Eye Tracking for Avatar Eye Gaze Control During Object-Focused Multiparty Interaction in Immersive Collaborative Virtual Environments

William Steptoe*[1], Oyewole Oyekoya[1], Alessio Murgia[3], Robin Wolff[2],
John Rae[4], Estefania Guimaraes[4], David Roberts[2], Anthony Steed[1]
[1]Department of Computer Science, University College London
[2]Centre for Virtual Environments, University of Salford
[3]Department of Computer Science, University of Reading
[4]Department of Psychology, Roehampton University

ABSTRACT

In face-to-face collaboration, eye gaze is used both as a bidirectional signal to monitor and indicate focus of attention and action, as well as a resource to manage the interaction. In remote interaction supported by Immersive Collaborative Virtual Environments (ICVEs), embodied avatars representing and controlled by each participant share a virtual space. We report on a study designed to evaluate methods of avatar eye gaze control during an object-focused puzzle scenario performed between three networked CAVE™-like systems. We compare *tracked gaze*, in which avatars' eyes are controlled by head-mounted mobile eye trackers worn by participants, to a *gaze model* informed by head orientation for saccade generation, and *static gaze* featuring non-moving eyes. We analyse task performance, subjective user experience, and interactional behaviour. While not providing statistically significant benefit over static gaze, tracked gaze is observed as the highest performing condition. However, the gaze model resulted in significantly lower task performance and increased error rate.

Keywords: Immersive Collaborative Virtual Environments, Eye Tracking, Avatars, Eye Gaze, Behavioural Realism.

Index Terms: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual Reality; H.4.3 [Information Systems Applications]: Communications Applications—Computer conferencing, teleconferencing, and videoconferencing; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation;

1 INTRODUCTION

Immersive Collaborative Virtual Environments (ICVEs) connect remote or colocated users of immersive projection technology (IPT) display systems such as the CAVE™within a spatial, social and informational context with the aim of supporting high-quality interaction [24]. ICVEs usually represent participants as avatars (a graphical representation of a human) in a shared computer-generated virtual environment (VE) [4]. Head orientation and pointing gestures are generally animated by head and hand trackers worn by participants, and can significantly contribute toward perception of other's visual attention and actions [17].

Eye gaze has been identified as a critical aspect of avatar design, and it has been shown that manipulating levels of behavioural and representational fidelity of avatar eyes can significantly impact the perceived quality of communication in ICVEs [12], as well as impacting objective and subjective performance measures during object identification tasks [21, 30]. Wolff et al. presented *EyeCVE*, a tracked-gaze ICVE platform using mobile eye trackers to drive the gaze of each participant's virtual avatar [36]. Steptoe et al. detailed trials in which participants took part in three-way conferences between remote CAVE-like IPT systems linked by *EyeCVE* [31]. Eye tracking data was used to evaluate interaction, confirming the system's support for the use of eye gaze as a communicational and management resource in multiparty conversational scenarios. Object-focused tasks were identified as an important application of ICVEs in which tracked eye gaze could benefit remote interaction.

ICVEs were first evaluated in relation to the experience of users in Schroeder et al. [27]. Two participants collaborated to complete a simplified Rubik's cube puzzle over varying technologies including IPTs and standard desktop displays. We take that study as primary inspiration for this paper's experimental design, which presents a structured three-party scenario, in which the operational importance of eye gaze as a communicational resource is emphasised. However, we shift the focus from a technology comparison to the investigation of how quality of communication is impacted by varying methods of avatar eye gaze control.

There are several novel elements of this research. *EyeCVE* is the first telecommunications system (ICVE, video-based or otherwise) in our judgement that is able to preserve communicational eye gaze [31] while allowing users to move freely (physically in the IPT and virtually in the VE). This paper presents the first investigation of a task which requires full body movement. In terms of ICVEs, *EyeCVE* is the first example to capture users' eye gaze for real-time representation and post-analysis. Finally, although three IPT systems have previously been networked to support object-focused interaction [24], this is the first time that interactional analysis has been performed on the experimental sessions with the additional channel of eye gaze. This not only has implications for our experimental design, but also for the remote interactions themselves, and subsequent analysis of data which is naturally richer than previous communication captured in ICVEs.

The current study investigates the impact of tracking the eye gaze of ICVE users to drive their avatar's eyes. We compare tracked gaze (eye and head tracking), a gaze model (gaze simulation and head tracking), and static gaze (head tracking but no gaze animation). The experiment was performed between three people in three networked IPTs: two confederates and one participant, in which the confederate pair issued a series of instructions to the participant towards solving a simplified Rubik's cube puzzle. We collected data from multiple sources at all three sites, and have generated a dataset that will be of interest to various research fields. We divide analysis into task performance, user experience and interactional behaviour. For issues of readability, throughout this paper we use the terms "eye gaze" and "gaze" interchangeably. The term "user experience" means how a person perceives and rates their time using the system as opposed to their level of expertise.

2 RELATED WORK

2.1 Small-Group Interaction in ICVEs

In an early evaluation of object-focused interaction supported by (non-immersive) CVEs, Hindmarsh et al. [14] suggested that some of the collaboration limitations observed in such systems could be alleviated by IPTs due to their intuitive head and hand tracking and a larger field-of-view (FOV). The aforementioned study by

*W.Steptoe@cs.ucl.ac.uk

Schroeder et al. [27] did not explicitly analyse interaction, but rather compared task performance over varying display technologies. The authors noted that performance (time to complete the task) in a networked IPT setting was almost as good as in the real world setting, and significantly superior to asymmetric sessions between an IPT and a desktop system. The mean completion time in the ICVE setting was 8.82 minutes, and no structure was imposed on the collaboration. In contrast, we have designed the experiment presented in this paper as a defined sequence of subtasks (thus allowing for finer post-analysis of interactional elements), within an involving and cohesive whole, albeit less collaboratively-natural. Hence, our paradigm combines elements of Schroeder et al.'s experiment, with the structured interaction seen in Roberts et al.'s study on constructing a virtual Gazebo between networked IPTs [24].

