is to allow people to efficiently and accurately communicate both what they are talking about and how they feel about it, within the context of a shared simulation. Today, this is much easier when people are physically together. In widening the set of NVC that can be contextualized within simulations shared across a distance, this work could ultimately impact across many domains.

An important contribution is opening the door to technology mediated social interaction in which the linkage between interpersonal distance, mutual gaze and facial expression are likely to play a role. Both the photographs in this article and the supporting video provide evidence that people observe natural rules of interpersonal distance (e.g. Fig. 6) and gestures to bring them to it (e.g. Fig. 6 and Fig. 8). However, only one of the avatars was reconstructed from video. The others were conventional CGI avatars, two driven from motion tracking and another one from a desktop interface. This is the only system to our knowledge that has both avatars created using video reconstruction and authored CGI models following motion tracked data. This might allow comparison of how the two approaches impact on the level of NVC supported during social interaction and outcomes such as trust, rapport, team cohesion and task performance.

A practical contribution is the demonstration of pragmatic approaches to share space between legacy display systems of distinct levels of immersion. For example, we showed how a wall display was extended with floor projection in order to improve the support for NVC in the social space of the users. The different displays allow the impact of display to be studied for the first time with video reconstructed avatars.

A key novel technological contribution of this article is a new method for background segmentation. Segmentation allows a person to be captured without their surroundings so that they can be "imported" live into a shared virtual context. Previous methods supported segmentation against plain color or static backgrounds. However, the legacy displays used by the CROSS DRIVE partners had distinct levels of immersion. This meant that some people needed to be segmented from a moving CGI background, some from a static background, and others from a combination of the two. While the principle and technical feasibility has been demonstrated, a complete solution has not yet been implemented. This is simply as doing so requires the purchase of more of the equipment that we have already used. Specifically, the current implementation covers only part of our largest immersive display, needing the reminder to be turned off.

The rigor of this work is in the iterative steering of technology development from psychological principles. Other work has tended to focus on the support of subsets of non-verbal communication necessary for particular classes of interaction. We are not aware of work from other groups that has looked specifically at supporting the all-important linkage between interpersonal distance, mutual gaze and facial expression. Furthermore, we have addressed an unprecedented spread of issues to balance visual, temporal and spatial qualities. We further argue that this work will contribute to the rigor of future work by allowing us providing more ecologically valid social interaction experimentation.

Immediate impact includes demonstration to the space science community, how such technology could improve distributed team cohesion and reduce cost of international collaboration. This approach could be implemented in many other fields that require remote participants to discuss information or models that they need to move around together, especially where emotions are part of the conversation Joint emergency services command and control of a disaster scene is a good example of where both spatial context and strength of feeling need to be communicated together. Health

Copyright (c) 2016 IEEE. Personal use of this material is permitted.
However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2580425, IEEE Transactions on Circuits and Systems for Video Technology

applications could include remote exposure therapy and a better understanding of the importance of linked non-verbal cues, in interactions with virtual humans during training and self-treatment. However, the widest impact may come from adding knowledge to telepresence research, on the conditions that need to be met before the above linkage plays its proper role in starting and mediating conversations. Understanding this could lead to general rather than niche approaches to bringing people together across a distance. This could radically reduce dependency on travel and improve quality of life.

APPENDIX

Video footage of the linkup test can be found in the following URL: https://vimeo.com/141524309

ACKNOWLEDGMENT

We thank Johannes Hummel, Fang Chen, Wito Engelke and Andreas Gerndt from DLR, and John O'Hare from the University of Salford for their support in the work described in this paper.

REFERENCES

[1] D. J. Roberts, A. J. Fairchild, S. P. Campion, J. O'Hare, C. M. Moore, R. Aspin, *et al.*, "withyou—An Experimental End-to-End Telepresence System Using Video-Based Reconstruction," *Selected Topics in Signal Processing, IEEE Journal of,* vol. 9, pp. 562-574, 2015.

