Running head:

Multimodal Data Capture and Analysis of Interaction in Immersive Collaborative Virtual Environments

William Steptoe and Anthony Steed
Department of Computer Science
University College London
Gower Street
London, WC1E 6BT
email: w.steptoe@cs.ucl.ac.uk, a.steed@cs.ucl.ac.uk

Abstract

Users of immersive virtual reality (VR) are often observed to act realistically on social, behavioral, physiological, and subjective levels. However, experimental studies in the field typically collect and analyze metrics independently, which fails to consider the synchronous and multimodal nature of the original human activity. This paper concerns multimodal data capture and analysis in immersive collaborative virtual environments (ICVEs) in order to enable a holistic and rich analysis based on techniques from interaction analysis. A reference architecture for collecting multimodal data specifically for immersive (VR) is presented. It collates multiple components of a user's nonverbal and verbal behavior in single log file, thereby preserving the temporal relationships between cues. Two case studies describing sequences of immersive avatar-mediated communication (AMC) demonstrate the ability of multimodal data to preserve a rich description of the original mediated social interaction. Analyses of the sequences using techniques from interaction analysis emphasize the causal interrelationships between the captured components of human behavior, leading to a deeper understanding of how and why the communication may have unfolded. In presenting our logging architecture, we hope that we will initiate a discussion of a logging standard that can be built by the community so that practitioners can share data and build better tools to analyze the utility of VR.

Multimodal Data Capture and Analysis of Interaction in Immersive Collaborative Virtual Environments

Introduction

The act of communication is a continuous process in which discrete packages of information are imparted by a sender, and subsequently decoded and responded to by a receiver (Walther, Parks, Knapp, & Daly, 2002). People engaged in collocated communication have access to the full gamut of sensory channels, of which visual (nonverbal) and aural (verbal) cues typically dominate, although haptics, olfactics, and body temperature often play important roles in shaping the holistic experience of a social encounter. Human action and interaction is, thus, naturally multimodal, and if it is to be understood thoroughly by practitioners, must be complemented with methods of both observation and analysis that recognize this multimodality. In particular, for certain situations we would like to be able capture as much of the action and interaction as possible for potential future analysis.

Collocated interaction is *unmediated*, as it involves direct exchange of information between participants, with no intervening agents or technologies. Empirical analysis of such unmediated interaction examines a variety of synchronous activities including both verbal and nonverbal communication, and has the aim of identifying how humans perform sequences of social action and systematic practices (Sacks, 1995).

Mediated communication relies on a process by which messages are transmitted via some form, or medium, external to direct face-to-face interaction (Pavlik & McIntosh, 2004). Because the mediation is often done by electronic devices it is possible to record elements of the communication by logging data from those devices. However, depending on a medium's characteristics, specialized approaches to data collection and analysis of activity are required. In this paper we are particularly interested in immersive collaborative virtual environments (ICVEs) connect remote users of immersive virtual reality (VR) systems, such as the CAVE™, within a spatial, social and informational context, with the aim of supporting high-quality interaction (Roberts, Wolff, Otto, & Steed, 2003). Generally, ICVE users 'puppeteer' embodied graphical humanoids, or avatars, within the shared virtual environment (VE) (Benford, Bowers, Fahlén, Greenhalgh, & Snowdon, 1995). This form of immersive avatar-mediated communication (AMC) presents an alternative paradigm for interaction and its related processes of data collection and analysis in comparison with unmediated communication.

AMC in ICVEs is a relatively new topic for investigation. However there are many studies of VR and individual response to VR using a variety of methods of data collection. Methods generally fall into one of four categories: *body tracking* which captures users' kinesic behavior as records of action and involvement, including body posture (Pan & Slater, 2007; Rovira, Swapp, Spanlang, & Slater, 2009), verbal signals (Pertaub, Slater, & Barker, 2002), oculesics (Steptoe, Steed, Rovira, & Rae, 2010), and proxemics (Bailenson, Blascovich, Beall, & Loomis, 2001); *questionnaires and interviews* assessing users' subjective sense of presence and their experience of being immersed in the VE (Slater & Usoh, 1993; Roberts et al., 2003); *performance metrics*, which aim to quantify action, and are particularly suited to assessing the success and failure of collaborative tasks (Schroeder et al., 2001; Wolff, Roberts, & Otto, 2004); and *physiological response* measures to stimuli experienced in a VE, including heart rate (Meehan, Insko, Whitton, & Brooks Jr, 2002), galvanic skin conductance (Hagni et al., 2008), and pupil dilation (Steptoe et al., 2010).

While the objective methods above are able to elucidate particular elements of behavior and response in VEs, they are generally both collected and analyzed independently of one another, often with no temporal alignment, or purely in a summative, per experience manner. During collaborative tasks uses produce and respond to a range of highly interrelated verbal and nonverbal signals (Argyle & Trower, 1979). Hence, while analysis of a single data source (for instance gaze direction (Steptoe et al., 2008)) is likely to uncover common generalizable practices, it fails to consider causal concomitant signals such as talk and gesture, and hence, is unable to form holistic description of action.

This paper thus explores the area of multimodal data collection and analysis of interaction in ICVEs in order to enable type of holistic analysis of action that are derived from the field of unmediated interaction analysis (Jordan & Henderson, 1995). Our objectives are three-fold. Firstly, through a critique of both the traditional approach to collocated interaction analysis and non-multimodal analysis of mediated action in the VE literature, to establish multimodal data collection as a fruitful approach to describing user behavior in ICVEs. Secondly, to provide a reference architecture for collecting multimodal data for ICVEs that collates multiple streams of a user's action in single log file thereby preserving the temporal relationships between cues. Thirdly, to demonstrate the ability of multimodal data to preserve a rich description of the original mediated social interaction through two case studies analyzing sequences of immersive AMC. The mid-term aim of the overall work is to standardize a log description that can provide enough detail to allow third parties to analyze log files without having to replicate the original system(s) that supported the ICVE.

The *Background* section provides a discussion of the predominant, video-based, approach to unmediated interaction analysis. Realistic action in VEs is then discussed, followed by common data collection methods. The *Capture Architecture* section firstly presents methods and requirements towards a multimodal approach to data collection in VE systems. A reference implementation of the technical architecture and process of multimodal data collection is presented. The architecture is generalizable and may be adapted for implementation in any ICVE system, and supports collation of multiple tracking and input data streams, representing an individual's natural behavior, in a single log file. The *Multimodal Interaction Analysis* section presents novel analyses of two data extracts, selected from two previously published experimental studies involving immersive AMC: dyadic conversation (Steptoe et al., 2010) and triadic object-focused interaction (Steptoe et al., 2009). These analyses aim to provide a rich description of the segments of interaction they represent, and emphasize the interrelation between multiple behavioral cues. The *Discussion* section explores how, analogous to analysis of collocated interaction as described by Jordan and Henderson (Jordan & Henderson, 1995), multimodal capture and analysis of AMC in ICVEs is able to provide an excellent foundation for analytic knowledge that is superior to examining cues individually. The final section draws conclusions and gives suggestions for further work.

