Estimating the Gaze of a Virtuality Human

David J. Roberts, John Rae, Tobias W. Duckworth, Carl M. Moore, and Rob Aspin

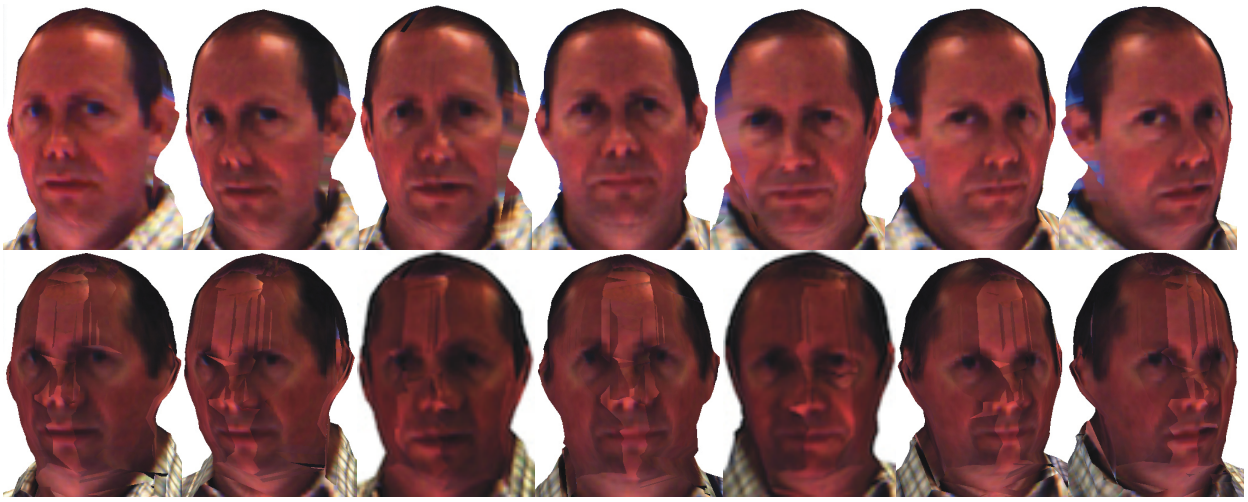

Fig. 1. In the experiment participants rotated static 3D virtuality humans (recorded with seven different combinations of eyes, head and body orientation, captured with a range of camera arrangements) until they felt that the virtuality human was gazing at them. The figure shows the point of gaze at the participant for two camera arrangements (one in the upper row, one in the lower row); all seven eye/head/body orientations are shown.

Abstract—The aim of our experiment is to determine if eye-gaze can be estimated from a virtuality human: to within the accuracies that underpin social interaction; and reliably across gaze poses and camera arrangements likely in every day settings. The scene is set by explaining why Immersive Virtuality Telepresence has the potential to meet the grand challenge of faithfully communicating both the appearance and the focus of attention of a remote human participant within a shared 3D computer-supported context. Within the experiment n=22 participants rotated static 3D virtuality humans, reconstructed from surround images, until they felt most looked at. The dependent variable was absolute angular error, which was compared to that underpinning social gaze behaviour in the natural world. Independent variables were 1) relative orientations of eye, head and body of captured subject; and 2) subset of cameras used to texture the form. Analysis looked for statistical and practical significance and qualitative corroborating evidence. The analysed results tell us much about the importance and detail of the relationship between gaze pose, method of video based reconstruction, and camera arrangement. They tell us that virtuality can reproduce gaze to an accuracy useful in social interaction, but with the adopted method of Video Based Reconstruction, this is highly dependent on combination of gaze pose and camera arrangement. This suggests changes in the VBR approach in order to allow more flexible camera arrangements. The work is of interest to those wanting to support expressive meetings that are both socially and spatially situated, and particular those using or building Immersive Virtuality Telepresence to accomplish this. It is also of relevance to the use of virtuality humans in applications ranging from the study of human interactions to gaming and the crossing of the stage line in films and TV.

Index Terms—Cinematography, virtual worlds, virtual environments, camera placement, hierarchical finite state machines

- *David J. Roberts is with the University of Salford. E-mail: d.j.roberts@salford.ac.uk.*
- *John Rae is with the University of Roehampton. E-mail: J.Rae@roehampton.ac.uk.*
- *Tobias W. Duckworth is with the University of Salford. E-mail: T.W.Duckworth@edu.salford.ac.uk.*
- *Carl M. Moore is with the University of Salford. E-mail: c.m.moore@edu.salford.ac.uk.*
- *Rob Aspin is with the University of Salford. E-mail: r.aspin@salford.ac.uk.*

Manuscript received 13 September 2012; accepted 10 January 2013; posted online 16 March 2013; mailed on 16 May 2013.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

1 Introduction

Immersive Virtuality Telepresence (IVT) is undergoing a major resurgence now that technology is catching up with aspirations. A decade on from the Teleimmersion Initiative and Office of the Future, it is reaching a maturity that allows meaningful perceptual studies to ascertain its importance and utility. A long-held challenge of computer science is to faithfully communicate both the appearance and the attentional focus of a remote participant within a shared 3D computer-supported context. IVT has the potential to meet this challenge, however this is not widely appreciated and there is a lack of empirical evidence for this. Communicating both appearance and attention allows viewers to judge both what someone looks like (for example what facial expression they have) and what they are looking at. This is important in understanding: Who thinks what about whom? What is this person looking at and what do they think of it? Consequently, this is relevant for building trust, empathy, and

rapport; and in managing the start and end of conversations and for turn taking within them. IVT could open the door to general solutions to distributing expressive socially and spatially situated meetings. Applications thus range from the ad hoc encounters to design sign-off meetings.

Using IVT of sufficient quality, a wide range of non-verbal communication could be spatially grounded, either across apparently aligned real places or a shared virtual place. As with **I**mmersive **C**ollaborative **V**irtual **E**nvironments, as a user moves, both their viewpoint into the shared space, and their embodiment seen from within it, move with them. As with VC, the real identity and non-verbal cues, even subconscious, are transmitted. Together this should allow interactions such as meeting someone's eye as walking past them, while noticing the dilation of the iris and the quiver of the lip.

Eyes give particularly important expressive cues in socially and spatially situated settings. Eye gaze has been widely studied in telepresence and has proven difficult to communicate across all mediums. As part of its capacity to capture the appearance of a remote participant, IVT has the potential to capture the appearance of the eyes and thereby to communicate the focus of visual attention. We thus focus this work on the communication of eye gaze through IVT. We call the avatars used in IVT *virtuality humans*. These are created through the process of **V**ideo **B**ased **R**econstruction. One such method, suited to producing a fully 3D *virtuality human* to a quality at which eyes are clearly visible, is Shape from Silhouette. The framing and turn of eyes in the head affect natural world gaze estimation. However, with shape form silhouette, the polygonal qualities of reconstructed head, and textural qualities of eyes and face upon it, depend on the arrangement of cameras used to capture colour images and silhouettes. Nothing is known about the impact on estimation of gaze of the relationship between method of reconstruction, turn of eyes, head and body, and relative position of cameras capturing texture and form.

Contribution The theoretical contribution is: to introduce the terms *virtuality human* and *immersive virtuality telepresence*; how the approaches these encompass are uniquely suited to the grand challenge of faithfully communicating both appearance and attention within a tele-shared 3D context; and what sub-challenges arise. The main contribution is, however, practical. This is to provide the first empirical evidence of the reliability of gaze reproduction in virtuality humans. Reliability of gaze estimation is measured across various gaze poses and camera arrangements, likely in every day social interactions and settings. This is the first work to provide evidence that social eye gaze can be reliably communicated through a virtuality medium. Furthermore, it provides the first insight into the importance of the relationship between turn of eyes, head and body, method of recreating and texturing 3D form, and arrangement of capturing cameras.