Placing users in a situation closer to the natural setting in which real world collaboration may take place, Steed et al. [29] investigated the use of ICVEs to perform various tasks over an extended period of time. Participants were able to collaborate intuitively, and the ICVE lent itself particularly well to highly spatial and interactive tasks. However, negotiation tasks were difficult due to the avatars' absence of facial expression, and understanding the intentions and activities of other people remained a hindrance. This complies with subsequent work by Hindmarsh et al. [15], which noted that although the lack of information about others' actions is alleviated somewhat by the surrounding nature of ICVEs, there are still problems due to the representation of other users. An observation particularly relevant to this work is that actions such as eye gaze are not captured or transmitted by the medium. Indeed, capture of nonverbal behaviour is considered essential in order to support remote interactions that are more similar to those in the real world [28].

2.2 Avatars

User embodiment is a fundamental issue for shared VEs, and is typically maintained using an avatar: a graphical representation of a human [4]. Avatars generally exhibit a humanoid form of varying representational fidelity. This form grants a direct relationship between the natural bodily movement of a user, and the corresponding behaviour of their avatar. This control metaphor becomes critical in multi-user VEs, as participants' representation is used directly as a communication mediator [8]. Avatar representation in ICVEs has additional function to those defined in single-user VEs [7, 5] including determining position, identification, visualisation of focus of attention and recognising gesture and actions [33].

Virtual humans are capable of eliciting appropriate responses from observers, and it has been shown that unwritten social norms such as proxemics and unease from close-range mutual eye contact with unknown partners occur in the virtual world similarly to the real world [3, 37]. Avatars exhibiting higher levels of visual and behavioural fidelity can potentially communicate more subtleties of human nonverbal communication, enhancing the perceived authenticity of the interaction [35]. The interaction effect between visual and behavioural fidelity indicates that the impact of identical behavioural traits change in relation to the avatar's appearance: higher visual realism benefits from consistently realistic behaviour, while the lower fidelities also benefit from such consistency [11].

2.3 Eye Gaze Control

Meaningful representation of eye gaze information has long been recognised as a requirement for natural communication in visual remote collaboration and conferencing systems [1]. This is a logical extension of Argyle's conviction that gaze is of central importance in social behaviour and nonverbal communication, where it is used as a bidirectional channel monitoring initiation, maintenance and termination of messages [2]. As the complexity of VEs and avatar behaviour increases, it becomes more difficult to maintain a direct correlation between the user's wishes and the avatar's actions [23]. Many aspects of nonverbal communication are currently too complex and temporal to be directly tracked, and therefore models controlling various behavioural channels have emerged [13].

Gaze models have been developed with the aim of simulating naturalistic eye movement for virtual characters. Parameters representing behavioural properties such as fixation point and duration, and saccade magnitude and velocity, generally act as input to a broader analytical model. These parameters implement statistical generalisations of human gaze behaviour derived from empirical studies of saccades and/or statistical models of eye tracking data [18]. Modifying the input parameters can impact the perceived mental state of an avatar (excited or sleepy [10], dominant or submissive [16]). In order for avatars to meaningfully contribute to communication, their animation needs to reflect some aspect of the interaction that is taking place [34]. Likewise, avatars exhibiting gaze behaviour that is directly related to the current interactional state have been shown to significantly improve the perceived quality of communication compared to static or random gaze [19].

Another driving force behind the emergence of behavioural modelling is the improved visual realism of virtual humans in computer graphics. As noted, consistency between behavioural and representational fidelity can benefit the perceived social presence of an avatar. Figure 1 shows the avatars used in Schroeder et al.'s original cubes study and those used in the current work. It is clear that the visual fidelity of the former is unable to support the enhanced behavioural fidelity of eye gaze, as the eyes are formed by two black points. In contrast, if the latter avatar does *not* exhibit eye movement, observers may consider it "lifeless" [18].

Figure 1: *Left*: Avatars used in Schroeder et al.'s [27] 2001 study (courtesy of Ralph Schroeder). *Right*: Avatars used in current study.

The intuitive method for controlling an avatar's eye gaze is using eye tracking to faithfully replicate a user's eye movements. Steptoe et al. [31] presented trials performed in such a tracked gaze ICVE system (*EyeCVE*) to support three-way conferences between networked IPT systems. Despite *EyeCVE*'s support for the use of eye gaze as a communicational resource, the trials were limited as participants were required to remain seated, and the interactions were not rigorously analysed due to the small quantity of data. In the current study, we present a multiparty object-focused scenario and multidisciplinary analysis methods to determine the benefit of using tracked gaze over the established (model) and no (static, just head orientation) methods of gaze control.

3 Motivation and Hypothesis

Faithful replication of nonverbal behaviour is considered essential in order to facilitate natural avatar-mediated collaboration in shared VEs [28]. Eye gaze is of central importance in social behaviour and nonverbal communication [2]. Therefore, representing participants' real-time eye gaze as a behaviour exhibited by their avatars has the potential to enhance remote interaction in ICVEs toward a level more similar to the "gold standard" of face-to-face encounters in the real world. We compare three methods of eye gaze control:

- *Static gaze:* centred eyes with no gaze control. Thus, eye gaze information may be inferred from head orientation.
- *Gaze Model:* a simulation of gaze behaviour (section 4.2).
- *Tracked Gaze:* head-mounted eye tracking captures a user's gaze to drive their avatar's eyes in real-time (section 4.1).

Our hypothesis is that tracked gaze will result in higher quality of communication, as measured by task performance, subjective user experience, and interactional analysis, in comparison to static gaze and our gaze model. Hence, we expect superior task performance, and participants to rate tracked-gaze avatar communication as more natural and less ambiguous than the static and model conditions. Finally, we expect participants to demonstrate more natural behavioural practices and be able to employ repair strategies more efficiently than static and model control.

A fundamental feature of tracked-gaze ICVE systems is the ability for users to move freely (within the IPT) while maintaining meaningful communicational eye gaze, and problems of parallax as experienced in spatially-aligned videoconferencing systems [22] do not occur. ICVEs also grant the ability to share and manipulate virtual objects in a consistent shared space. These features are yet to be investigated in a tracked-gaze ICVE [31]. Consequently, this study explores the impact of varying eye gaze behaviour on communication in ICVEs during a multiparty object-focused task that requires free movement. We add to existing knowledge regarding the role of gaze in ICVEs, and present interactional analysis in which eye gaze data plays a central role.