Background

In the background we discuss unmediated data collection and analysis, realistic behavior in VEs, and data collection and analysis in VEs. We start with collocated interaction analysis for two reasons. Firstly this is very related to the domain of use of ICVEs that we explore in our cases studies in Section : we attempt to use ICVEs to simulate collocated interaction. Secondly we discuss interaction analysis because immersive VR involves whole body interaction and the literature on interaction analysis provides a number of techniques and strategies for analysis that we can build upon. We describe the types of realistic behavior in VEs to motivate the consideration of whole user response to a system. Finally we discuss other approaches to logging in VR systems.

Collocated Interaction Analysis

Goodwin and Heritage consider social interaction as the primary means through which the business of the social world is transacted, the identities of its participants are affirmed or denied, and its cultures are transmitted, renewed, and modified (Goodwin & Heritage, 1990). Shared meaning, mutual understanding, and the coordination of human conduct are achieved through processes of social interaction. Collocated interaction is recognized both as the central form of social interaction and as a strategic site for the analysis of human activity, as interactants have at their disposal a full range of verbal and nonverbal resources. Jordan and Henderson define *Interaction Analysis* as an interdisciplinary method for the empirical investigation of the interaction of human beings with each other and with objects in their environment. The approach examines human activities such as talk and nonverbal behavior in an attempt to describe routine practices, structure, and sequential patterns throughout an unfolding interaction. Its roots lie in ethnography, socio-linguistics, ethnomethodology, conversation analysis, kinesics, proxemics, and ethology. Such detailed descriptive analyses have been vital in elucidating the fundamentals and the nuances of human behavior: for instance, providing evidence for the role that nonverbal behavior plays in the interactive process (Kendon, 1967), and demonstrating the role of gaze direction during speaker selection in multiparty conversation (Lerner, 2003). A detailed introduction to the foundations and practice of the method can be found in (Jordan & Henderson, 1995).

Since the 1960s, the analysis of unmediated collocated communication has largely been performed through manual transcription and review of audiovisual recordings (Kendon, 1967). These records allow for general and, depending on video fidelity and camera position, detailed qualitative observation of action, including nonverbal cues. Although some more recent work has employed body tracking technology in order to collect quantitative data describing collocated interaction (Vertegaal, Slagter, Veer, & Nijholt, 2001; Keysar, Barr, Balin, & Brauner, 2000; Shockley, Baker, Richardson, & Fowler, 2007), these studies often aim to investigate particular phenomena, rather than more general social activity. Moreover, the requirement and logistics of setting up wearable and often intrusive tracking equipment, such as eye tracking and motion capture technology, in an attempt to record what is intended to be ‘natural’ communication may have a distortional effect on participant behavior. Hence, video technology, which provides a relatively non-intrusive means of data capture but fails to record nonverbal behavior in a precise and quantifiable manner, remains the primary approach to data collection and analysis of unmediated interaction (Jordan & Henderson, 1995).

Only electronic recording produces the kind of data corpus that allows close interrogation required to perform the process of interaction analysis (Jordan & Henderson, 1995). Hence, development of video technology has been instrumental to the evolution of a field which depends on audiovisual recording for its primary records, and on playback capability for analysis. It is rare to find examples, even in the contemporary literature, in which analysis of collocated interaction has been performed with any other technology apart from video: Vertegaal et al. report on gaze behavior captured using eye tracking in round-table conversation (Vertegaal et al., 2001), Keysar et al. analyzed people's tracked eye and head movements when following a confederate's instructions to manipulate objects (Keysar et al., 2000), and Shockley et al. investigated postural coordination between conversing dyads using magnetic motion tracking (Shockley et al., 2007). In each case, the researchers aimed to investigate a specific cue in a particular scenario, rather than to more broadly investigate topics such as collaborative design practice, the situated nature of skill and knowledge acquisition, and human-machine interaction.

Video is likely to remain the primary mode of data collection and analysis of collocated interaction, and it has considerable ability to unintrusively record complex events. However, there are several drawbacks to the approach. Firstly, transcribing and interpreting video is an intensely manual process that generally must be performed in real-time (as the video plays), requiring frequent pauses and replay of action. This process results in analysis that is imprecise in both temporal and objective terms: video recording and playback devices are prone to minor timing errors, and observation and comprehension of action is difficult to quantify. Furthermore, a critical drawback of the video-based approach is the typically static positioning of cameras (in order to minimize movement distraction) amidst the highly dynamic and spatial nature of human interaction: as Jordan and Henderson state, compared to direct participant observation, video recordings replace the bias of the analyst with the (consistent) bias of the machine (Jordan & Henderson, 1995). Although this issue of a static viewpoint may be alleviated somewhat by the use of multiple cameras, this has the compounding effect of escalating the amount of data transcription and analysis during this highly manual process.

Hence, when studying collocated interaction, there is a trade-off between the use of audiovisual recording and body tracking. The former is able to provide a relatively unintrusive means of data capture but fails to record nonverbal behavior in a precise and quantifiable manner, while the latter is able to provide numerical evidence of phenomena but requires participants to wear intrusive equipment, which may have a distorting effect on the interaction itself. The intrusive nature of wearable body tracking devices is less of a concern when investigating interaction in ICVEs, as those devices are often a prerequisite of system use.

Realistic Behavior in Immersive Virtual Environments

The central feature of users' response to VR is *presence*, which is defined as a user's psychological response to patterns of sensory stimuli, resulting in the user having the impression of "being there", in a computer-generated space (Slater, Usoh, & Steed, 1994). There are multiple methods for measuring a user's sense of presence while immersed in a VE. Perhaps the most common approach is using questionnaires aiming to elicit subjective responses regarding the experience. This was first demonstrated in (Slater & Usoh, 1993), and a refined questionnaire later appeared in (Slater & Wilbur, 1997). The questionnaires generally feature ordinal Likert scales that anchor responses between two extremes, and have been shown to be effective in eliciting meaningful responses in many cases. However, there are a number of criticisms to questionnaire-based measurement of presence: they been shown to be unstable, in that prior information can influence the results (Freeman, Avons, Pearson, & IJsselsteijn, 1999), they may be unable to discriminate between presence in a VE and physical reality (Usoh, Catena, Arman, & Slater, 2000), and they may be prone to methodological circularity (asking questions about PI may foster the very phenomenon that the questionnaire is supposed to be measuring) (Slater, 2004).