Definition of terms We introduce the term *Virtuality Human* to describe 2½D or 3D live augmented virtuality avatars, derived from multiple video streams, toward faithful ***replication*** of appearance and attention within a 3D context. Virtuality humans may be incorporated into an otherwise AR, AV or VR scene. We introduce the term **I**mmersive **V**irtuality **T**elepresence to describe the incorporation of *virtuality humans* into telepresence toward faithful ***communication*** of appearance and attention within a ***shared*** 3D context. IVT combines the visual qualities of video with the spatial qualities of Immersive Collaborative Virtual Environments.

2 Background

The majority of research toward the communication of attention and appearance has concentrated on the communication of gaze across mediums that have distinct spatial and visual qualities. We thus introduce the principles and relate the background work within this context.

2.1 Eye Gaze

Gaze is a fundamental and probably the most studied resource in human social interaction, used for eliciting [8] and directing attention [3], and managing flow of conversation, especially multi-party [23]. Estimation of gaze relies on a comparison of eyes and head/face [25, 14] and when one is hidden, accuracy is reduced [21]. Isolating face from eyes caused turn of gaze to be underestimated [13]. The kappa angle is the true orientation of gaze with respect to the head and is different from the binocular axis. This may be why binocular gaze was estimated more accurately than monocular in [14]. The sclera (white of the eye) and iris are particular prominent in humans which might be linked to the importance of gaze as a social resource [2]. The Mona Lisa effect is that of the feeling of eyes of a picture following you around a room. The effect is confined to the horizontal plane [19]. This is because what is displayed is identical from each viewing perspective. As a person walks past a painting, photograph, TV or video wall, anyone depicted on it will appear to remain looking at them as long as the former is looking at the camera, and regardless of observer movement. The Mona Lisa effect is addressed in 2D and 3D images in [15] but as yet such approaches have not been taken up. While estimation of gaze seems to become more accurate with distance, gaze is not typically used as a social resource beyond four metres [11].

2.2 Communicating Gaze through Telepresence

We now summarise how eye gaze has been supported and measured through video, VR and AV mediums.

Communication of eye gaze through a computer medium has long been a goal of computer science. Telepresence has often been cited as the killer application of VR and its importance is marked by the well-known journal that covers the two. 3D mediums are important to social communication through telepresence because 2D mediums cannot communicate spatially situated nonverbal cues, such as gaze, between freely moving people; and in the natural world these cues are not only important to seated interactants. Gaze is interesting as it is perhaps both the most useful and hardest to support of the spatially situated nonverbal cues. While traditional VR has concentrated on believable virtual humans, emotive communication is helped through faithful reproduction of cues. Immersive Virtuality Telepresence is thus interesting as it has the potential to combine the best of Video Conferencing and Immersive Collaborative Virtual Environments, to faithfully communicate both attention and appearance.

In video conferencing the Mona Lisa effect makes all observers feel looked at when the remote participant looks at the camera [1]. The impression of mutual gaze can be achieved through video conferencing by aligning the camera(s) through which a participant views a remote environment, with the video impression of his head [17]. However, the head must be kept in line with the camera, which restricts movement [22]. Aligning camera and face without reducing quality of captured or displayed image has proved hard to achieve [24]. Thus cameras are often placed around or simply above the screen(s). People used to such systems learn to look at the camera closest to the picture of the person to whom they want to give the impression of looking in the eye. In order to communicate gaze between people moving around in respective places, both their viewpoint into the apparently shared space, and their embodiment seen from within it, must move with them. This is not possible in a 2D medium as parallax is not supported. Because the viewpoint is virtualised and attached to that of a moving person, parallax is maintained, allowing a face to be viewed from changing perspectives while walking past it. Thus mutual gaze, for example, can be supported between people moving past each other. ICVEs have these properties and we previously extended one by driving avatar eyes from eye trackers in stereo glasses across three linked CAVEs [22]. However, in ICVEs the appearance of the participants is captured off line and thus does not remain faithful. Video provides a method for faithful reproduction of appearance. There are various methods for

virtualising viewpoint from multiple video streams. These scale from choosing nearest camera [24] to moving a virtual camera within an Augmented Virtuality [9]. IVT combines the latter with remotely observed live reconstructed embodiment, and does so in both directions. It therefore attempts to combine the reproduction qualities of embodiment of VC with the special qualities of embodiment of ICVE. In doing so it thus attempts to faithfully communicate both attention and appearance [22]. Using IVT, the shared space might arise from making two remote spaces seem adjacent or merged, or may be completely virtual. Making spaces adjacent as if divided by a window can be achieved though 2½D virtuality, such as [20]. 3D virtuality humans are more suitable when people can move around the shared space seemingly together, as in [4] AR interfaces can be used to merge real spaces, while VR surround IPT or HMDs are suited to merging spaces. Another way to occupy each other's space is through physical embodiment; be it projecting onto person sized 3D displays [12], furniture or robots [15]. While virtuality humans are not strictly necessary for this, they do offer a way of massively reducing the required number of video feeds due to the virtualisation of the camera. IVT evidently communicates some level of faithfulness in appearance and claims relating to its potential to faithfully communicate eye gaze are growing almost as fast as the spread of prototypes. However, there is little empirical evidence to back these up.

2.3 Measurement of Estimation of Gaze Communicated through Telepresence

The ability of participants to estimate gaze has been measured in VC [22], ICVE [22] and unidirectional IVT [12]. The latter study focused on the impact of display stereoscopy and parallax while projecting the virtual human on/within a cylinder. With parallax and stereoscopy enabled the reported accuracy of estimation of gaze was very close to that reported for the natural world [10]. However, both the turn of the eye and camera placement were unrealistically favorable. Given eyes always faced forward in head, it could be agued that only head gaze was measured. While the impact of the interface of IVT on head gaze estimation has thus been studied, we have found no study of the impact eye gaze, or the relationship between turn of eyes, head, body and relative position of cameras, and the VBR process.

Measuring estimation of gaze is harder when the tracked observer can move about and harder still in immersive settings. Live link ups, such as [22] are complicated not least as it is hard to maintain the same relative orientations of eyes, head and body when observed by many test subjects. All three eye trackers tested gave different accuracies depending on the colour of the eyes. Changing light levels form both the surround moving images, and movement with respect to them, impacts on accuracy [22]. In terms of the captured subject: spatial temporal accuracy of reproduction of moving gaze is impacted by latency, for example network, simulation, rendering and, at each end, motion tracking. This all suggests that to measure the impact of capture and reconstruction, it is sensible to do so with static reconstructions and not tracked immersive interfaces. However, it should be remembered that the latter has been shown to improve accuracy of estimation [12].

2.4 Virtuality Humans Used in Telepresence

We have introduced the term *virtuality human* to describe 2½D or 3D live augmented virtuality avatars, derived from multiple video streams, toward faithful replication of appearance and attention within a 3D context. We now describe the more common techniques for creating *virtuality humans* used within IVT and what aspects of the process are likely to impact on the estimation of gaze.

Image-based reconstruction is a process for recreating the form and appearance of an object from multiple images. VBR extends this temporarily to also reproduce movement, using synchronised streams of video. Colour images, used as textures, are usually paired with additional information from which form can be derived. There are two VBR approaches predominant in telepresence research: shape from silhouette and depth based. The former is well suited to 3D *virtuality humans* as the errors reduce with the number of cameras. An example of the use of a shape from silhouette in IVT is [4]. Depth based approaches, which include stereopsis and depth based cameras that use structured light, have traditionally been used more for 2½D humans as it is more tricky scale them in terms of numbers of cameras. A contemporary example of the use of a depth-based approach in IVT is [12].

A major advantage with depth-based approaches is that depth can be used to segment the background. Shape from silhouette approaches typically use chromo keying for this, which is impractical in most telepresent settings. However, other more practical methods of background segmentation are being researched.