4 Technical Description

4.1 ICVE Platform: *EyeCVE*

EyeCVE is built on OpenSG®, and uses RakNet™ for the network layer. At the universities of Salford and Reading, the system is run on SGI Prism™ machines, while UCL uses a Windows® XP cluster. Salford runs the centralised server, which maintains the state of the virtual world and distributes changes to all clients. Audio communication held over Skype™ conference calls external to *EyeCVE*. For a full technical description see Wolff et al. [36].

Binocular eye trackers from Arrington Research, Inc. were mounted on the IPTs' CrystalEyes® 3 shutter glasses as shown in Figure 2. The eye trackers provided robust data capture under varying lighting conditions with all subjects and minimal calibration time. Scene cameras were mounted on the modified shutter-glass frame, providing a very wide 150° FOV, granting peripheral view of the scene close to human limitations (generally considered to approach 180° forward-facing [9]). The wearer's foveal fixation point is overlaid on the scene-camera video, allowing for real-time observation and post-session analysis of gaze behaviour once synchronised with the separately recorded audio.

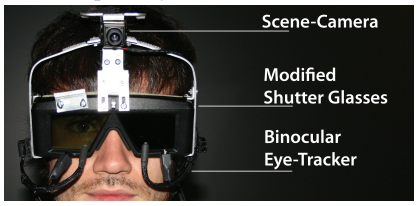

Figure 2: Mobile eye tracker on IPT shutter glasses.

Data from the calibrated eye tracker (2D gaze coordinates) are streamed to each local *EyeCVE* client. Secondary calibration of the eye tracker within the VE itself updates the eye gaze of each avatar to reflect the wearer's fixation point in 3D-space. The overall latency from eye tracker to graphical update on remote clients was measured at 150ms over the sites' standard Internet connections, thus supporting perceptually natural communicational gaze [36].

4.2 Eye Gaze Model

We developed a simple eye gaze model in order to compare against static- and tracked-gaze avatars. In previous studies into dyadic conversational scenarios in ICVEs, using a gaze model that is informed by interactional states (i.e. speaking and listening) has been found to significantly improve the perceived quality of the avatar-mediated communication [12, 18]. Therefore, in the current study, it was important to compare static gaze and the novel tracked gaze to a simulation of gaze behaviour (essentially autonomous eye animation). In order to provide meaningful gaze to a given interaction, a model should be tailored (or configurable) to the properties of a VE scenario. In this case, the current study presents a multiparty object-focused task including conversation and free movement. Hence, we developed a simple gaze model which would take into account an array of objects and avatars in a user's current FOV (determined by head tracking) to generate gaze behaviour. The basic model developed for this study intends to provide a baseline performance measure for future work into gaze modelling.

The position of the head and the eyes are the two contributing factors in the perception of where a person is looking [32]. Our gaze model takes head orientation as recorded by the IPT's head tracker, to determine the controlling participant's approximate FOV, defined as 70° from the head-centric vector. Therefore, the maximum horizontal and vertical foveal rotation to a fixation point is 35°, which is specified as the maximum ideal rotation by Boff and Lincoln [6]. Saccades and fixations are randomly distributed between targets (cubes and avatars' eyes) within the current FOV. Thus, as users move their heads, potential targets enter and exit the FOV, and new saccades and fixations will be generated. Fixation duration is therefore dependent on timing and velocity of head movement: reduced activity generates fewer saccades with longer fixations, while rapid motion results in greater numbers of saccades with shorter fixation times. In summary, our gaze model is informed by a user's current FOV inferred from head orientation to generate eye gaze animation throughout an unfolding interaction.

5 Experimental Design

The overall goal of our experiment was to determine the differences between varying methods of avatar eye gaze control as measures of task performance, subjective user experience and interactional behaviour. The conditions were tracked gaze, our gaze model and static gaze. A within-subjects, repeated measures design was conducted over the three conditions, which were resequenced to negate learning bias. Twelve normal or corrected-to-normal vision participants (no previous IPT experience) were recruited and paid £10 to perform the study. The experimental design involved two confederates and one participant. The same confederates performed all experiments, and were male. Therefore, to eliminate experimental bias, we recruited only male participants. Confederates were blind to their own and other avatars' eye gaze condition, which intended to negate any inter-experimental effects resulting in differing behaviour. Participants were not informed of the significance of other avatars' eye movements, nor the ability to communicate during the one tracked-gaze session using their own eyes. The setup procedure preceding each of the three sessions was identical, so that the participant's experience would be the same for each condition, and so the confederates would have no information to guess the condition.

For all conditions, the participant was required to *follow a defined series of instructions* issued by the two confederates to arrange eight cubes to form a single larger cube with each side displaying exactly one colour. Therefore, the puzzle's goal state is identical to the Rubik's cube task introduced by Schroeder et al. [27], which required two naive participants to work together to complete the puzzle, and should hence be considered as a symmetrical collaborative task. Accordingly, their task design sought to answer the paper's titular question: "is collaboration in ICVEs as good as being there together?" Contrastingly, the current study redefines that experimental task as a structured scenario, divided into a specific sequence of paired instructions. The instruction pair is composed of a *grab instruction* and a *position instruction*, which accordingly, requires the participant to take action to identify and grab an indicated cube and subsequently to identify and place the cube in the indicated position. Hence, this study does not present a symmetrical

collaborative task; rather, we aim to determine how the behaviour of avatar eye gaze may influence task performance, user experience and interactional behaviour as metrics of quality of communication.

5.1 Virtual Environment

Three VEs featuring different starting configurations for the three experimental sessions were built. As stated, conditions were resequenced to negate learning bias, while the VE order presented to all participants remained constant, in order to negate the impact of any variation in difficulty of the starting configurations. The spatial volume of the VEs was approximately equal to that of the IPT (3x3x2.2 metres). This allowed participants and confederates to physically move in the IPT without requiring the use of the wand to navigate. Each VE was populated with a configuration of eight puzzle cubes and five spare cubes. In each VE, three of the puzzle cubes were initially configured correctly to form the base of the solution, while the remaining five were scattered around the VE. The five spare cubes were also scattered around the VE close to the puzzle cubes. The purpose of the spare cubes was to increase task difficulty, and their positioning and quantity were informed by preparatory pilots. The spare cubes appeared coloured, and thus indistinguishable from the puzzle cubes in the participant's IPT, but appeared black in the confederates' IPTs in order to reduce confusion when performing the experiment script.