More robust and compelling measures of presence involve quantitative analysis of behavioral response to the VE stimuli. If participants in a VE behave as if they are in an equivalent real environment, then this is a sign that they are experiencing presence. For example, the *pit room* is a classic VE which researchers have evolved through several iterations to assess aspects of presence (Usoh et al., 1999; Insko, 2001; Meehan et al., 2002). The experiment involves a participant having to navigate around an apparent drop over a cliff. It elicits a stress response from the users which is quite marked. In an experimental scenario, a participant's reactions to the visual cliff have been quantified (separately) through physiological measurements of heart rate, respiratory rate, and galvanic skin response. When compared to the baseline level elicited by the first room, people experience significantly heightened physiological response. Meehan and Razzaque investigated the influence of passive haptics on response to the pit, in which real (but small) ledges were aligned to their virtual counterparts (Meehan et al., 2002). This added to the effect of standing over a real pit, and significantly increased the heart rate of participants compared with when the ledge was absent. Thus, if the normal physiological response of a person to a particular situation is observed in a VE, this is a sign that the user is experiencing a high level of presence. The obvious drawback of such measurement methods is that they are limited to situations in which there is a significant physiological response in reality, so are less useful in mundane situations (Sanchez-Vives & Slater, 2005).

Realistic behavior and response during interactions with avatars have been demonstrated through a variety of metrics and scenarios, including measurement of interpersonal distance between users and avatars (Yee, Bailenson, Urbanek, Chang, & Merget, 2007), lack of willingness to administer virtual electric shocks to avatars (Slater et al., 2006), and intervention in response to witnessing avatars as victims of virtual violence (Rovira et al., 2009). This realistic behavior has also been observed in social scenarios comparing video-mediated communication (VMC) and AMC (Garau, Slater, Bee, & Sasse, 2001; Steptoe et al., 2008). In particular, the study from which the first of the forthcoming analysis segments is extracted ((Steptoe et al., 2010)) compared the degree of social presence experienced by users when engaged in AMC or VMC during truthful and deceptive situations. Analysis of gaze and pupil size revealed that characteristic behavioral patterns correlated with truth telling and lying were performed similarly during interaction in both communication mediums.

Data Collection in Immersive Virtual Environments

The conclusion to draw from the previous section is that human action and interaction in VEs is often observed to be ‘realistic’ on social, behavioral, physiological, or subjective levels. Thus any specific experiment might record any number of different variables, and most of the VE systems and experiments discussed in the previous two sections would have used some form of logging capability. Types of data that might be logged include body movement (either partial using a specific sensor (Frisoli et al., 2008), or in full, using motion capture (Slater, Spanlang, & Corominas, 2010)), oculesic behavior (through eye tracking (Steptoe et al., 2010)), vocal signals (using microphones (Pertaub et al., 2002)), and physiological response (including heart rate (Meehan et al., 2002) and galvanic skin response (Hagni et al., 2008)). Tracking devices used in VE systems are generally used for real-time input, such as head tracking to inform both perspective-correct stereo rendering (Cruz-Neira, Sandin, & DeFanti, 1993) and avatar position (Benford et al., 1995), and hand tracking to both interact with the VE (Bowman & Hodges, 1997) and drive avatar gesture (Badler, Hollick, & Granieri, 1993). However, an orthogonal application of user tracking is precise data collection, which may be subsequently used to analyze users’ behavior, response, and performance. For example, in (Bideau et al., 2004) precise measures of human motion are made in order to evaluate the performance of goalkeepers, and (Frisoli et al., 2008) characterized the main features of stroke gesture performed during in-door rowing.

Most of the devices that measure such data would provide a logging capability. For example, a motion capture systems typically export to motion formats such as BioVision Hierarchy, and a physiology measurement device such as NeXus-4 from Mind Media B.V. can log to its own format. Such formats are useful, but for integration with other sensor data a neutral format is often used. Perhaps most notably, VRPN provides a device-independent and network-transparent interface to VR peripherals (Taylor et al., 2001). VRPN includes a robust logging system, capable of synchronizing and recording multiple data channels from the tracking devices operating via the system. VRPN remains a common platform amongst researchers (Slater et al., 2010; Murray et al., 2007).

Another area of data collection in VR systems is logging of the system behavior. Whilst any VR system can output log messages, a few have developed functionality sufficient to record the whole behavior of the system and replay it. The Virtual Mail system was an early example of multimodal logging for purposes of replay, which allowed both voice and gesture to be recorded and later replayed as a form of VR email (Imai, Johnson, Leigh, Pape, & DeFanti, 1999). Recording and replaying action in VEs was subsequently explored in considerable detail by Greenhalgh et al. in the MASSIVE (Greenhalgh, Purbrick, Benford, et al., 2000), who introduced the concept of 'temporal links' within the MASSIVE-3 framework (Greenhalgh, Purbrick, & Snowdon, 2000), which provide a flexible mechanism for replaying past or recent recordings of VEs within other real-time VEs. The authors suggest applications for such replay including the support for post-production tools for VEs, post-exercise debriefing in training simulators, and asynchronous communication such as VR email. In a similar manner the DIVE system, which is a multi-user distributed VR system, (Frcon, Smith, Steed, Stenius, & Sthl, 2001), can log events in such a way that a network session can be replayed. It can do this by having any one client application log data, or a separate logging process can run on the network. This logging facility was used to analyze the performance of the network system under load (Greenhalgh, Bullock, Frécon, Lloyd, & Steed, 2001) and also to identify usability issues in environments (Steed, Vinayagamoorthy, & Brogni, 2005).

Finally, we noted that most interaction analysis is currently done with respect to video and audio recordings of interaction. We would advise anyone doing VR experiments to also video the experiment (with permission) as it can provide useful information about the user's reaction that might otherwise be opaque to the logging. However there are restrictions to the use of video recording as mode of data collection in immersive VR. Firstly, a fundamental feature of many immersive VR systems is head tracking, which couples the perspective of the graphical display to the user's head movements so that the rendered displays match as closely as possible the changes that would be expected in reality when the same motions are made (Cruz-Neira et al., 1993). Hence, a video camera is unable to capture both the user's physical body and the graphical display with sufficient quality due to perspective misalignment. Thus it is difficult to capture a view of the user and the user's view of the environment simultaneously. One workaround is to make multiple video recordings, with some being taken from video feeds to display devices, or observer cameras renderings from the virtual environment.

Capture Architecture

Requirements

In a system in which an individual's multiple behavioral channels are captured from several sources, the central requirement for multimodal data collection is the processing and collation of the multiple streams in a manner which preserves the temporal and synchronous nature of the action. We would contend that the systems described in the previous section do not fulfill all the requirements we outline below.