We believe shape-from-silhouette to still have the edge in terms of visual quality. However, if images in recent papers, such as [12], are representative of live performance, the quality gap between the approaches is becoming barely distinguishable. Shape-from-silhouette approaches scale better as sources for errors are reduced rather than increased with number of cameras. This makes them well suited to a study that accurately measures relative impact of many of cameras. Depth based approaches suffer noise that is non-trivial to address in real time, for all but a minimal number of cameras. Shape-from-silhouette approaches do not capture concavities, such as those of the eyes, without extensions. For our experiment we have chosen shape-from-silhouette as we believe it to still define state of the art in terms of visual quality. Of such approaches, we consider the results of the EPVH algorithm [7] to remain representative of state of the art for real-time reconstruction. EPVH creates a polygon form that is then textured. We have parallelised our own implementation of the EPVH algorithm to allow close to interactive frame rates for combined reconstruction and texturing of form on single commodity computers [5]. Problems arise in VBR as the virtual viewpoint moves from that of the texturing camera. In shape-from-silhouette these include apparent stretching of the texture and shadow artefacts such as a hand being textured on a leg that the hand occludes from the camera. While texturing approaches, such as [6] have made good progress in addressing such problems, we have yet to see a report that includes both a clear view of the eyes or temporal performance. Colour texture images require far more bandwidth to communicate than the black and white silhouette images used for the form. Thus while using surround cameras to create the form, some researchers have then only textured it from the camera closest to the viewpoint [9]. The impact of reducing the number of cameras on the accuracy of gaze estimation remains an open question. Shape from silhouette produces the maximal volume within silhouettes and thus does not capture concavities. Eyes are within concavities of the face and thus shape from silhouette can only approximate the placement of eye textures on the face.

3 METHOD

The aim of our experiment was to: Determine if eye-gaze can be estimated from a virtuality human: to within the accuracies that underpin social interaction; and reliably across gaze poses and camera arrangements likely in every day settings. Twenty-two Experimental participants were each presented with a series of life size static virtuality reconstructions of a previously captured human subject and asked to rotate the viewpoint around each until feeling most looked at. The dependent variable was the accuracy with which the participant matched their view of the subject to the line of the subject's gaze. The independent variables were: (1) The relative orientations of head, body and eyes of the captured subject; and (2) the subset of cameras used for texturing. A within subjects design was used such that all participants attempted to align a total of 35 combinations of 5 camera arrangements and 7 gaze poses. (3 asymmetric camera arrangements and thus 12 combinations were also tested but are not reported due to space). The order of presentation was randomized. To simplify the experiment, and in

particular the analysis, and to maximise repeatability, we used static reconstructions. To demonstrate results were valid for not only image based reconstruction but also VBR, the *virtual human* was kept still by reusing a single frame from each camera at an interactive frame rate. Eight of the cameras were placed above the screens of a large octagonal immersive display system in which a captured subject stood. Another two were hung from the roof at a similar height. Five different arrangements of texture cameras were used. Each was a subset of the ten cameras from which silhouettes from images were used to create form. Seven gaze poses compared different relative orientations of body, head and eyes.

3.1 Hypotheses

In the natural world, the turn of eyes in the head impacts more on estimation of gaze than the turn of head on the body. Thus:

H1 The relative orientation of eyes but not body to head will significantly impact on the accuracy of estimation of gaze from a virtuality human.

Social gaze in the natural world is used up to about 4m. At this distance two people stood shoulder to shoulder would be 4° apart.

H2 Gaze can be estimated through VBR to accuracy underpinning social gaze in the natural world. i.e. 4°@≈4m.

Colour images used for texture require far more bandwidth to communicate than silhouettes taken from them, used to create form. Furthermore, a pilot study showed that reducing cameras used for texture, had far more impact on estimation of gaze. Thus it is particularly useful to know the impact of reducing texture cameras, isolated from that of reducing form cameras.

H3 Reducing the number of texture cameras will significantly impact on gaze estimation.

Telepresent systems may have large display walls that participant can walk up to, and place cameras above these walls to avoid occluding the image. Thus it is useful to know the impact of steepness of texture camera to face.

H4 Increasing the steepness of texture camera to face will significantly impact on gaze estimation.

3.2 Scope

This study investigates the ability of a medium evidently able to communicate some level of faithfulness to also communicate gaze. It measures how aspects of visual faithfulness impacts on faithfulness of gaze and provides visual evidence of this. It does not directly measure faithfulness of appearance, for example through subjective user impression or objective image correspondence.

Many methods for VBR are being researched and vary in terms of visual quality, temporal performance and primary impacting factor on scalability. We are specifically interested in approaches that can reproduce the entire human form to a visual quality from which the whites of the eyes are clearly visible, within interactive frame rates. Literature suggests the EPVH algorithm to posses suitable qualities and we have been able to realize these requirements by implementing a parallelised version of it. EPVHs are typically textured by blending images of cameras close to a polygons norm. However, there are various ways of doing this, when we have tried it, the result has been to blur the eyes, and the real-time performance depends on number of textures merged. In order to provide a clear benchmark with easily repeatable results we have chosen not to merge textures in this initial experiment. We have chosen a lab-based study to gain clear results that are not affected by a changing environment. Furthermore our experiment was carried out in an immersive display and capture space designed for telepresence. Both capture and display was carried out in our octagonal display and capture space, known as the octave. This consists of eight viewing screens of 4m wide and 2m high arranged into a regular octagon. However, as we wanted to decouple impact of immersive display we have only used one screen for display and did not enable stereoscopy or parallax.

3.3 Independent Variables

The independent variables were: (1) The relative orientations of head, body and eyes of the captured subject; and (2) the subset of cameras used for texturing.

3.3.1 Relative Orientations of Head, Body and Eyes

To test hypothesis H1 we varied relative orientations of head, body and eyes. Seven gaze poses represented all possible combinations of eyes, head and body; being either centred or turned, Figure 2. It is important to note that the orientation of body, head and eyes with respect to the cameras differs across the poses. This is reflected diagrammatically in the second row of figure 1 and in some later figures. In each of the seven gaze poses, the captured subject's eyes either looked straight ahead or were at their maximum comfortable orientation (subject looked at targets either straight ahead or at 18° from the mark on which they stood).

We describe each gaze pose both through a triple of body, head and eye orientation and diagrammatically showing the normal of each with respect to the full set of cameras. The ordered triple describes (rotation of body relative to the eyes, rotation of head relative to the eyes, rotation of eyes relative to the centre line). For example (R,0,L) indicates that the body is rotated to the right from the eyes, that the head and eyes are in line, and that the eyes are rotated to the left. Note that the first two values thereby give the relative arrangement of the body, head and eyes within the target participant and the third value gives the spatial orientation of the assembly as whole. As explained below, the increments of rotation are nominally 18°, however, because this is not the exact value, and to simplify matters, L and R will be used to show a nominal rotation of 18° to the left and right respectively and L' and R' to show nominal rotations of 36°. Left and right are those of the captured subject rather than an observer. The seven gaze poses are shown both diagrammatically and from photographs of the captured subject, in Figure 2. It should be noted that these photographs are taken from an SLR camera without frame locking to the reconstruction system.

In addition to the relative orientations of eyes, head and body within the captured subject, the different gaze poses also involve different orientations relative to the cameras. Consequently it is relevant to describe how the different gaze poses were achieved. We chose to place three viewing targets at 36° apart when viewed from 4m in front of that in the centre. We had looked for a gaze angle that appeared natural, was comfortable for the observer, and at the whites of the eye were barely discernible on both sides.

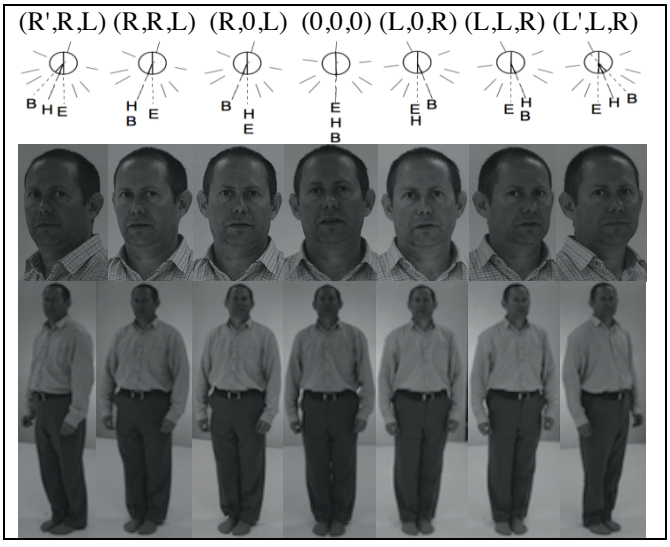

Fig. 2. The seven gaze poses, each shown diagrammatically above photographs of head and whole body from the gaze target. Normal of body head and eyes (B H E) are shown in relation to the cameras.