[2] R. Westerteiger, A. Gerndt, and B. Hamann, "Spherical Terrain Rendering using the hierarchical HEALPix grid," 2011.

[3] A. Garcia, D. Roberts, T. Fernando, C. Bar, R. Wolff, J. Dodiya, *et al.*, "A collaborative workspace architecture for strengthening collaboration among space scientists," in *Aerospace* 2015.

[4] P. Milgram and F. Kishino, "A Taxonomy of Mixed Reality Visual-Displays," *Ieice Transactions on Information and Systems,* vol. E77d, pp. 1321-1329, Dec 1994.

[5] M. L. Patterson, "An arousal model of interpersonal intimacy," *Psychological Review,* vol. 83, p. 235, 1976.

[6] D. J. Roberts, J. Rae, T. W. Duckworth, C. M. Moore, and R. Aspin, "Estimating the gaze of a virtuality human," *Visualization and Computer Graphics, IEEE Transactions on,* vol. 19, pp. 681-690, 2013.

[7] D. Roberts, R. Wolff, O. Otto, and A. Steed, "Constructing a Gazebo: supporting teamwork in a tightly coupled, distributed task in virtual reality," *Presence: Teleoperators and Virtual Environments,* vol. 12, pp. 644-657, 2003.

[8] V. Vinayagamoorthy, M. Garau, A. Steed, and M. Slater, "An eye gaze model for dyadic interaction in an immersive virtual environment: Practice and experience," in *Computer Graphics Forum*, 2004, pp. 1-11.

[9] M. Slater, A. Rovira, R. Southern, D. Swapp, J. J. Zhang, C. Campbell, *et al.*, "Bystander responses to a violent incident in an immersive virtual environment," *PloS one,* vol. 8, p. e52766, 2013.

[10] Faceshift. (2015, 01/09/15). *Markerless facial motion tracking system [Online]*. Available: http://www.faceshift.com/

[11] Vicon. (2015, 01/09/15). *Vicon optical motion tracking system [Online]*. Available: http://www.vicon.com/

[12] H. Fuchs, G. Bishop, K. Arthur, L. McMillan, R. Bajcsy, S. Lee, *et al.*, "Virtual space teleconferencing using a sea of cameras," in *Proc. First International Conference on Medical Robotics and Computer Assisted Surgery*, 1994.

[13] 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," in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 1998, pp. 179-188.

[14] D. Roberts, R. Wolff, J. Rae, A. Steed, R. Aspin, M. McIntyre, *et al.*, "Communicating eye-gaze across a distance: Comparing an eye-gaze enabled immersive collaborative virtual environment, aligned video conferencing, and being together," in *Virtual reality conference, 2009. VR 2009. IEEE*, 2009, pp. 135-142.

[15] M. Hansard, S. Lee, O. Choi, and R. P. Horaud, *Time-of-flight cameras: principles, methods and applications*: Springer Science & Business Media, 2012.

[16] P. M. Will and K. S. Pennington, "Grid coding: a preprocessing technique for robot and machine vision," presented at the Proceedings of the 2nd international joint conference on Artificial intelligence, London, England, 1971.

[17] G. C. Stockman, S. W. Chen, G. Hu, and N. Shrikhande, "Sensing and recognition of rigid objects using structured light," *Control Systems Magazine, IEEE,* vol. 8, pp. 14-22, 1988.

[18] A. Maimone and H. Fuchs, "Reducing interference between multiple structured light depth sensors using motion," in *Virtual Reality Short Papers and Posters (VRW), 2012 IEEE*, 2012, pp. 51-54.

[19] A. Maimone and H. Fuchs, "A First Look at a Telepresence System with Room-Sized Real-Time 3D Capture and Large Tracked Display," presented at the International Conference on Artificial Reality and Telexistence (ICAT), Osaka (Japan), 2011.

[20] A. Maimone and H. Fuchs, "Real-Time Volumetric 3D Capture of Room-Sized Scenes for Telepresence," presented at the Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), Zurich (Switzerland), 2012.