The first requirement is that the log file should be able record various types of data from the types listed in the previous section (e.g. tracking, physiological sensors, application events, experimenter annotation, etc.) and described in detail in the following case study section. The second requirement is that because ICVEs generally distribute computation over several networked machines, there should be a central process that is accessible to all machines over that network, and responsible for collation of all data streams. There may be multiple logging processes (see below), but there is an important role for a central logging process. The third requirement is that to be conducive to subsequent processing and analysis, the output log file must me human readable, with data elements appearing in consistent tab- or comma-separated columns. This is so that log files are readable in a broad range of applications from Microsoft Excel, through to custom parsers. The fourth requirement is that the logging system should, as far as possible, record sensor data in a way that makes for efficient for later analysis. As we discuss below, this includes logging sufficient for needs, but also deriving otherwise difficult to generate data from within the system, to provide a rich description of action. An example is complementing eye gaze direction, as recorded from an eye tracker and usually expressed using 2D coordinates, with the concurrent gaze target within the VE, such as another avatar's hand or virtual object. The fifth requirement is that time be stored in a robust way so that subsequent analysis can rely on the temporal nature of events. Further, we may want an external time source to synchronize with other data sources. A subset of these requirements were discussed in (Friedman et al., 2006), but a reference implementation as detailed below was not presented.

Reference Implementation

This section presents a method for multimodal data capture, taking EyeCVE (Wolff et al., 2008) as the case ICVE system. In EyeCVE, a user's performance is captured by several tracking devices, each monitoring a separate channel of nonverbal or verbal behavior. These data streams include eye tracking to measure the oculesic behaviors of gaze, blinks and pupil size; head and hand tracking measuring gross body movement including head direction, arm gesture, and interaction with objects; and audio input detecting verbal signals. Additionally, it is possible for an experimenter to input arbitrary markup data, critical in an experimental scenario. Hence, the central challenge of recording such synchronous multimodal data is the collation of streams in a manner which preserves the holistic temporal characteristics of the original action.

Our approach directs all data streams to a logging process located on the machine operating an eye tracking process, collating the data in a single log file at an update rate of 16 ms (60 Hz). While the network architecture potentially allows for any process to collate and record the multimodal data, there are computational performance reasons why, in this case, we designate the process to the machine running the eye tracker. This decision should be informed by the characteristics of the tracking devices in use. In this case, while the cycle-rate of eye tracking data acquisition (60 Hz) may be less than some other tracking processes, the natural speed and frequency of human oculesics ensures that the amount of discreet and consequential data that is recorded by eye tracking exceeds that produced by all other tracking sources combined. Hence, the location of the logging process should seek to minimize network bandwidth, so it is sensible to locate the logging process on the eye tracker machine. A further motivation is that oculesic data is also more sensitive to temporal misalignment than other less rapid channels of human movement.

Also it is important to note that this log file from this logging process is not the only source of logging in the system. We also make audio recordings and video recordings. These assist in later analysis and it is important that they be synchronized together by some sort of virtual clapperboard, especially in the case when independent recorders are used that don't take time input, such as consumer camcorders. Further, as we discuss later EyeCVE can make a system log of events so that the 3D VR simulation can be replayed.

∗**Figure 1 around here**∗

Figure 1 presents an overview of the approach, and will be the focus of this explanatory section. The diagram features six interrelated columns, each representing a stage of the information processing pipeline. This pipeline progresses from left to right, and accordingly, from initial communication performed by a human user, to the final collated output. Each stage of the pipeline will be discussed in turn.

Whether engaged in collocated or in remote interaction mediated by a telecommunication system, human action is always expressed through natural verbal and nonverbal channels. Telecommunication in EyeCVE is both visual and aural, so data collection must reflect these synchronous modes. The first column, headed *Human* in Figure 1, illustrates the particular cues tracked by EyeCVE. These include the oculesic behaviors of gaze, blinks, and pupil dilation, head and hand movement, and vocal signals. Additionally, in an experimental scenario, it is often critical for an experimenter to be able to input markup data alongside these tracking streams in order to delineate, for example, experimental stages and timing, or add miscellaneous notes. Figure 1 indicates this divide between data originating from an EyeCVE user, and markup data input by an experimenter.

The second column illustrated in Figure 1 is concerned with *Mediation*, which is the initial capture and conversion of natural human behavior to electronic signals ready for processing. Head-mounted eye tracking is used to monitor oculesic behavior including gaze, blinks, and pupil size. Head and hand movement is inferred from tracking sensors attached to a user's head and hand respectively. In immersive VR, head tracking has the primary function of updating the graphical rendering perspective, and hand tracking is generally used as an input device. While EyeCVE does employ the devices for that purpose, the data is also recorded, measuring head and hand movement, posture, and interaction input during system use. Alongside these nonverbal elements, a wearable wireless microphone monitors the verbal component of the social and collaborative interactions supported by the system. Finally, input of additional markup data by an experimenter is performed using a standard keyboard.

When discussing the third column in Figure 1, entitled *Processing*, it should be noted that each of the illustrated software processes may be distributed over various machines that are connected via the same local network. This design is important in order to distribute the significant computation that arises from the synchronous processing of the mediated tracking streams, which may otherwise result in performance bottlenecks. In the approach described here, the eye tracker machine is critical both to processing oculesic data, and system-wide data collation and logging. This is indicated by the direct (i.e. non-networked) link from Processing to Collation stages. Calibrated for each individual prior to system use, the eye tracker (from Arrington Research) outputs 2D gaze direction, blink signals, and pupil size at a rate of 60 Hz. Head and hand tracking was performed using VRCO trackd coupled with InterSense IS-900 sensor devices, outputting six degrees-of-freedom position and rotation values at a rate of 96 Hz. Voice detection is implemented using OpenAL. The original aural component of the interaction is recorded externally to the multimodal log file, and features embedded time stamping consistent with the log file. A binary value indicating the presence of speech is recorded in the log file. The audio detection software has a cycle rate of 100 Hz, and uses a volume-based threshold to filter vocal signals likely to have been emitted from the monitored individual from those likely to originate from other sources. Finally, a keystroke detector operating on a machine accessible to the experimenter is used to input additional markup data. Individual key presses may be assigned to specific text strings, which are transmitted to the data logger at a rate of 100 Hz. For example, at the start of an experiment, the experimenter may input '*s*', which is processed by the keystroke detector, and output in the log file as "*START*".