After experimenting with various pairs of observer and observed from within the research team, we found 20˚ to be a good compromise. As it happens, at 4m perpendicular to any one of our large display screens the edges were at 18˚ off centre. This gave a clear marker for target placement, which was close enough to still meet our requirements. A laser measuring tool and trigonometry were used together to place a marker on the floor at 4m from the perpendicular line from the screen centre bottom and equidistant to the left and right markers. The target participant stood on this marker. The distance from the captured subject to the front gaze target was 4m, with that to the left and right around 5m.

3.3.2 Cameras used for texturing

To test hypothesis H3-H4 we varied the subset of cameras from which images were used for texturing.

Table 1. Cameras arrangements

Name	Diagram	Cameras	Description
Shallow		1	Single front ~ 15˚ Vertical to face
Pair		2,8	~ 38˚& 342˚ H & ~ 15˚ V to face
Arc		1,2,3,7,8	~ 38˚- 342˚ Horizontal
Surround		0-9	Around & above
Steep		9	Single front ~ 30˚ Vertical to face

To evaluate the impact of reducing texture cameras we compared camera arrangements of *Single*, *Pair*, *Arc* and *Surround*; with one, two, five and ten cameras respectively. To evaluate the impact of steepness of texture camera to face we compared two single frontal cameras, both just above 2m from the floor with one about 2m and the other 4m in front of the subject. The five camera arrangements are described in Table 1. (wrt Table 1 - dot in centre of diagram for *Surround* is camera looking down on head).

Each gaze pose was captured in ten synchronised frames, each from a different camera facing the target participant from a unique angle. Ten silhouettes were created, one from each image. The form was created using all ten silhouettes and textured from various subsets of the ten images. The placement of cameras with respect to the captured subject and display walls is shown in Figure 3. Example images and corresponding silhouettes are shown in Figure 4.

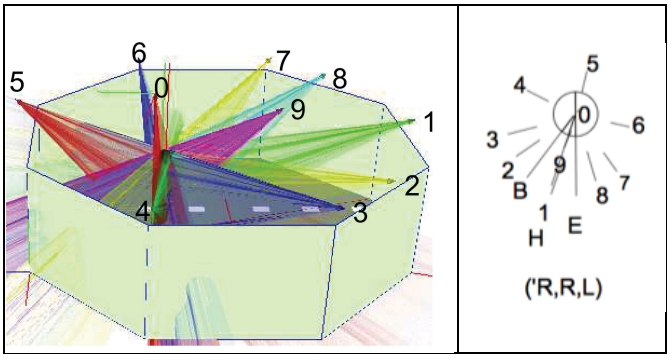

Fig. 3. Camera placement with silhouette cones to captured subject (left) and schematic of cameras with respect to a given gaze pose (right). All but two of the cameras are fixed on top of the display walls, with the other two mounted on above.

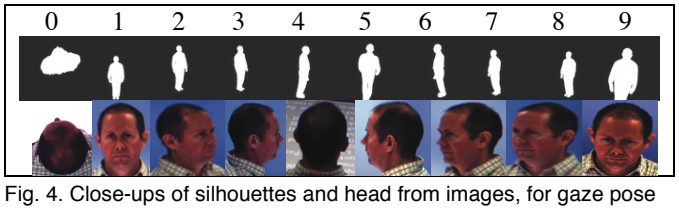

Fig. 4. Close-ups of silhouettes and head from images, for gaze pose (0,0,0) where head, eyes and body face front.

3.4 Dependent Variable

The accuracy with which participants could orient the *virtuality human* such that they felt it was looking directly at them was measured in degrees. Thus if the participant oriented her view directly along the line between the captured subject and his gaze target, the error in accuracy would be 0˚. The captured subject and participant are both in real space but not at the same time. The participant rotated around the *virtuality human* by controlling the virtual viewpoint of their display. The ideal situation is where the captured subject remains centred on the same point across the gaze poses and the participant rotates around the resultant *virtual human* centred at that point. In practice, in the absence of using physical restraint, the exact position of the subject being captured cannot be guaranteed across the different gaze poses. Rather than assuming that is constant, the actual position is c from the position of *virtuality human*, and a correction factor calculated using trigonometry.

3.5 Experimental Participants

Twenty-two participants were recruited from staff and students within the university and colleagues and friends of the researchers. They came from a wide range of working backgrounds: seven administrators, two computer scientists, one sociologist, two computer support, one media industry, one computer science student two psychology students, the rest chose not to say. There was a reasonably even spread of age between twenty and forty and one in mid forties. All were given an information sheet and consent form prior to the experiment. None reported illness or disability that might result in untoward outcomes or affect the results. One reported autism but his performance was typical. All had normal, or corrected-to-normal vision. Participants who were not university staff were paid £10. University Ethical Approval was received. The results from 4 of the 22 were discarded, as they did not complete the experiment.

3.6 Procedure

The test subjects all experienced and moved the viewpoint around the *virtuality human* in an identical test environment. This rotation was controlled through arrow keys on a hand held wireless keyboard, to provide a familiar, simple and accurate interface. The rotation speed was set to 0.25º per rendering frame at around 40Hz but could be changed by the participant. The reconstructed target participant was displayed life size and at the correct perspective. The distance at which the participants viewed the captured subject was set to 3.5 metres. This was to keep it just within the bounds of natural social gaze (4m). Each participant was instructed to stand with the balls of their feet touching at the centre of a cross, aligned to the required eye position. Motion tracking was not used as it introduces much potential for errors and can be a distraction to test subjects in it's fitting, wearing and errors. Test participants' movement was mostly horizontal and mostly sway of body rather than moving of feet. The differential rotation affect, more colloquially known as the Mona Lisa affect, means that horizontal orientation [19] of an observer to a screen has little impact on their impression of where a person depicted on it is looking. Observer eye height was measured and entered into the system at the start of each trial both to match the eye height of the *virtuality human* and to calculate the view frustum.

3.7 Technical Details

The approach to creating the form is based on the well-known EPVH algorithm [6]. We have optimised this through parallelisation to achieve real-time frame rates while reconstructing from 10 HD cameras on a single computer [5]. We used our software environment [5] throughout. The volumetric reconstruction was formed, textured and viewed within it. For this experiment, we adapted it to allow the reconstructed form to be viewed from any angle while changing which two subsets of cameras were used to create and texture the form. Here we only report variations in texture. The participants used the environment to view and rotate around a

series of reconstructed models, pressing the space bar when they felt each most closely looked at them. The authors used it both to plan camera placement and during analysis to understand the cause of visual effects and subsequent differences in gaze estimation. The interface allows the various stages of the reconstruction process to be visualised. For example in Figure 3, textured model, viewing lines, rotation of axis and the surround display environment are selected and shown.

The subject was captured within an octagonal CAVE-like display system designed for telepresence research. This allowed us to easily test camera placements representative of a highly immersive telepresence system. However, we wanted to decouple impact of immersive display as this has been tested elsewhere [12] and not decoupling would significantly complicate analysis. We thus only used one screen for display and did not enable stereoscopy or parallax. The used display combined a 4x2m simulation grade active surface with a Christie S+3K Stereo DLP, running mono at a resolution of 1400x1050 at 102Hz. The user trials were run on either an Apple Mac Mini or Macbook Pro.