[21] A. Maimone and H. Fuchs, "Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras," presented at the Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on, 2011.

[22] D. S. Alexiadis, D. Zarpalas, and P. Daras, "Real-time, full 3-D reconstruction of moving foreground objects from multiple consumer depth cameras," *Multimedia, IEEE Transactions on,* vol. 15, pp. 339-358, 2013.

[23] C. Zhang, Q. Cai, P. Chou, Z. Zhang, and R. Martin-Brualla, "Viewport: A distributed, immersive teleconferencing system with infrared dot pattern," *MultiMedia, IEEE,* vol. 20, pp. 17-27, 2013.

[24] P. E. Debevec, C. J. Taylor, and J. Malik, "Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach," presented at the Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, 1996.

[25] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 32, pp. 1362-1376, 2010.

[26] B. G. Baumgart, "A polyhedron representation for computer vision," in *Proceedings of the May 19-22, 1975, national computer conference and exposition*, 1975, pp. 589-596.

[27] J.-S. Franco and E. Boyer, "Exact polyhedral visual hulls," in *British Machine Vision Conference (BMVC'03)*, 2003, pp. 329--338.

[28] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan, "Scanning 3d full human bodies using kinects," *Visualization and Computer Graphics, IEEE Transactions on,* vol. 18, pp. 643-650, 2012.

[29] S.-Y. Lee, I.-J. Kim, S. C. Ahn, H. Ko, M.-T. Lim, and H.-G. Kim, "Real time 3D avatar for interactive mixed reality," in *Proceedings of the 2004 ACM SIGGRAPH international conference on Virtual Reality continuum and its applications in industry*, 2004, pp. 75-80.

[30] R. Aspin and D. Roberts, "Projective multi-texturing for integrated real-time 3D reconstruction and rendering of a person," 2011.

[31] M. Eisemann, B. De Decker, M. Magnor, P. Bekaert, E. De Aguiar, N. Ahmed, *et al.*, "Floating textures," in *Computer Graphics Forum*, 2008, pp. 409-418.

[32] S. Beck, A. Kunert, A. Kulik, and B. Froehlich, "Immersive group-to-group telepresence," *Visualization and Computer Graphics, IEEE Transactions on,* vol. 19, pp. 616-625, 2013.

[33] B. Petit, T. Dupeux, B. Bossavit, J. Legaux, B. Raffin, E. Melin, *et al.*, "A 3d data intensive tele-immersive grid," in *Proceedings of the international conference on Multimedia*, 2010, pp. 1315-1318.

[34] K. Higuchi, Y. Chen, P. A. Chou, Z. Zhang, and Z. Liu, "ImmerseBoard: Immersive Telepresence Experience using a Digital Whiteboard," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 2383-2392.

Copyright (c) 2016 IEEE. Personal use of this material is permitted.
However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2580425, IEEE Transactions on Circuits and Systems for Video Technology

[35] J. Zillner, C. Rhemann, S. Izadi, and M. Haller, "3D-board: a whole-body remote collaborative whiteboard," in *Proceedings of the 27th annual ACM symposium on User interface software and technology*, 2014, pp. 471-479.

[36] R. Vasudevan, G. Kurillo, E. Lobaton, T. Bernardin, O. Kreylos, R. Bajcsy, *et al.*, "High-quality visualization for geographically distributed 3-d teleimmersive applications," *Multimedia, IEEE Transactions on,* vol. 13, pp. 573-584, 2011.

[37] Y. J. Benezeth, P.-M.; Emile, B.; Laurent, H.; Rosenberger, C., "Review and evaluation of commonly-implemented background subtraction algorithms," presented at the Pattern Recognition, 2008.

[38] C. Schultz, "Digital Keying Methods," ed. University of Bremen Center for Computing Technologies, 2006.

[39] L. Li, W. Huang, I. Y. Gu, and Q. Tian, "Foreground object detection from videos containing complex background," in *Proceedings of the eleventh ACM international conference on Multimedia*, 2003, pp. 2-10.