The *Simulation* stage of the pipeline, the fourth column in Figure 1, is the concern of the ICVE system itself. As the diagram illustrates, six user-tracked signals (gaze, blinks, pupil size, head movement, hand movement, and voice) are transferred over a network, and 'simulated' by EyeCVE, which transposes the actions to have meaning within the context of the VE. These simulated versions of the data are represented by mirrored green icons. The primary use of the simulated data streams is to animate the user's avatar in real-time: gaze is simulated and replicated by the avatar's eye movement; blink signals instigate a model of blinking; changes in pupil size are approximated by the avatar's pupils; head movement is replicated by the avatar, and is also used to update body orientation and posture; hand and arm movement is animated using an inverse kinematics algorithm; and vocal signals cause the avatar's mouth to open and close. However, the 3D spatial and synthetic environment in which interaction takes place in ICVEs means that, in order to preserve the semantic context in which immersed users perform actions, certain modes of tracking data must be processed before logging takes place. When determining whether an input stream requires simulation to provide this context before logging, the answer depends on the dimensionality of the signal. Data featuring more than one dimension must generally be simulated prior to logging, while data that can be represented on a single dimension generally does not require simulation. Consider, for example, pupil size and gaze. Figure 1 indicates that, while both pupil size and gaze are simulated (for avatar animation), only the simulated gaze data is output to the data logger, while the simulated pupil size is not. The reason for this lies in the dimensionality of the signals. Pupil size may be represented by a single scalar value: for instance as a percentage, with 0% representing full constriction and 100% indicating full dilation. Similarly to blinks and vocal input, which can be further simplified by binary values (i.e. blinking or not blinking, and speaking or not speaking), this scalar value has sufficient intrinsic meaning that is present at the processing stage of the data collection pipeline, which occurs prior to the simulation stage. Hence, pupil size requires neither the context of, or calibration in, the 3D spatial environment of EyeCVE. The same may partially be said for gaze. It is correct to state that, to some degree, gaze does have intrinsic meaning at the processing stage of the pipeline, defining a user's 2D direction of gaze relative to the viewpoint of the eye tracker's scene camera. However, while useful for general gaze analysis, these 2D coordinates do not relate explicitly to the immediate visual stimuli being looked at in the 3D VE. Hence, in the case of gaze, the simulation phase consists of casting an intersection ray from the position and orientation of a user's tracked eye, into the VE. The ray returns the hit-point of the virtual geometry (for instance another user's avatar or an object) that is currently being looked at, and returns the object name, along with its position in the VE. Analytically, this data is clearly more meaningful than 2D gaze coordinates, which, as the user is able to physically move in the VR system, are recorded from a varying perspectives. Similar simulation is performed with head tracking data, which returns the hit-object from the position and orientation of a user's head, which is useful as an indicator of attention. Hand tracking simulation records the names of the virtual objects and avatars with which a user interacts. Thus, through simulation in EyeCVE, multidimensional tracking data may be recorded and semantically enriched beyond the prior stage's processed tracking data, thereby enhancing potential for analysis. In contrast, scalar data streams for which no simulation is required, including additional markup data, may be sent directly to the logging process.

The penultimate stage of the data collection pipeline concerns *Collation*, and is represented in the fifth column of Figure 1. This stage is responsible for preparing the individual data streams prior to final output. As the diagram indicates from the unbroken connection from the *Eye Tracker Software* at the Processing stage to the *Logger* at the Collation stage, this computation is performed by the machine local to the eye tracker. Other data streams originate from other processes operating on different machines, and are streamed over the local network to the logger process. The output string, including gaze (both raw and simulated), blink signals, pupil size, simulated head and hand position and rotation, vocal signals, and experimental markup data are is prepared for output in the desired format.

Finally, as indicated in the sixth column of Figure 1, the collated data is *Output* to a single line of a tab-separated log file. In order to preserve the highly temporal and interrelated nature of human behavior, the logger process writes a line of collated data a log file every 16 ms. It should be acknowledged that the varying cycle rates and data transfer throughout the pipeline are likely to result in some temporal variability between the different modes of input as measured from the initial stage of human action to the final stage of log file output. However, this delay is in the order of milliseconds, and is both unavoidable, and in practice, negligible to the intended application of interaction analysis. From experience of logging all cues illustrated in Figure 1 including time stamping at 60 Hz, the logging process writes around 8000 bytes/second, or 0.48 MB/minute. In summary, the data collection pipeline has been designed to operate alongside EyeCVE's distributed system architecture, strategically locating computation to enable collation and output of rich multimodal data describing a user's behavior when engaged in AMC.

Discussion. We made several architectural decisions in defining this reference architecture. Our main requirements were to make a single log file that aggregated diverse data in such a way to enable analysis of effects that might depend on more than one channel of information. The log files produced by the reference architecture do not record all the data produced by the systems involved in the ICVE system. However, this does not mean that other log files couldn't be written and synchronized. In the example analyses in the next section we use three additional sources of information: audio and video recordings and a scene-graph log from EyeCVE that allows the VR simulation to be replayed.

We have focused on a single user in the multi-user system. We could instance a similar logging architecture at multiple sites for different users, or we could unify all logging across the whole system. Our choice would be the former: we are logging data intensive processes, and although we are only logging at 60Hz, some effects, such as glances and gestures are sufficiently fast that the additional latency of logging over the a wide-area network rather than a local-area network would perhaps confuse the issue. On a local-area network, the overhead of communication should be very low (< 1ms), and very reliable. Also, this means that the logfile records what the local user experiences as closely as possible, without introducing latency or ambiguity due to network effects.

Other issues concern the update rate of sensors and time synchronization. The logging rate is lower than some devices report, in particular motion tracking systems. If this was a concern, and the speed of movement was an issue, two alternatives present themselves: run the centralized logging at a higher rate, or make a specific log file for that data. Under the latter configuration a robust time signal synchronization such as NTP or a GPS clock would be needed. A central log file is thus very useful for maintaining a global synchronization: it can be used to recover synchronization points through matching lines of data in the central log file against the external file. For example, this can be used to detect time drift, where a device's clock drifts noticeably over the course of a session.

Multimodal Interaction Analysis

Two sequences of AMC, performed in EyeCVE and captured using the multimodal data collection method detailed in the previous section are now presented. The first analysis presents a dyadic conversational scenario, and features data extracted from an experiment involving truthful and deceptive dyadic conversation (Steptoe et al., 2010). The second analysis examines an object-focused task featuring three users (Steptoe et al., 2009). In each extract, all interactants engaged in the AMC were located in CAVE™-like VR systems with the same tracking modalities described by the reference architecture. The two examples are typical to the type of collaborative interaction performed in ICVE systems. The analyses are novel, both in terms of method and content, and do not appear in the original publications. The segments are analyzed similarly to traditional interaction analysis: by identifying significant moments of interaction, looking for causal relationships between observed behavior, and, through verifiable observation, making judgments of how the interaction may have occurred (Jordan & Henderson, 1995). To simplify the extracts for presentation, each analysis only considers metrics that are insightful with regards to describing the unfolding interaction. These metrics vary accordingly between the conversational and object-focused cases.