During capture, defuse/ambient light from the immersive displays fourteen projectors (6 ceiling and 8 wall) was combined with defuse spotlights to achieve clear lighting and contrast of the face. Defuse spot lighting was bounced off the floor in front of the captured subject in order to remove much of the shadow around the eyes, without causing glare in their face and reflected highlights in the pupils. Spotlights were also placed behind the screens to increase ambient light. The room's strip lighting was turned on. The brightness and colour of image capture was manually adjusted. This perhaps surprising decision was taken as pilots showed noticeable colour differences not to be an important factor in gaze estimation and actually very helpful in analysing images of the reconstruction in terms of texturing cameras. During display, only one projector and some safety lights in the entrance were turned on.

3.8 Method of Analysis

We use three metrics for significance: statistical; practical; and qualitative. In the quantitative analysis we investigate statistical and practical significance, whereas significant corroborating evidence is investigated in the qualitative analysis.

Accuracy of gaze estimation in terms median and interquartile range of absolute errors is shown in box plots. Where statistical significance lies, it is described in the text. In the box plots: horizontal line marks 4º accuracy; upper and lower box, respective quartile; line in box, mid quartile; whiskers, min and max; numbered dots and circles, outliers. Statistical significance is measured using the Wilcoxon Signed Rank Test. We take: ***p* values of less than .05 to indicate significance**. Where data is aggregated for an overall view, for example combining all gaze poses where eyes are centred and all that are turned, or combining all camera arrangements, we report median of medians. Practical significance compares accuracy of gaze estimation to that typical of social gaze in the real world. Our test subjects were stood 3.5m from the virtual humans. 4m is the extent of social gaze distance and we have chosen 3.5m as it falls comfortably but not excessively within this. From this distance, centre head to end of shoulder are about 4˚ apart. Thus we argue 4˚ is the critical accuracy necessary for two people stood shoulder to shoulder to determine which is being looked at from this distance. Both lighting and eyesight impact upon the accuracy of estimation. Experience in undertaking [1] suggested that allowing for a 1˚ variance would be prudent. For this experiment we consider **practical significance to be the movement by more than 1˚ of one of the quartiles or median error in estimation that takes it across the 4˚ threshold.** For this experiment we defined qualitative significance as any visible reduction in quality of representation of eyes or face that correlates to a reduction in both statistical and practical significance. Images from both source cameras and virtual cameras were compared. Virtual viewpoints were also interactively moved in some cases for the authors to experience the same rotation as the test subjects.

4 Results and Analysis

Our quantitative analysis relates eye orientation and camera arrangement to the performance of estimation of gaze. This is followed by a qualitative analysis, which examines images of the reconstructions and from the cameras in search of corroborating evidence and underlying reason.

4.1 Quantitative Analysis

We now examine the impact on accuracy of gaze estimation of (1) the relative orientations of eyes, head and body (2) the set of cameras used for texturing (3) the size of horizontal array and (4) steepness of vertical angle.

4.1.1 Impact of relative orientations of eyes, head and body

Firstly we consider the overall differences in estimating gaze direction from different gaze poses, averaging across the five camera arrangements, Figure 5. Unsurprisingly, the best accuracy is achieved for (000), where the eyes are centred in the head and the head is centred on the body. Whether or not the head and eyes are turned relative to each other results in a significant difference with eyes centred (L0R and R0L), (median 3.82°, IQR 1.68°) performing significantly better than eyes turned (L'LR, LLR, R'RL, and RRL) (median 5.9288°, IQR 2.67°) Z=-3.2 p= .001. However, the relative orientation of head and body did not make a statistically significant difference. Head and body aligned (LLR and RRL) (Median 5.45°, IQR 2.67°) showing no significant difference with head turned (L'LR and R'RL) (Median 5.87° IQR 4.89°) Z=-.631 p= .528. An asymmetry between left- and right-oriented gaze poses can be noted with R'RL performing better than RRL but L'LR performing worse than LLR. Consequently, it is clearly necessary to distinguish between gaze poses in which eyes are centred in head and those in which eyes are turned; yet it is not necessary to distinguish between poses according to the relative rotation of head and body. Figure 5 further shows the median performance for each gaze pose only achieved the 4° criterion when the eyes were centred (L0R, 000 and R0L).

4.1.2 Impact of Camera Arrangement

We now report the impact of practical camera arrangement on accuracy of estimation of gaze. Firstly we consider each camera arrangement's overall performance averaging across all gaze poses, Figure 6. A practically significant different occurs with *Shallow*, *Pair*, *Arc* and *Surround* all performing around the 4° criterion but the *Steep* viewing angle performing very much worse. Over all gaze poses a median accuracy of 4° is achieved by *Surround* and *Arc* with both *Shallow Single* and *Pair* being within 1° of this limit. The following subsections compare the performance of the *Shallow*, *Pair*, *Arc* and *Surround* across different gaze poses and then the performance of *Shallow* and *Steep*.

4.1.1 Impact of Number of Cameras across Gaze Pose

The overall performance at around the 4° mark is markedly different across all frontal array camera arrangements, when eye centred and eyes turned gaze poses are separated (Figure 7). When eye centred and eyes turned gaze poses are combined, texturing from *Surround* (Median 3.230° IQR 2.261°) significantly outperformed both *Pair* (Median 4.755° IQR 1.133°, Z=2.591 p=.031), and *Shallow* (Median 4.191° IQR 3.135°, Z=2.156, p=.010). This appears to be due to significant differences for the eyes centred gaze poses. *Surround* eyes centred (Median 1.735° IQR 1.404°) being better than *Pair* eyes centred (Median 3.136° IQR 2.384°, Z=2.243, p=.025); and better than *Shallow* eyes centred Median 3.675° IQR 3.581°, Z= 2.461, p=.025. *Arc* significantly outperformed *Shallow* only when eyes centred and not overall (*Arc* centred (Median 2.340° IQR 1.718°), *Shallow* centred (Median 3.675° IQR 3.581, Z=-2.112, p=.035). Overall, Figure 7 suggest that an error of around 4° can be achieved for all the frontal camera arrangements.

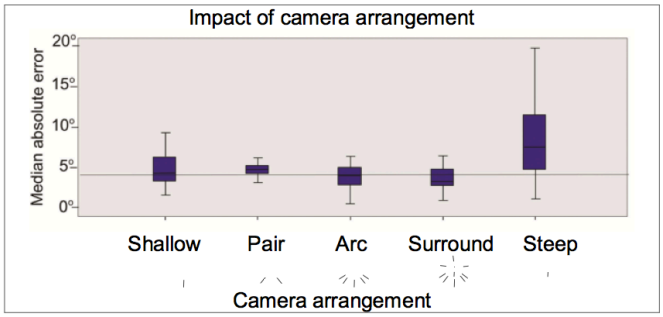

Fig. 5. Taken across all camera arrangements, median of median estimations always and only below 4° when eyes centred (x0x).

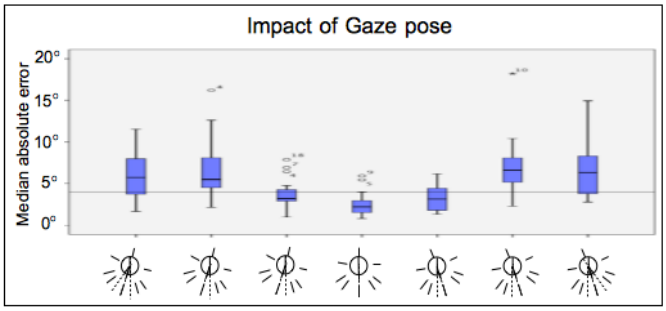

Fig 6. Over all gaze poses a median accuracy of 4° is achieved by *Surround* and *Arc* with both *Shallow Single* and *Pair* within 1 of this limit. *Steep Single* stands out as clearly worst performer.