5.2 Procedure

Upon entering the IPT, participants were greeted by the confederate avatars within one of the three VEs. A short training session on navigation (i.e. walking and moving) and grabbing, rotating, and positioning cubes was given by the confederates. Once the participant was comfortable, the confederates explained that they needed help from the participant to finish the semi-completed puzzle. The goal state was defined as forming a larger cube from eight single cubes so that each side (including the bottom) of the completed cube consisted of only one colour. It was then made clear to the participant that they were *not* required to solve the puzzle themselves; rather, they would be guided by the confederates over a sequence of instructions which, if followed correctly, would result in the correct solution to the puzzle.

Confederates then commenced to give a defined sequence of instructions which naturally varied for each of the three VE configurations. Instructions were defined as a pair comprising of one grab instruction and one position instruction for each of the five missing cubes, resulting in a total of ten instructions per session.

- *Grab Instructions* required the participant to identify and pick up a specific cube from within the assortment of puzzle cubes and spare cubes scattered around the VE.
- *Position Instructions* then required the participant to place the correctly grabbed cube into a specific position in the puzzle.

Confederates took turns to issue a grab instruction and position instruction pair. The confederate at Reading always started the sequence, and hence issued three instruction pairs per session while the confederate at Salford issued two. During the task, confederates were limited to gesticulation using their eyes and heads, while hand pointing (and consequently grabbing cubes) was not allowed. Confederates were able to move freely similarly to participants within the IPT, but movement was restricted to periods between giving instructions. Verbal communication was restricted such that providing information regarding the colours of the indicated cubes was not allowed, and only using "this" (rather than "that", which may infer proximity) was allowed. Hence, all instructions were formed by variations on the two-part clause of *a) "Pick up this cube"* and, once the participant had grabbed the correct cube, *b) "Place the cube here"*. Confederates were free to answer questions that did not reveal the indicated cube or position, and would frequently talk between themselves and the participant during the task.

Figure 3: Screenshots of an experimental session taken using replay and analysis tool [20]. *Top*: The avatars (participant located in the middle) greet each other before commencing the task. The initial base can be seen between two groups of loose cubes including five spare cubes which appeared black to the confederates but coloured to the participants. The puzzle base is located in the centre of the IPT allowing for navigation simply by walking. *Middle*: The participant responds to a grab instruction by reaching for the indicated cube. *Bottom*: The participant responds to the final position instruction by positioning the cube to complete the puzzle.

Therefore, in a typical instruction pair, a confederate issues the grab instruction with a verbal utterance while indicating naturally with eye gaze and head orientation which cube they are intending. The participant responds either by choosing the correct cube or an incorrect cube. If the participant identifies an incorrect cube, the confederate informs him, and continues to target the intended cube as before. Once the correct cube is grabbed, the confederate issues the position instruction. The defined verbal grammar combined with eye gaze and head orientation is similarly employed to indicate the correct placement of the cube within the puzzle. Accordingly, if the participant incorrectly identifies the cube's placement, the confederate informs him and continues to indicate the intended position with the defined cues. Once the cube is correctly positioned, the instruction ends and the by-standing confederate begins the next (often following brief talk). This process continues until all five instructional pairs are complete, and the puzzle is solved. Figure 3 illustrates states during an experimental session.

Using this structured design, eye gaze (and head orientation) was emphasised as the predominant nonverbal resources supporting the pseudo-collaborative interactions. Consequently, differences in task performance, user experience and interactional behaviour can

be correlated with the varying behaviour of eye gaze. In future work, we intend to investigate symmetrical collaborative scenarios by introducing eye tracking to the unrestricted use of standard head and hand tracking as communicational resources in ICVEs. In the current study however, it was essential to isolate eye gaze as the central communicational resource in order to investigate the influence of varying methods of eye gaze control.

5.3 Recorded Data

With ethical clearance, we recorded the following data at all sites:

- *Video & Audio:* recorded from the eye tracker scene cameras worn by the participant and both confederates. The wearer's eye gaze is overlaid on the video, identifying fixation point.
- *Replay Logs:* session logs recorded for our replay and analysis tool [20]. The tool provides identical visualisation of experimental sessions, free camera movement, and additional data such as eye gaze heat-maps, and operates in the IPT.
- *Eye Tracker Logs:* detailed logs of eye gaze behaviour.
- *Gaze Model Logs:* eye gaze generated by our gaze model.

Following each session, the participant completed an oral questionnaire designed to elicit responses regarding subjective experience, copresence, and self-performance ratings, on a 1..7 Likert scale. The questionnaire was performed orally due to the lengthy setup and eye tracker calibration procedure and in order to maintain the participant's position within the IPT. Finally, we performed an informal interview with each participant after the third session. We targeted any general differences they may have noticed between the sessions, and then narrowed the topic down to the behaviour of the avatars and finally the eyes.

6 Analysis

We approach the analysis of the data captured from the experimental sessions from three perspectives: *Interactional Analysis*, which explores differences and trends in participant behaviour, focusing particularly on eye gaze patterns, bodily movement, and repair strategies, *User Experience*, which considers questionnaire responses and interviews, and *Task Performance*, of which the following metrics are identified:

- *Grab Instruction Errors and Timings:* Number of incorrect cube identifications following a grab instruction, and the time taken from when the instruction is issued until the *correct* cube is identified by the participant.
- *Position Instruction Errors and Timings:* Number of incorrect cube placements following a position instruction, and the time taken from when the instruction is issued until the cube is placed in the *correct* position by the participant.
- *Puzzle Completion Time:* Time elapsed between the issue of the first grab instruction and the completion of the fifth position instruction.

6.1 Task Performance

Our experimental design subdivided each session into five pairs of grab instructions and position instructions. Hence, the following task performance analysis firstly considers grab instructions, and secondly position instructions, with the associated metrics of errors and time. Finally, puzzle completion time is analysed, together with the impact of training effects as participants become more experienced at the task. In this way we are able to quantify performance over the three eye gaze conditions.