Conversational Scenario

Table 1 shows a sequence of interaction drawn an experimental study investigating user behavior when engaged in truthful and deceptive conversation in AMC (Steptoe et al., 2010). Figure 2 shows two experimental interactants and their avatar representations. During the experimental interactions, a confederate issued questions to users, who responded either with truths or with lies. Six lines, manually chosen and extracted from a recorded log file, are presented in Table 1. Each line pertains to a significant moment during the short interaction. The leftmost column states each line's time of occurrence. For ease of reading the first line has been reset to 0. The extract finishes 7.33 seconds later with the sixth line. The remaining columns refer to: markup data (input by the experimenter as the action unfolds); head direction (the object returned by the head-centric intersection ray cast from the position of the user's head); gaze direction (the object in the VE at which the user is looking); pupil size (ranging between 0 indicating full constriction, and 1 indicating full dilation); whether the user is currently blinking (1 indicates blink, 0 indicates no blink); and whether the user is currently talking (1 indicates that they are talking, 0 indicates that they silent).

∗**Figure 2 around here**∗ ∗**Table 1 around here**∗

The sequence begins at *T=0.00*. The "General/Lie/Q01" markup indicates the experimental stage, in which the user is about to be asked the first of a series of general questions, to which they must respond to deceptively, by lying. At this time, the user is gazing a little away from their partner (the questioner), but their head direction indicates that general attention is focused on the questioner. The user's pupil size is close to normal given the environmental luminance levels, indicating that they are relaxed, and neither cognitively or emotionally aroused (pupil dilation is an indicator of such affective processing (Partala & Surakka, 2003)). The user is not talking at this time. At *T=1.81*, the markup data indicates that a question has been issued to the user (in this case the question is "What is your first name?"). Both head and gaze direction indicate focus of attention on the questioner's face, and a slight increase in pupil size may indicate psychological arousal. At this time, the user also begins to blink. At *T=2.65*, the user begins to answer the question, directing both gaze and head direction away from the questioner. The user's pupils dilate significantly as the talk begins, suggesting cognitive load. At *T=5.17*, the user finishes delivering the verbal answer, and returns head direction and gaze to the questioner. The user's pupils are now less dilated. Shortly after, at *T=5.65*, the user again averts gaze from the questioner's face, fixating downwards on their right arm. Tellingly, pupil size increases again, suggesting that the prior eye contact following the lie-telling evoked the user's negative emotional arousal due to the social discomfort of lying. Finally, at *T=7.33*, the markup data indicates that the current question is over, and the next is soon to be issued. The user's pupil size has almost returned to normal, indicating a more relaxed state. A blink occurs, and gaze indicates that the user's attention is again focused on the questioner.

Object-Focused Scenario

Table 2 shows a sequence of three-party object-focused interaction drawn from an experimental study investigating immersive AMC (Steptoe et al., 2009). The task revolves around constructing a simplified Rubik's cube from eight smaller cubes, so that each face of the finished cube consists of a single color. In the original experiment, the collaborative interactions involved three people, represented as avatars in their partners' VR systems, working together to find specifically-colored cubes, picking up those cubes in turn, and positioning each cube in its correct position as part of the larger cube. Seven lines, manually chosen and extracted from a log file are presented, each one pertaining to a significant stage during the interaction. Each line's time of occurrence is indicated in the leftmost column, which is reset to 0 at the first line, and finishes 9.32 seconds later at the seventh. Figure 3 shows eye gaze direction at three moments during the analyzed excerpt from the perspective of the participant. The key difference between this object-focused segment and the conversational segment is the analysis of hand tracking data, which identifies any currently-grabbed object, while pupil size and blink data are omitted.

∗**Figure 2 around here**∗ ∗**Table 1 around here**∗

At *T=0.00*, the user is searching for a particular cube, and appears to be examining "Cube-04" (left in Figure 3). Soon after, at *T=2.17*, the user has grabbed "Cube-04". The user's head is oriented toward the grabbed object, but gaze is directed at their partner's body. Voice data indicates that the user is speaking. In the context of the experiment, this is likely to be querying the correctness of the currently-grabbed cube for the intended placement position in the puzzle. This initial choice appears to have been incorrect as, at *T=4.81*, the user has released "Cube-04", and is again visually searching the VE for the correct cube. At *T=5.65*, the user's gaze falls on "Cube-05", and this time, caution is exerted before grabbing as, at *T=6.17*, gaze is directed to their partner's face, and another vocal query is uttered (middle in Figure 3). At *T=7.48*, following what appears to have been an affirmative response, "Cube-05" is grabbed, and subsequently positioned, at *T=9.32* where the segment ends (right in Figure 3).

Discussion

The two example analyses aim to demonstrate the ability of multimodal data capture to preserve detailed information describing action performed during AMC in ICVEs. During analysis of the conversation segment, the user's state of psychological arousal and cognitive load was inferred from behavioral data, in particular from changes in pupil size. However, independent analysis of such characteristic changes in pupil size is unable to reveal the broader interactional circumstances under which they occurred. Hence, analysis of concomitant changes in gaze and head direction, together with talk signals were essential in order to describe the overall sequence. In this case, the act of establishing mutual gaze (*T=5.17* in Table 1), and the subsequent response of increased pupil dilation (*T=5.65* in Table 1) to the uncomfortable social situation of lying to another person, is observable in the data. More generally, the collation of multiple cues is able to reveal the occurrence of social practices common to collocated conversation. For instance a common gaze practice is for a listener to look directly at a speaker's facial region, particularly at the eyes, while being questioned. Once the question has been issued, response formation and delivery begins with the respondent looking away from the questioner, indicating cognition, and signaling a hold of the conversational floor. The response ends by redirecting gaze back to the questioner as the answer reaches completion, returning control of the floor to the questioner (Duncan, 1974). The three stages of this behavioral sequence may be observed accordingly at *T=1.81*, *T=2.65*, and *T=5.17* in Table 1.

Analysis of the object-focused segment centered on a user's physical action of searching, grabbing, and positioning cubes when solving the puzzle. Hence, hand and body movement in combination with both verbal and gaze behavior were used to describe the logistic process of the task. In this case, the user demonstrates a behavioral learning effect in response to a previous mistake. Early in the segment (*T=2.17* in Table 2), the user is seen to grab an incorrect cube without first asking their partners' approval. Significantly, the subsequent grab attempt (*T=7.48*) only occurs following a verbal query (*T=6.17*) seeking confirmation that the cube being pointed to is correct. Throughout the sequence, the combination of head and gaze direction, talk, and particularly hand action are able to elucidate the captured collaboration.

In each analysis, the benefit of being able to reference synchronized elements of nonverbal and verbal telecommunication is evident when attempting to explicate sequences of interaction.