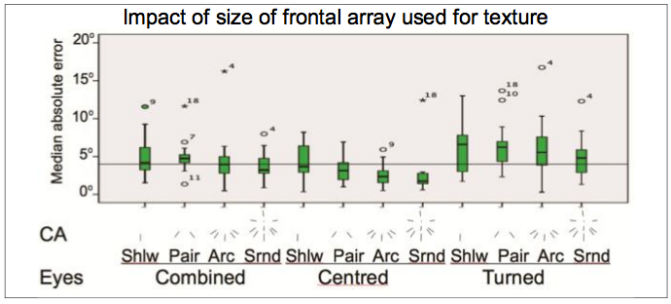

Fig. 7. Median accuracy increases with number of texturing cameras both when eyes centred and turned, being within 4° when centred.

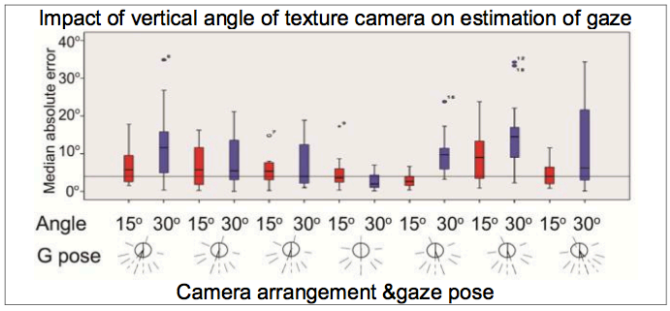

Fig. 8. Estimations when texture camera close to eye level outperform those when a texture camera looking down at a steeper angle. At worse when eyes turned and camera angle steeper.

4.1.2 Impact of Vertical Viewing Angle across Gaze Poses

Comparing *Shallow* and *Steep* viewing angles for a single texturing camera shows that *Shallow* outperforms *Steep* for gaze poses where the eyes are turned, but that they perform comparably when the eyes are centred, Figure 8. Using *Shallow* rather than *Steep* resulted in statistically significantly better performance overall, *Shallow* (Median=4.19° IQR= 3.13) outperforming than *Steep* (Median=7.45° IQR=7.34°, Z=2.156, p=.031) and approaching the 4° criterion. This difference appears to be due the difference between gaze poses with eyes turned, *Shallow* eyes turned, (Median=6.60° IQR=4.89°) was less than *Steep* eyes turned (Median=8.96° IQR=9.01°, Z=2.504 p=.012).

4.2 Qualitative Analysis

We now look for visual differences in the various models that may explain differences in performance of estimation of gaze. We also explore underlying reasons that relate these visual effects to camera arrangements. Figure 9 shows close ups of headshots of all reconstructions taken from the target of gaze. The difference in brightness of the textures helps to show that the eye closest to the viewpoint is taken from a different camera in the two conditions. That used in *Surround* is close to the viewpoint these pictures were taken from. Thus we are seeing a stretch in texture when the texture camera is far from the virtual viewpoint. It is notable that the inability of shape-from-silhouette to capture concavities has exasperated stretch by sloping the eyes toward the nose and thus increasing the difference between the normal of the polygon and the centre camera.

We now look at why the impact on subset of cameras was not significant when eyes are turned. Reconstructions of gazes with eyes turned right suffered stretching of the dark of the rightmost eye, considerably greater than that seen when eyes were centred.

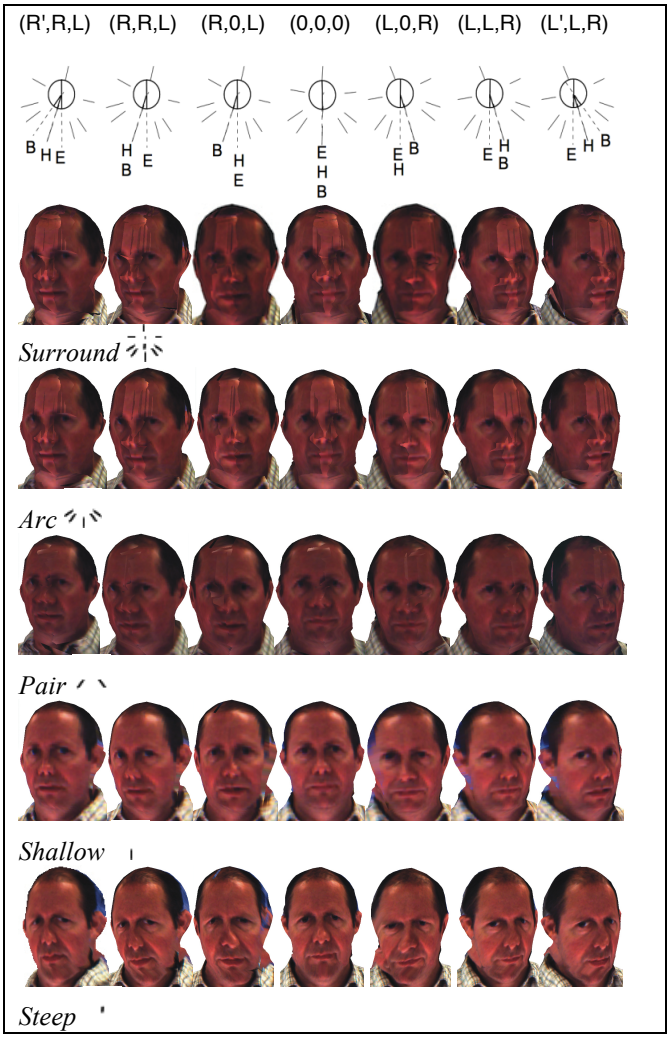

Fig. 9. Close up of head from target of gaze. Ranked from top to bottom in order of accuracy of gaze estimation across all poses.

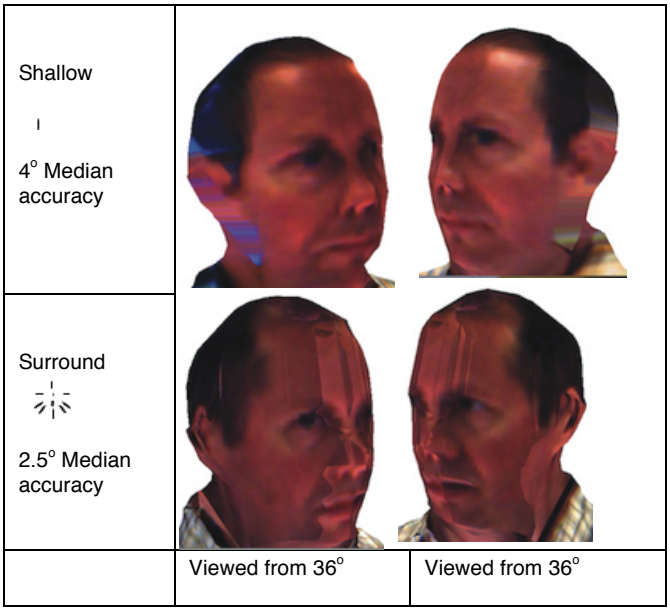

Fig. 10. Rotating past the gaze target exaggerates the differences between *Surround* and *Single* texture. The dark of the eye appears stretched when taken from a single texture but not when taken from a texture from the camera close to this viewpoint. (0,0,0).

Closer inspection of images from source cameras and of reconstruction showed the cause to be a combination of shadow in the corner of the eye and insufficient camera resolution to clearly distinguish this from the dark and white of the eye, Figure 11. During analysis we reproduced a similar result through blurring in an image editor, Figure 11. Interestingly, the problem was most apparent for the single central camera.