6.1.1 Grab Instruction Errors and Timings

Repeated measures two-way analysis of variance (ANOVA) calculations were separately performed for grab instruction errors and timings, with the three eye gaze conditions and the five grab instructions as factors. Regarding grab instruction errors, a significant difference between conditions ($F(2,4)=5.30;P<.01$) and also between instructions ($F(2,4)=4.91;P<.001$) was found. Post-hoc calculations determined that these differences lay between the gaze model and the other two classes of static gaze ($F(1,4)=7.36;P<.01$) and tracked gaze ($F(1,4)=6.84;P=.01$). No interaction effect was found between eye gaze condition and grab instruction.

Similarly, regarding timings, a significant difference between conditions ($F(2,4)=3.67;P<.05$), and between grab instructions ($F(2,4)=5.0;P<.001$) was found. Likewise, post-hoc calculations determined that the significant differences lay between the gaze model and the other classes of static gaze ($F(1,4)=4.35;P<.05$) and tracked gaze ($F(1,4)=4.5;P<.05$), and no interaction effect was found between eye gaze condition and grab instruction.

In summary, tracked gaze and static gaze resulted in significantly fewer errors and faster operation time than the gaze model while performing grab instructions. Tracked gaze and static gaze elicited an equal number of errors over all participants, while tracked gaze resulted in slightly faster operation time while performing grab instructions. Table 1 shows mean grab instruction errors and time per session for all participants.

6.1.2 Position Instruction Errors and Timings

Repeated measures two-way ANOVA calculations were separately performed for position instruction errors and timings, with the three eye gaze conditions and the five position instructions as factors. Regarding position instruction errors, no significant difference between the conditions ($F(2,4)=0.64;P=0.53$) was found. However, a significant difference was found between instructions ($F(2,4)=5.03;P<.0001$). No interaction effect was found between condition and position instruction. Correspondingly, regarding timings, no significant difference between conditions ($F(2,3)=0.91;P=0.4$) was found, but a significant difference between position instructions was found ($F(2,3)=3.94;P<.0.01$). No interaction effect was found between eye gaze condition and position instruction.

In summary, no significant differences between conditions were found regarding errors and time while performing position instructions. Table 1 shows slightly superior results of tracked gaze over static and model gaze for position instruction errors and timings.

Table 1: Mean task performance per session in *number of errors* and *time in seconds* (and standard deviation) for each condition.

Condition	Static	Model	Tracked
Grab Errors	1.67 (1.44)	3.42 (1.78)	1.67 (1.37)
Grab Time	48.0 (22.3)	68.1 (34.2)	47.6 (20.6)
Position Errors	0.75 (0.87)	0.50 (0.90)	0.47 (0.67)
Position Time	24.2 (18.5)	26.6 (14.3)	20.9 (9.7)
Total Time	270.1 (67.2)	301.9 (133.4)	310.3 (132.8)

6.1.3 Puzzle Completion Time

A one-way ANOVA evaluation of the overall puzzle completion time of each session did not expose any significant difference between conditions ($F(2,33)=0.41;P=0.67$), with values shown in Table 1. It should be noted here that it is reasonable to expect that avatar gaze behaviour is not a major determinant of overall completion time, which was rather dominated by a multitude of factors regarding the participant's behaviour in the VE as discussed in section 7. Hence, we performed subsequent analysis into how completion time varied with experience. In this case, we performed a one-way ANOVA on each session's puzzle completion time, but used session *order* (first, second, third) as factors (rather than eye gaze conditions as factors). A significant difference between orders ($F(2,33)=9.06;P<.001$) was found. The mean completion times in seconds (and standard deviation) for the first, second and third sessions were 370 (101.8), 304.3 (102) and 207.9 (75.1). As stated in section 5.1 the three VEs were not resequenced, so any influence of varying difficulty of starting configuration difficulty cannot be

made. However, the highly significant difference between session order and lower puzzle completion times suggest that performance improves greatly with experience.

6.2 Subjective User Experience

Questionnaires after each session sought to elicit judgement regarding user experience, self-performance ratings and copresence. Repeated measures two-way ANOVA taking the three eye gaze conditions and the fourteen questions exposed no significant difference between conditions ($F(2,13)=0.24; P<.0.79$). Although not statistically significant, participants generally rated the tracked-gaze session higher than static gaze and in particular the gaze model. This was supported by the summary interview, in which the tracked-gaze sessions were most commonly identified as the "easiest", while and sessions under the gaze model were most commonly identified as the "most confusing". However, under all conditions, the subjective judgement of the experience and self-performance ratings were consistently high, indicating the effectiveness of the avatar-mediated remote interactions and high involvement in the task.

6.3 Interactional Analysis

The following section identifies specific behavioural practices and resources that participants are seen to employ during interaction, including repair strategy. The foundation of this analysis has been drawn primarily from performing conversation analysis (CA): an inductive process for analysing how human interaction is organised into sequences of action or systematic practices [26]. Due to space, we are unable to show any CA transcripts, but examples will be made available online. Firstly we present typical cases of how interaction unfolds during correct and incorrect responses to a grab instruction. We then identify general behavioural strategies and resources adopted by participants. Finally, we performing post-hoc analysis on a specific gaze practice used extensively by all participants as a repair strategy. It should be noted that this section primarily analyses interactional behaviour associated with grab instructions: analysis of position instructions was not so revealing, being easier (moreso as the puzzle progressed), and with participants often attempting to identify placement position before the confederate issued an instruction, especially with experience.

6.3.1 Responses to a Grab Instruction

Participants responded to grab instructions either by correctly or incorrectly identifying the target cube. A typical case of a correct and unproblematic response to a grab instruction begins with a confederate summoning the participant's attention and eliciting his gaze. Next the confederate produces a pre-instruction such as "so, my turn", which prompts the the participant to respond with "OK", and maintain his focus on the speaker. When the grab instruction is given, the confederate typically turns and looks at the designated cube as the deictic term "this" is uttered (i.e. "Can you pick up *this* cube."). The participant then follows the confederate's gaze towards the group of targeted cubes, glancing between the cubes and the confederate (usually focusing on the eyes). Once he has identified which cube he believes is being indicated, he looks straight at that cube and touches it while simultaneously requesting confirmation and briefly returning his gaze to the instructing avatar. The speaking confederate gives confirms the choice as correct, and the participant completes the instruction by grabbing the cube.