Additionally, the multimodal log file grants the opportunity to calculate extra columns with derived data that are targeted at analysis. For instance the distance between participants, or whether the participants are engaged in mutual gaze may easily be calculated. Such derived information has potential to enrich a given analysis so that a fuller representation of the interaction and it's context may be reconstructed. It must be noted that the logging process writes 60 lines per second to an output file. Hence, in order to distill an interaction as presented in the two examples, a significant amount of both automated and manual processing must be performed prior to actual analysis. While the logical format of the log file enables automated parsing, it is often necessary to reference an audio or visual replay of the performance in order to be certain of the context in which the activity took place. To this end, two solutions are available in EyeCVE. Firstly, the aural component of interaction is recorded in the Ogg Vorbis (Moffitt, 2001) bitstream format, which supports multiplexing of a number of separate codecs including audio and text. Alongside this multiparty talk, a textual time stamp, matching that of the main collated log file, is embedded. Hence, audio recording and log files are synchronized, and allow for random access. Secondly, a log file player application has been developed and is presented in (Murgia et al., 2008), which is able to reconstruct and replay the recorded virtual action together with audio . The player application allows a session to be replayed, paused, and randomly accessed, and is also capable of visualizing additional data, including gaze targets as the interaction proceeds. Additionally, the player application operates within the immersive CAVE™-like systems used during the original experimental AMC. This allows for observation of, and movement within, the same spatial VE in which the original interaction occurred. Hence, an analyst is able to scrutinize the pre-recorded interaction similarly to how they may do so a video replay, but with the critical advantage of navigating with an adjustable camera viewpoint. Such virtual replay has been useful in performing usability analysis through identification of aspects of the VE that might need attention and refinement (Steed et al., 2005).

For simplicity, the example analyses explored the behavior of a single user engaged in social interaction supported by immersive AMC. Clearly, this only part of the complete interaction, which involved either one or two other remote users. The proposed approach to data collection captures performance at a single site. However, log files capturing activity from multiple users are easily generated. Our favored approach to multi-user capture is to log locally at each site as described by the capture architecture, firstly ensuring that clocks are synchronized between machines at all sites using a network time protocol such as IEEE 1588 or an external radio clock. The separate log files generated at each site may then be merged based on time stamp to generate log files relating to multiple users engaged in AMC. This approach is preferable to the alternative of all sites transmitting significant amounts of data across a wide-area network to be collated and logged by a central process, which would introduce additional and variable latency.

Conclusions

The diversity of research investigating both single- and multi-user VEs ensures that a variety of data collection and analysis methods are employed. Researchers identify these methods depending on the aims of each experimental study. We suggest that, even if multimodal analysis is not initially intended, multimodal data capture of as many time-synchronized tracking sources and events supported by a VE system is a low-cost and high-impact investment. The approach documented in this paper is able to capture precise multimodal data describing users' nonverbal and verbal behavior in a manner that preserves temporal characteristics and emphasizes the causal interrelationships between synchronous tracking streams. As demonstrated by the analyses, this data provides a rich description of the original mediated social interaction that occurred, leading to a broad understanding of how and why action may have unfolded, including users' intent and state of psychological arousal.

All forms of data collection and analysis of human behavior must acknowledge the difference between the reality of original events, and the transformation of that reality that takes place through data capture. This transformation of natural human interaction is further altered due to the mediated nature of ICVEs. Transformations are always less rich than the original events they represent, and involve some loss of information. When studying AMC in ICVE systems, in which users generally act in a socially-realistic manner, we consider multimodal collection and analysis to provide an excellent foundation for analytic knowledge that is superior to examining cues individually. More generally, multimodal analysis is also likely to be a fruitful means of identifying usability issues relating to VE systems. Whether arising from technical or human-centered factors, moments of problematic or non-optimal action are likely to be revealed through such analyses.

Future work includes automated processing of log files to reduce manual analysis effort. This is a broad research area that should likely employ the use of existing models of human behavior and interaction such as proxemics (Hall, 1968), and gaze distribution (Oyekoya, Steptoe, & Steed, 2009). Automatic identification of significant interactional events, such as when mutual eye contact is established between two interactants, or identification of common gestures such as head-nodding and pointing would further enrich the recorded data. Considering the wider VE research community, standardization of logging format is desirable, and possible due to the data types having canonical representations.

References

Argyle, M., & Trower, P. (1979). *Person to Person: Ways of Communicating*. HarperCollins Publishers.

Badler, N., Hollick, M., & Granieri, J. (1993). Real-time control of a virtual human using minimal sensors. *Presence*, *2*(1), 82–86.

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

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

Bideau, B., Multon, F., Kulpa, R., Fradet, L., Arnaldi, B., & Delamarche, P. (2004). Using virtual reality to analyze links between handball thrower kinematics and goalkeepers reactions. *Neuroscience Letters*, *372*(1-2), 119–122. Available from `http://www.ncbi.nlm.nih.gov/pubmed/15531100`

Bowman, D., & Hodges, L. (1997). An evaluation of techniques for grabbing and manipulating remote objects in immersive virtual environments. In *Proceedings of the 1997 symposium on interactive 3d graphics.*

Cruz-Neira, C., Sandin, D., & DeFanti, T. (1993). Surround-screen projection-based virtual reality: the design and implementation of the CAVE. In *Proceedings of the 20th annual conference on computer graphics and interactive techniques* (p. 142).

Duncan, S. (1974). Some signals and rules for taking speaking turns in conversations. *Nonverbal Communication: Readings with Commentary.*

Frcon, E., Smith, G., Steed, A., Stenius, M., & Sthl, O. (2001). An overview of the coven platform. *Presence: Teleoperators and Virtual Environments*, 109-127.

Freeman, J., Avons, S., Pearson, D., & IJsselsteijn, W. (1999). Effects of sensory information and prior experience on direct subjective ratings of presence. *Presence: Teleoperators & Virtual Environments*, *8*(1), 1–13.

Friedman, D., Brogni, A., Guger, C., Antley, A., Steed, A., & Slater, M. (2006). Sharing and analyzing data from presence experiments. *Presence: Teleoperators and Virtual Environments*, *15*(5), 599–610.

Frisoli, A., Ruffaldi, E., Filippeschi, A., Avizzano, C., Vanni, F., & Bergamasco, M. (2008). In-door skill training in rowing practice with a vr based simulator. In *Proceedings of 10th european workshop of ecological psychology.*

Garau, M., Slater, M., Bee, S., & Sasse, M. (2001). The impact of eye gaze on communication using humanoid avatars. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 309–316).

Goodwin, C., & Heritage, J. (1990). Conversation analysis. *Annual review of anthropology*, *19*(1), 283–307.

Greenhalgh, C., Bullock, A., Frécon, E., Lloyd, D., & Steed, A. (2001, April). Making networked virtual environments work. *Presence: Teleoperators and Virtual Environments*, *10*(2), 142–159. Available from `http://dx.doi.org/10.1162/105474601750216777`

Greenhalgh, C., Purbrick, J., Benford, S., Craven, M., Drozd, A., & Taylor, I. (2000). Temporal links: recording and replaying virtual environments. In *Proceedings of the eighth ACM international conference on multimedia* (pp. 67–74).