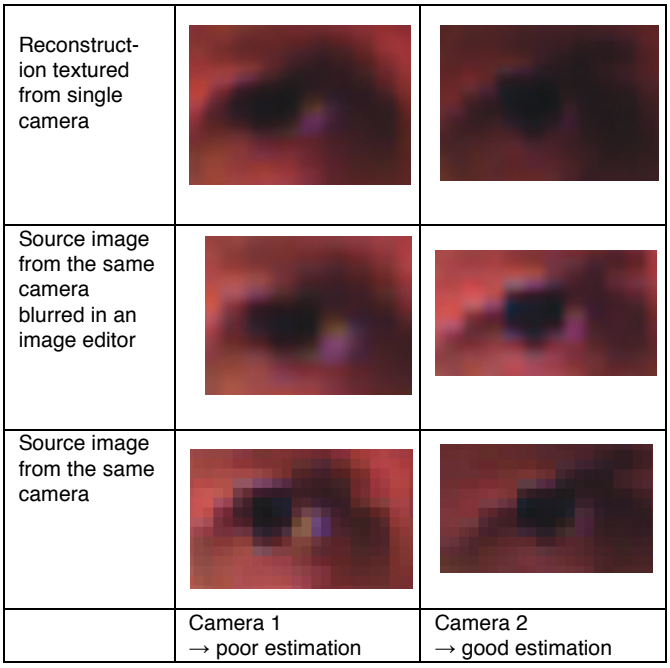

Fig. 11. Close up of right eye, Reconstructed, blurred through an image editor, and source. The image used for texture from one camera but not the other appears to stretch the dark of the eye. The underlying cause seems to be insufficient camera resolution to resolve dark of eyes from shadow when image blurred by texturing. The effect of texturing is similar to that of blurring in an image editor.

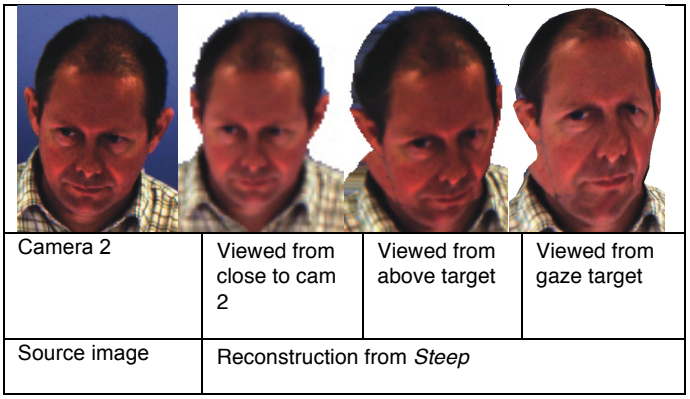

Fig.12. Reconstruction textured from camera looking down steeply on face looks reasonable from a similar vertical angle but distorts as viewpoint lowered to eye level. (L,L,R).

4.2.1 Steepness of Texture Camera to Face

Texturing from a camera angled steeply to the subject's face caused a blatant drooping effect to the eyes and an apparent twist to the end of the nose as if pulling it toward the viewpoint. The bridge of the nose, the inner corners of the eyes and the top of the mouth are not drooped, whereas the droop of the eye increases gradually toward the outer corners, and drooping appears at the end of the nose and under the chin.

In all cases the cause appears to be texturing beneath an overhanging polygon that should have occluded the camera. The reason eye droop gradually increases from nothing at the bridge of the nose, is from shape-from-silhouette's inability to capture the concavity of the eye socket. When viewed from the steepness of the texture camera, the reconstruction looks reasonably correct, Figure 12.

5 Discussion

Over all gaze poses, our first hypothesis **H1 proved to be clearly true**. In terms of practical significance the median of median estimation errors when eyes centred were 3° better, and only and always within the 4° limit of social gaze perception, when eyes centred. p=.001 was statistical significance. Stretching of the dark of the eyes was only clear when eyes turned. This confirmed that it was worth taking relative eye to head orientation into account in the rest of the analysis but not accounting for that of relative body orientation.

H2 was proved to be clearly true. The *Surround* camera arrangement was found that allows participants to estimate gaze to within 4° for upper quartile of samples. This tells us that VBR is capable of supporting the level of accuracy of gaze estimation used social gaze. Furthermore, when eyes were centred, 4° fell between upper and mid quartiles, and when turned, between mid and lower. Together these findings reinforced the use of *Surround* as a useful benchmark.

H3 proved to be partially (barely) true. In terms of practical significance, while the median was moved across the 4° line, this was by less than 1° and not always in the right direction (median of *Shallow* better than that of *Pair*). 1° is almost certainly within the noise of the experiment and at approaching 4m, depending on eye sight and lighting, unlikely to be clearly discernible in the natural world. The statistical difference between *Surround* and *Pair* (p=.031) were significant, however, others between different camera arrangements were not. With close inspection, stretching of the dark of eyes was generally worse when fewer cameras were used.

H4 was proved to be clearly true. In terms of practical significance, over all gaze poses, the lower quartile was moved across the 4° line by a steeper camera angle and the median was moved by around 3°. Statistically estimations when texturing from a shallow vertical angle significantly (p=.031) outperformed those using of a steeper. Visual

distortion of the eyes, and when turned, stretching of the dark of the eyes, was obvious.

The overriding cause of stretching of the dark of the eye appears to be different for eyes centred and turned. In both cases it only occurred in certain combinations of gaze pose and camera configuration. The primary cause when eyes centred appears to be interplay between the poor modelling of the eye socket, the curve of the head, and the difference between viewpoint of real and virtual camera. The poor modelling of the eye socket is because shape-from-silhouette does not capture concavities. This problem appears to be the overriding impacting factor when eyes centred. When eyes are turned, some cameras had problems in discerning the dark of the eye from shadow on the skin. We notice the same problem in images from [22] which used commodity cameras rather than shape-from-silhouette. In the real world, gaze is harder to gauge from turned eyes. In the absence of other convincing evidence we are left presuming that this had a larger impact than the quality of the reconstruction, and that this is why the number of texturing cameras only had a significant impact when eyes centred. H4 Proved to be true in that increasing the steepness of a texture camera to face: significantly (*Shallow* Median=4.19° IQR= 3.13º, *Steep* Median=7.45° IQR=7.34º, Z=2.156, p=.031) reduced the accuracy of gaze estimation; and visibly reduced the quality of reproduction (through drooping of face beneath overhanging polygons – above eyes, and bottom of nose and chin). To make sure that the droop did not come from poor calibration, we later tried three cameras at different steepness to the one in the centre, which was calibrated to lowest RMS. Droop was still increased as steepness of angle.

There are two ways in which our reconstructed humans could be said to fall short of state of the art for real-time reconstruction. Firstly the manual pre run adjustment of colour and brightness falls short in terms of quality of automated techniques commonly used. However, we decided to keep this method as results from pilots and the main study showed that participants performed well with conditions where polygons on the face were textured at different brightness, and the differences proved very helpful in our understanding and we hope presentation of the results. We find it unlikely that improving uniformity of colour would result in worse estimation of gaze. However, the differences in light in the corners of the eyes that came from different camera brightness levels did. This suggests that when choosing a brightness balance, the potential for extenuating shadow in the corner of the eyes should be considered. The second way in which we chose to fall short of state of the art was in the texturing approach. Secondly, it is likely that our texturing each polygon from a single camera and not calculating a view dependent projection may have contributed to the stretching the dark of the eye in some cases. However, close examination of the model while rotating it leaves us believing that the slanting of the polygons across where the eye socket should have been to the nose also impacted.

There is a discrepancy between the distance from participant to *virtual human* and that used to calculate the angular difference from head to shoulder. The former was 3.55 metres and the latter 4. Our reasoning at the time was that we wanted to measure accuracy within the limit of social distance but compare it to worst case of the limit. In hindsight it may have made more sense to keep both distances the same.

We did not use parallax or stereoscopy in order not to impact the results by fundamental and technical characteristics of immersive interface. This was in part as a study had already looked at these but not at medium or camera arrangement [12]. In that study parallax and stereoscopy were shown to improve gaze estimation. Thus it is reasonable to assume that the results we obtained would be further improved by running the display in immersive mode, although more noise may have been introduced.