By contrast, cases of grab instructions in which the participant selects an incorrect cube often provide examples of repair strategies to identify the correct cube, and occasionally even to reconfirm which avatar is currently issuing an instruction. In such a case, confederate *A* issues a grab instruction, but the participant is unsure about which avatar is speaking (avatars did not have mouth-movement as discussed in section 7), and looks from confederate *A* to confederate *B* and back. Confederate *A* establishes that he is speaking by moving towards and gazing at the participant, and reissues his grab instruction. As he does this, the participant shifts his gaze to the potential cubes. After glancing between the speaking avatar and the cubes, the participant chooses a cube and grabs it without confirmation that it is the correct choice. He then gets feedback from confederate *A* that his choice is incorrect. Finally, the participant moves closer to the confederate avatar to examine his gaze, glancing to and from the potential cubes, before selecting the correct cube on a second attempt.

In summary, we can determine that incorrect responses to instructions have the potential to significantly reduce the participant's task performance, in terms of number of errors, time and possibly user experience (judging the interaction as confusing). We also see participants' strategic use of other avatars' gaze, following an incorrect identification to attempt to repair the mistake. This clarifies the the need for accurate and unambiguous indication.

6.3.2 Movement as a Resource

Confederates' proximity to cubes seems very useful to participants when determining which cube to grab. Correspondingly, confederates made extensive use of the ability to move freely to position themselves at an appropriate distance and clear angle (vertically and horizontally) to the intended cube. Above we noted how participants often moved toward the instructing avatar to more clearly determine his direction of gaze. This is one category of a number forms of locomotion that participants engaged in. In many cases, participants look at the avatar's face and then take up an opposite position behind the cubes in order to identify the target cube. Another strategy adopted by participants is taking a position similar to the one of the instructing avatar, by standing beside them or looking over their shoulder, presumably to see from the confederate's perspective. In this way, participants made use of several resources provided by the ability to move freely and maintain communicational eye gaze to attempt to perform the task more effectively.

6.3.3 Eye Gaze as a Resource

Eye gaze is employed as a general social resource as previously observed [31], and the participant shares their gaze between both confederate avatars during periods of silence and conversation, and when carrying out instructions. Additionally, participants are frequently seen to inform their actions on where they perceive the instructing avatar to be gazing, and estimations are developed after looking at the head, and in particular the eyes of the confederate avatars. The gaze model in particular caused difficulties when responding to instructions, frequently coming to the surface of the interaction via participant vocalisation types including questions (i.e. "Are you looking at it now?"), complaints (i.e. "You're looking up! Not down.") or accounts (i.e. "It's a little difficult for me because your eyes keep moving about."). Such behaviour highlights the importance of eye gaze in avatar-mediated communication, where it seen to be used as a bidirectional nonverbal communication resource similarly to face-to-face interaction, thus promoting tracked gaze as potentially the most effective means of a controlling an avatar's eye movement.

6.3.4 Glances Per Grab Instruction

As mentioned in Section 6.3.1, a strategy adopted by all participants, and for 98.9% of grab instructions (176/180 over all participants and sessions) was to alternate their gaze between the instrucing avatar and the potential cubes. In this way, participants were able to develop their judgment as to which cube was being indicated, and the method was also employed as a repair strategy following an incorrect action. Hence, fewer glances required before correctly responding to a grab instruction not only reduced the amount of time required per instruction, but also indicated the abil-

ity of the eye gaze control method to unambiguously and efficiently indicate specific objects without the need for such repair strategies.

A repeated measures two-way ANOVA calculation for glances per grab instruction was performed, with the three eye gaze conditions and the five grab instructions as factors. A significant difference between conditions ($F(2,4)=8.23$;$P<.0005$) and also between instructions ($F(2,4)=4.54$;$P<.002$) was found. As found when analysing grab instruction errors and timings, post-hoc tests determined that the differences lay between the gaze model and the other two classes of static gaze ($F(1,4)=8.81$;$P<.005$) and tracked gaze ($F(1,4)=10.95$;$P<.001$). No interaction effect was found between condition and grab instruction. Table 2 shows the mean number of glances performed per grab instruction under each condition. Although not statistically superior to static gaze, tracked gaze is identified as the highest-performing method of eye gaze control.

Table 2: Mean number (and standard deviation) of glances to and from the instructing avatar and target cubes per grab instruction.

Condition	Static	Model	Tracked
# of Glances	2.42 (1.5)	3.58 (2.74)	2.25 (1.57)

7 Discussion

We analysed our experimental data from the perspectives of task performance, user experience and interactional behaviour, as measures of quality of communication. Results calculated in each category demonstrated high levels of correlation, indicating that our gaze model is significantly outperformed by static gaze and especially by tracked gaze. However, differences between static gaze and tracked gaze were not found to be statistically significant. Thus, we cannot support our initial hypothesis that tracked-gaze avatars are able to enhance the quality of communication when performing object-focused scenarios in ICVEs.

Despite this, metrics of task performance (Table 1) and interactional behaviour (Table 2) as well as trends from our informal post-experimental interviews (section 6.2) ranked tracked gaze as the superior method of eye gaze control. Intuitively, this of course should be the case. However, the experimental interactions were highly complex, and responses to grab instructions and position instructions were informed by an amalgamation of factors aside from how other avatars' eyes behaved. We now discuss factors influential to the perception and performance of the experimental interactions (including paths for future investigation) which are likely to have detracted from the enhanced behavioural fidelity and additional information communicated by tracked-gaze avatars. Finally, we examine the poor performance of our eye gaze model.