[40] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, pp. 28-31.

[41] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern recognition letters,* vol. 27, pp. 773-780, 2006.

[42] O. Grau, "Studio production system for dynamic 3D content," in *Visual Communications and Image Processing 2003*, 2003, pp. 80-89.

[43] C. Moore, T. Duckworth, R. Aspin, and D. Roberts, "Synchronization of images from multiple cameras to reconstruct a moving human," in *Distributed Simulation and Real Time Applications (DS-RT), 2010 IEEE/ACM 14th International Symposium on*, 2010, pp. 53-60.

[44] J. O'Hare. (2015, 01/09/15). *Octave - technical information, University of Salford [Online]*. Available: http://www.salford.ac.uk/computing-science-engineering/facilities/octave-technical-information

[45] A. Griesser, S. De Roeck, A. Neubeck, and L. Van Gool, "GPU-Based Foreground-Background Segmentation using an Extended Colinearity Criterion," in *Proceedings of Vision, Modeling, and Visualization (VMV) 2005*, 2005, pp. 319-326.

[46] A. Prati, I. Mikic, M. M. Trivedi, and R. Cucchiara, "Detecting moving shadows: algorithms and evaluation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 25, pp. 918-923, 2003.

[47] T. Duckworth and D. J. Roberts, "Parallel processing for real-time 3D reconstruction from video streams," *Journal of Real-Time Image Processing,* vol. 9, pp. 427-445, 2014.

[48] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *Robotics and Automation, IEEE Journal of,* vol. 3, pp. 323-344, 1987.

[49] 7-ZIP. (2015, 01/09/15). *LZMA algorithm [Online]*.

[50] P. Ekman and D. Matsumoto, "Facial expression analysis," *Scholarpedia,* vol. 3, p. 4237, 2008.

[51] J. N. Bailenson, J. Blascovich, A. C. Beall, and J. M. Loomis, "Equilibrium theory revisited: Mutual gaze and personal space in virtual environments," *Presence,* vol. 10, pp. 583-598, 2001.

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

Simon P. Campion is a Project Manager and 3D generalist at the THINKLab, where he manages a portfolio of commercial Virtual Reality projects. He is also undertaking a part time PhD under the supervision of David Roberts. His PhD is assessing the impact of medium and interface on non-verbal communication.

Arturo S. García is a Research Fellow at the THINKLab. He received his European PhD in Computer Science at the University of Castilla-La Mancha (UCLM), Spain, in 2010. His research interests include the development of Collaborative Virtual Environments (CVEs) and interaction in VEs and CVEs.

Robin Wolff leads the 3D Interaction group at Simulation and Software Technology at the German Aerospace Center (DLR), where he has worked as senior researcher since 2010. He received a PhD in Immersive Collaborative Virtual Environments at the University of Salford, UK, in 2007 and was working there in several research projects.

Terrence Fernando is the Director for the ThinkLab. He has a broad background in conducting multi-disciplinary research programmes involving large number of research teams in areas such as distributed virtual engineering, virtual building construction, driving simulations, virtual prototyping, urban simulation, and maintenance simulation. He was the technical director for the EU funded CoSpaces IP project which studied the challenges in creating a reference architecture that can support a range of collaborative environments. In this work he also investigated the challenges in developing co-located, distributed and mobile workspaces that can offer range of collaboration styles through innovative virtual interfaces.

David Roberts is a Professor of Telepresence at the university of Salford. He builds and studies the use of immersive VR, until now mostly for telepresence. c. 100 publications mostly in the area of telepresence or distributed simulation. The telepresence area of his work now focuses on video based reconstruction of humans in real time. He is part of the EU CROSS DRIVE project. David recently lead the EPSRC funded Eyecatching project which developed a telepresence approach that could support mutual eye gaze between moving people in different displays. He chaired IEEE Distributed Simulation and Real Time applications for 6 years.

Copyright (c) 2016 IEEE. Personal use of this material is permitted.
However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.