6 Conclusion

The first contribution of this paper was to explain how various aspects of Mixed Reality can be combined to overcome limitations of today's approaches to telepresence. Specifically we argued that while Video Conferencing faithfully communicates appearance but not attention, and Immersive Collaborative Virtual Reality does the opposite, Immersive Virtuality Telepresence has properties suited to doing both. This is important as it allows communication of what a person feels about what, through even subconscious cues. This significantly opens the door to supporting trust, empathy, rapport, and ad hoc or situated meetings. The point of departure for the experiment was that: both the framing of eyes in the head and the relative turn of each, impacted on gaze estimation in the real world; the quality to which this was reproduced in IVT was likely to be impacted by the relationship between process of Vision Based Reconstruction and camera arrangement; and the impact of parallax of display of IVT on head gaze estimation had been studied but only under unrealistically favourable conditions of camera placement. We call this head gaze because eyes were always fixed forward. This was the first work to accurately measure the impact on gaze estimation of a VBR medium used in IVT. In particular it is the first to measure the impact of the relationship between method of VBR, camera arrangement and turn of eye, head and body. Thus it tells us much more of the reliability of the medium in practical situations. It tells us if CGI avatars created in real time from multiple videos can today support eye gaze. It further tells us under what conditions and how the way in which the form is created and textures applied to it impact on the requirements for camera placement.

As fewer texturing cameras lead to less bandwidth consumption, we wanted to know how the number of cameras covering the face impacted on estimation of gaze. As cameras are often mounted directly above the display walls in telepresent systems, thus allowing for the possibility of steep viewing angles as the walls are approached, we also measured impact of the steepness of the texturing camera to the face. Within the paradigm adopted, it was not possible to make a direct comparison with face-to-face interaction but we adopted accuracy to within 4° as a criterion. We showed that this level of accuracy could be achieved by implementing a shape-from-silhouette approach to VBR. However, there was a notable difference across gaze poses: when the target's eyes were turned relative to the head, this criterion was achieved by less as than half of the participants, even with the best performing camera arrangement. We conclude that unless cameras are placed correctly with respect to likely gaze poses, erroneous attributes of form and texture and in particular the relationship between the two, significantly impacts on accuracy of gaze estimation. Specifically estimation can achieve sufficient accuracy to allow a participant to determine if they are being looked at, provided that the viewing angle is not too steep, however, when the target person's eyes are turned to one side (relative to their head) a comprehensive array of surrounding cameras is preferable.

Acknowledgments

The authors wish to thank the EPSRC and OMG VICON for funding of PhD students, HEFCE for funding the equipment under SRIF, and John O'Hare of Salford for technical support.

References

[1] S.M. Anstis, J.W. Mayhew and T. Morley, 1969, The Perception of Where a Face or Television 'Portrait' Is Looking. The American Journal of Psychology, 82:474-489. 1969.

[2] M. Argyle and M. Cook, 1976, Gaze and mutual gaze. Cambridge: Cambridge University Press. ISBN 10: 0521208653 / 0-521-20865-3.

[3] M. Argyle, and J. Graham, 1977, The Central Europe Experiment - looking at persons and looking at things. Journal of Environmental Psychology and Nonverbal Behaviour, 1, pp. 6-16.

[4] Allard, J., Franco, J., Menier, C., Boyer, E., Raffin, B.: The GrImage Platform: A Mixed Reality Environment for Interac- tions. IEEE International Conference on Computer Vision Sys- tems, 2006 ICVS'06, pp 46–46 (2006)

[5] T. Duckworth and D.J. Roberts, 2013, Parallel processing for real-time 3D reconstruction from video streams, Springer-Verlag Journal of Real-Time Image Processing, 10.1007/s11554-012-0306-1.

[6] M. Eisemann, B. De Decker, M. Magnor, P. Bekaert, E. de Aguiar and N. Ahmed, C. Theobalt, A. Sellent, 2008, Floating Textures, Computer Graphics Forum 27(2), 2008, pp409-418. DOI: 10.1111/j.1467-8659.2008.01138.x

[7] J. Franco and E. Boyer, 2009, Efficient polyhedral modeling from silhouettes. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 13(3), 414 – 427, 2009. DOI:10.1109/TPAMI.2008.104

[8] Goodwin, C., 2000, Action and Embodiment Within Situated Human Interaction, Journal of Pragmatics, 32, pp. 1489-522.

[9] O. Grau, A. Hilton, J. Kilner, G. Miller, T. Sargeant and J. Starck, 2007, A free-viewpoint video system for visualisation of sport scenes, SMPTE Motion Imagining Journal. May 1, 2007, vol 116, no5-6, pp. 213-219. DOI:10.5594/J11445

[10] 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, 2003, Blue-c: a spatially immersive display and 3D video portal for telepresence, ACM Transactions on Graphics (TOG), Vol. 22, Issue 3, pp. 819-827. DOI:10.1145/882262.882350

[11] E. Hall, (1966). The Hidden Dimension. Anchor Books. ISBN 0-385-08476-5.

[12] 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. of the 2012 ACM annual conference on Human Factors in Computing Systems, New York, NY, USA, 2012, CHI '12, pp. 2531–2540, ACM. DOI:10.1145/2207676.2208640

[13] N.L. Kluttz, B.R. Mayes, R.W. West and D.S. Kerby, 2009, The effect of head turn on the perception of gaze, Vision Research, Volume 49, Issue 15, 22 July 2009, pp. 1979-1993, ISSN 0042-6989, DOI:10.1016/j.visres.2009.05.013.

[14] S.R.H. Langton, 2000, The mutual influence of gaze and head orientation in the analysis of social attention direction. Quarterly Journal of Experimental Psychology, 53:825-845. DOI:10.1080/713755908.

[15] P. Lincoln, G. Welch, A. Nashel, A. State, A. Ilie, and H. Fuchs, 2011. "Animatronic shader lamps avatars," Virtual Real., vol. 15, no2-3, pp. 225–238, June 2011. DOI:10.1007/s10055-010-0175-5

[16] S. Al Moubayed, J. Edlund, and J., Beskow, 2012 "Taming mona lisa: Communicating gaze faithfully in 2d and 3d facial projections," ACM Trans. Interact. Intell. Syst., vol. 1, no. 2, pp. 11:1–11:25, Jan. 2012. DOI:10.1145/2070719.2070724.

[17] N. Negroponte, Being Digital. Alfred A. Knopf, Inc., New York, ISBN-10: 0679762906 | ISBN-13: 978-0679762904, NY, USA, 1995.

[18] D. Nguyen, and J. Canny, 2005, MultiView: Spatially Faithful Group Video Conferencing, in Proc. CHI 2005, ACM Press, pp. 799-808. DOI:10.1145/1054972.1055084

[19] E. Oyarskaya and H. Hecht, 2009, The Mona Lisa Effect: Is it confined to the horizontal plane?. In Perception 38 ECVP, 2009. p31.

[20] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, & H. Fuchs, The Office of the Future: A Unified Approach to Image-Based Modeling and Spatially Immersive Displays, in Proc. 25th conf. Computer graphics and interactive techniques, SIGGRAPH'98, pp. 179-188. DOI:10.1145/280814.280861

[21] Rae, J P & Roberts, D J 2011, Some Implications of Eye Gaze Behavior and Perception for the Design of Immersive Telecommunication Systems, in: 'IEEE/ACM Proceedings of 15th Int. Symp. On Distributed Simulation and Real Time Applications', IEEE, Salford, UK, pp.120-125.

[22] D. 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 IEEE Virtual Reality 2009. pp.135-142

[23] R. Vertegaal, G van der Veer, H. Vons, 2000, Effects of Gaze on Multiparty Mediated Communication, In: S. Fels, P. Poulin (eds) Proceedings of Graphics Interface 2000, (Montreal, Canada, 15-17 May 2000). Morgan Kaufmann Publishers. ISBN 0-9695338-9-6. pp. 95-102.

[24] R. Vertegaal, I. Weevers, C. Sohn and C. Cheung., 2003, GAZE-2: conveying eye contact in group video conferencing using eye-controlled camera direction, in Proc. of CHI'03 SIGCHI Conference on Human factors in computer systems, ACM Press, New York, pp. 521–528. DOI:10.1145/642611.642702

[25] W.H. Wollaston, 1824, On the Apparent Direction of Eyes in a Portrait. Philosophical Transactions of the Royal Society of London, 114:247-256.