7.1 Naïve Users and Task Complexity

The experiment situated inexperienced IPT users at the focal role in an unfamiliar multiparty puzzle scenario mediated by avatars in an ICVE. They were required to communicate with confederates, navigate, make judgements, manipulate objects, and conduct questionnaires over three experimental sessions lasting up to two hours. In this demanding situation, it can be reasonably proposed that there were many factors influencing participant performance and perception, and that eye gaze was not the sole determinant of action. However, while this may be the case, the poor performance during the gaze model sessions confirm that eye gaze can be highly influential to participant action. It appears that head orientation (static gaze) and body positioning provided participants with sufficient nonverbal information in order to perform the task effectively. While this finding contradicts a previous object-identification test performed in a non-immersive stereo display [21], it is consistent with our parallel study in which a discussion of the impact of the IPTs' display properties can also be found [25]. Thus, varying task complexity, and performing the experiment with experienced ICVE users may reveal the potential of using tracked-gaze avatars.

7.2 Avatar Representation

Central to avatar-mediated communication common to ICVEs, is the visual representation of the avatars themselves. In this experiment, we used a single avatar model to represent all three participants, which, while featuring photographic textures, did not have particularly high representational fidelity in terms of geometry and overall realism. Also significant was the lack of eyelid movement and blinking as promoted in [30], causing the sclera to appear very large, and lack of lip-synchronisation, causing difficulty for participants to determine the speaking confederate. Recalling the interaction effect between visual and behavioural fidelity [11], investigation into varying the representation of our avatar models in combination with the behavioural fidelity of eye gaze, may expose significant differences in quality of communication.

7.3 Gaze Model Performance

Participants performed consistently and significantly worse in the gaze model condition than with static gaze and tracked gaze. This result confirms the interactional-importance of eye gaze in avatar-mediated interaction in ICVEs. Figure 4 reveals why the gaze model underperformed, by comparing the proportion of equality of the actual eye gaze fixation point (eye-target, ET), the forward vector of the head orientation (head-target, HT), and the gaze model fixation point (model-target, MT) over all participants and sessions. The low equalities shown for the model-target demonstrate that the model rarely generated gaze fixation points equal to the actual eye-target (mean 9.2%, SD 2.8%) or the head-target (mean 13.56%, SD 2.32%) of the participant. Hence, while informed by head orientation to select objects within the current FOV, the gaze model usually did not pick targets that were equal to the object actually being looked at, making target-identification by observers uncertain. Based on this wrong eye gaze information, our analysis showed that participants made more errors, required more time, and relied more on repair strategies during the task. Also interesting to note is the wide variation of gaze behaviour employed by different participants as denoted by each point.

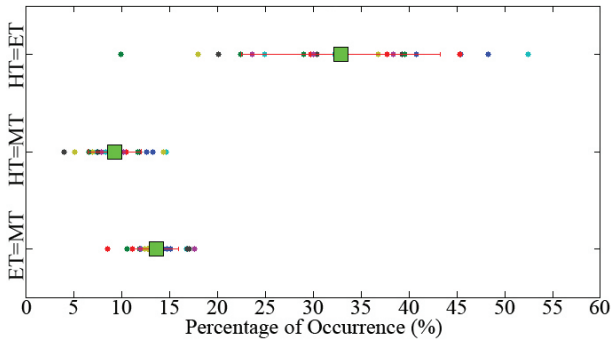

Figure 4: Percentage equality of eye-target (ET), head-target (HT) and model-target (MT) in all sessions. Small points denote individual participants. Square is mean. Range is standard deviation.

8 Conclusions and Future Work

We investigated of how quality of communication is impacted by varying methods of avatar eye gaze behaviour control in an object-focused multiparty ICVE scenario. We compared tracked gaze to static gaze, and a simple gaze model by measuring task performance, subjective user experience, and analysing interactional behaviour. Tracked- and static-gaze avatars were seen to support significantly higher quality of communication than avatars exhibiting gaze model behaviour, as measured by task performance and interactional analysis. However, our questionnaire failed to elicit significant differences in subjective user experience. While tracked gaze consistently performed better than static gaze, a statistically significant difference was not found. Factors influencing the perception

and performance of the experimental interactions that were likely to have detracted from the enhanced behavioural fidelity and additional information communicated by tracked gaze were discussed, including naïvity of users, task complexity, display properties and avatar representation, and all identified for future work.

Results elicited during the gaze model sessions demonstrate how displaying incorrect eye gaze can be a hindrance during object-focused interaction in ICVEs. This both emphasises the importance of avatar eye gaze, and the tension between animating the eyes so that the avatar looks alive, and confusing the viewer when the gaze target is critical to the task. Hence we view the development of effective and reusable gaze models as an open problem which we will address in future work. Finally, while we cannot fully support our initial hypothesis, our results suggest that the benefit of tracked gaze is likely to be demonstrable with redirected experimental and avatar design. Videos and CA transcripts of the *EyeCVE* sessions can be found at http://www.cs.ucl.ac.uk/staff/W.Steptoe.

Acknowledgements

The UK's EPSRC provides funding for this study which is part of the *Eye Catching* project (EP/E007406/1).

References

[1] S. Acker and S. Levitt. Designing Videoconference Facilities for Improved Eye Contact. *J. Broad. & Elec. Media*, 31:181, 1987.

[2] M. Argyle and M. Cook. *Gaze and Mutual Gaze*. Cambridge University Press Cambridge, 1976.

[3] J. Bailenson, J. Blascovich, A. Beall, and J. Loomis. Equilibrium Theory Revisited: Mutual Gaze and Personal Space in Virtual Environments. *Presence: Teleoperators & Virtual Environments*, 10(6):583–598, 2001.

[4] S. Benford, J. Bowers, L. Fahlén, C. Greenhalgh, and D. Snowdon. User embodiment in collaborative virtual environments. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 242–249, 1995.

[5] S. Benford, J. Bowers, L. Fahlén, C. Greenhalgh, and D. Snowdon. Embodiments, avatars, clones and agents for multi-user, multi-sensory virtual worlds. *Multimedia Systems*, 5(2):93–104, 1997.

[6] K. Boff and J. Lincoln. User's Guide Engineering Data Compendium Human Perception and Performance, 1988.

[7] T. Capin, I. Pandzic, H. Noser, N. Magnenat-Thalmann, and D. Thalmann. Virtual Human Representation and Communication in VLNET Networked Virtual Environments. *IEEE Computer Graphics and Applications*, 17(2):42–53, 1997.

[8] T. Capin, J. Wiley, and I. Sons. *Avatars in Networked Virtual Environments*. John Wiley & Sons, 1999.

[9] J. Costella. A Beginner's Guide to the Human Field of View. *School of Physics, The University of Melbourne, November*, 1995.

[10] Z. Deng, J. Lewis, and U. Neumann. Automated Eye Motion Using Texture Synthesis. *IEEE COMPUTER GRAPHICS AND APPLICATIONS*, pages 24–30, 2005.

[11] M. Garau. Selective fidelity: Investigating priorities for the creation of expressive avatars. *Avatars at work and play: Collaboration and interaction in shared virtual environments*, pages 17–38, 2006.

[12] M. Garau, M. Slater, V. Vinayagamoorthy, A. Brogni, A. Steed, and M. Sasse. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. *Human factors in computing systems*, pages 529–536, 2003.

[13] M. Gillies and D. Ballin. Integrating Autonomous Behavior and User Control for Believable Agents. In *International Conference on Autonomous Agents*, volume 1, pages 336–343, 2004.

[14] J. Hindmarsh, M. Fraser, C. Heath, S. Benford, and C. Greenhalgh. Fragmented interaction: establishing mutual orientation in virtual environments. *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, pages 217–226, 1998.

[15] J. Hindmarsh, M. Fraser, C. Heath, S. Benford, and C. Greenhalgh. Object-focused interaction in collaborative virtual environments. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 7(4):477–509, 2000.

[16] M. Kipp and P. Gebhard. IGaze: Studying Reactive Gaze Behavior in Semi-immersive Human-Avatar Interactions. *Intelligent Virtual Agents: 8th International Conference, IVA 2008, Tokyo, Japan, September 1-3, 2008, Proceedings*, page 191, 2008.

[17] S. Langton, R. Watt, and V. Bruce. Do the eyes have it? Cues to the direction of social attention. *Trends in Cog Sci*, 4(2):50–59, 2000.

[18] S. Lee, J. Badler, and N. Badler. Eyes alive. *ACM Transactions on Graphics (TOG)*, 21(3):637–644, 2002.

[19] S. Masuko and J. Hoshino. Head-eye Animation Corresponding to a Conversation for CG Characters. In *Computer Graphics Forum*, volume 26, pages 303–312. Blackwell Synergy, 2007.

[20] A. Murgia, R. Wolff, W. Steptoe, P. Sharkey, D. Roberts, E. Guimaraes, A. Steed, and J. Rae. A Tool For Replay And Analysis of Gaze-Enhanced Multiparty Sessions Captured in Immersive Collaborative Environments. *ACM DS-RT*, 8, 2008.

[21] N. Murray and D. Roberts. Comparison of head gaze and head and eye gaze within an immersive environment. *the 10th IEEE International Symposium on Distributed Simulation and Real Time Applications. IEEE Computer Society, Los Alamitos, CA*, 2006.

[22] K. Okada, F. Maeda, Y. Ichikawaa, and Y. Matsushita. Multiparty videoconferencing at virtual social distance: MAJIC design. *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 385–393, 1994.

[23] S. Penny, J. Smith, P. Sengers, A. Bernhardt, and J. Schulte. Traces: Embodied Immersive Interaction with Semi-Autonomous Avatars. *Convergence*, 7(2):47, 2001.

[24] 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 & Virtual Environments*, 12(6):644–657, 2003.

[25] D. Roberts, R. Wolff, J. Rae, A. Steed, R. Aspin, M. McIntyre, A. Pena, and W. Steptoe. Communicating Eye-gaze Across a Distance. *Virtual Reality Conference, 2009. VR'09. IEEE*, 2009.

[26] H. Sacks. *Lectures on Conversation*. Blackwell, 1995.

[27] R. Schroeder, A. Steed, A. Axelsson, I. Heldal, Å. Abelin, J. Wideström, A. Nilsson, and M. Slater. Collaborating in networked immersive spaces: as good as being there together? *Computers & Graphics*, 25(5):781–788, 2001.

[28] A. Steed, D. Roberts, R. Schroeder, and I. Heldal. Interaction between Users of Immersion Projection Technology Systems. *HCI International 2005, the 11th International Conference on Human Computer Interaction*, pages 22–27, 2005.

[29] A. Steed, M. Spante, I. Heldal, A. Axelsson, and R. Schroeder. Strangers and friends in caves: an exploratory study of collaboration in networked IPT systems for extended periods of time. *Proceedings of the 2003 symposium on Interactive 3D graphics*, pages 51–54, 2003.

[30] W. Steptoe and A. Steed. High-Fidelity Avatar Eye-Representation. *Virtual Reality Conference, 2008. VR'08. IEEE*, pages 111–114, 2008.

[31] W. Steptoe, R. Wolff, A. Murgia, E. Guimaraes, J. Rae, P. Sharkey, D. Roberts, and A. Steed. Eye-tracking for avatar eye-gaze and interactional analysis in immersive collaborative virtual environments. In *Proceedings of the ACM 2008 conference on Computer supported cooperative work*, pages 197–200. ACM New York, NY, USA, 2008.

[32] R. Stiefelhagen and J. Zhu. Head orientation and gaze direction in meetings. In *Conference on Human Factors in Computing Systems*, pages 858–859. ACM New York, NY, USA, 2002.

[33] D. Thalmann. The Role of Virtual Humans in Virtual Environment Technology and Interfaces. In *Proceedings of Joint EC-NSF Advanced Research Workshop, Bonas, France*, 1999.

[34] H. Vilhjálmsson and J. Cassell. BodyChat: autonomous communicative behaviors in avatars. *Proceedings of the second international conference on Autonomous agents*, pages 269–276, 1998.

[35] V. Vinayagamoorthy, A. Steed, and M. Slater. Building Characters: Lessons Drawn from Virtual Environments. *Toward Social Mechanisms of Android Science: CogSci Workshop*, pages 119–126, 2005.

[36] R. Wolff, D. Roberts, A. Murgia, N. Murray, J. Rae, W. Steptoe, A. Steed, and P. Sharkey. Communicating Eye Gaze across a Distance without Rooting Participants to the Spot. *ACM DS-RT*, 8, 2008.

[37] N. Yee and J. Bailenson. The Proteus Effect: Self Transformations in Virtual Reality. *Human Communication Research*, 2007.