Greenhalgh, C., Purbrick, J., & Snowdon, D. (2000). Inside MASSIVE-3: flexible support for data consistency and world structuring. In *Proceedings of the third international conference on collaborative virtual environments* (pp. 119–127).

Hagni, K., Eng, K., Hepp-Reymond, M., Holper, L., Keisker, B., Siekierka, E., et al. (2008). Observing Virtual Arms that You Imagine Are Yours Increases the Galvanic Skin Response to an Unexpected Threat. *PLoS One*, *3*(8).

Hall, E. (1968). Proxemics. *Current Anthropology*, *9*(2/3), 83.

Imai, T., Johnson, A., Leigh, J., Pape, D., & DeFanti, T. (1999). The virtual mail system. *vr*, 78.

Insko, B. (2001). *Passive haptics significantly enhances virtual environments.* Unpublished doctoral dissertation, Citeseer.

Jordan, B., & Henderson, A. (1995). Interaction analysis: Foundations and practice. *The Journal of the learning sciences*, *4*(1), 39–103.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychol (Amst)*, *26*(1), 22–63.

Keysar, B., Barr, D., Balin, J., & Brauner, J. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 32–38.

Lerner, G. (2003). Selecting next speaker: The context-sensitive operation of a context-free organization. *Language in Society*, *32*(02), 177–201.

Meehan, M., Insko, B., Whitton, M., & Brooks Jr, F. (2002). Physiological measures of presence in stressful virtual environments. In *Proceedings of the 29th annual conference on computer graphics and interactive techniques* (pp. 645–652).

Moffitt, J. (2001). Ogg VorbisOpen, Free AudioSet Your Media Free. *Linux journal*, *2001*(81es), 9.

Murgia, A., Wolff, R., Steptoe, W., Sharkey, P., Roberts, D., Guimaraes, E., et al. (2008). A tool for replay and analysis of gaze-enhanced multiparty sessions captured in immersive collaborative environments. In *Proceedings of the 2008 12th IEEE/ACM international symposium on distributed simulation and real-time applications-volume 00* (pp. 252–258).

Murray, N., Roberts, D., Steed, A., Sharkey, P., Dickerson, P., & Rae, J. (2007). An assessment of eye-gaze potential within immersive virtual environments. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, *3*(4), 8.

Oyekoya, O., Steptoe, W., & Steed, A. (2009). A saliency-based method of simulating visual attention in virtual scenes. In *Proceedings of the 16th ACM symposium on virtual reality* *software and technology* (pp. 199–206).

Pan, X., & Slater, M. (2007). A preliminary study of shy males interacting with a virtual female. In *Presence: The 10th annual international workshop on presence.*

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, *59*(1-2), 185–198.

Pavlik, J., & McIntosh, S. (2004). *Converging media: An introduction to mass communication*. Allyn & Bacon.

Pertaub, D., Slater, M., & Barker, C. (2002). An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators & Virtual Environments*, *11*(1), 68–78.

Roberts, D., Wolff, R., Otto, O., & Steed, A. (2003). Constructing a Gazebo: Supporting Teamwork in a Tightly Coupled, Distributed Task in Virtual Reality. *Presence: Teleoperators & Virtual Environments*, *12*(6), 644–657.

Rovira, A., Swapp, D., Spanlang, B., & Slater, M. (2009). The Use of Virtual Reality in the Study of People's Responses to Violent Incidents.

Sacks, H. (1995). *Lectures on Conversation*. Blackwell.

Sanchez-Vives, M., & Slater, M. (2005). From presence to consciousness through virtual reality. *Nature Reviews Neuroscience*, *6*(4), 332–339.

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

Shockley, K., Baker, A., Richardson, M., & Fowler, C. (2007). Articulatory constraints on interpersonal postural coordination. *Journal of Experimental Psychology*, *33*(1), 201–208.

Slater, M. (2004). How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *13*(4), 484–493.

Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., et al. (2006). A virtual reprise of the Stanley Milgram obedience experiments. *PLoS One*, *1*(1).

Slater, M., Spanlang, B., & Corominas, D. (2010). Simulating virtual environments within virtual environments as the basis for a psychophysics of presence. *ACM Transactions on Graphics (TOG)*, *29*(4), 92.

Slater, M., & Usoh, M. (1993). Presence in immersive virtual environments. In *1993 IEEE virtual reality annual international symposium, 1993.* (pp. 90–96).

Slater, M., Usoh, M., & Steed, A. (1994). Depth of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, *3*(2), 130–144.

Slater, M., & Wilbur, S. (1997). A Framework for Immersive Virtual Environments(FIVE)-Speculations on the role of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, *6*(6), 603–616.

Steed, A., Vinayagamoorthy, V., & Brogni, A. (2005). Using breaks in presence to identify usability issues. In *Proceedings of the 11th international conference on human computer interaction* (pp. 22–27).

Steptoe, W., Oyekoya, O., Murgia, A., Wolff, R., Rae, J., Guimaraes, E., et al. (2009). Eye Tracking for Avatar Eye Gaze Control During Object-Focused Multiparty Interaction in Immersive Collaborative Virtual Environments. In *Proceedings of the 2009 IEEE virtual reality conference* (pp. 83–90).

Steptoe, W., Steed, A., Rovira, A., & Rae, J. (2010). Lie tracking: social presence, truth and deception in avatar-mediated telecommunication. In *Proceedings of the 28th international conference on human factors in computing systems* (pp. 1039–1048).

Steptoe, W., Wolff, R., Murgia, A., Guimaraes, E., Rae, J., Sharkey, P., et al. (2008). 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* (pp. 197–200).

Taylor, I., Russell, M., Hudson, T., Seeger, A., Weber, H., Juliano, J., et al. (2001). VRPN: a device-independent, network-transparent VR peripheral system. In *Proceedings of the ACM symposium on virtual reality software and technology* (p. 61).

Usoh, M., Arthur, K., Whitton, M., Bastos, R., Steed, A., Slater, M., et al. (1999). Walking > walking-in-place > flying, in virtual environments. In *Proceedings of the 26th annual conference on computer graphics and interactive techniques* (pp. 359–364).

Usoh, M., Catena, E., Arman, S., & Slater, M. (2000). Using presence questionnaires in reality. *Presence: Teleoperators & Virtual Environments*, *9*(5), 497–503.

Vertegaal, R., Slagter, R., Veer, G. van der, & Nijholt, A. (2001). Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 301–308).

Walther, J., Parks, M., Knapp, M., & Daly, J. (2002). *Handbook of interpersonal communication.* Sage Thousand Oaks, CA.

Wolff, R., Roberts, D., Murgia, A., Murray, N., Rae, J., Steptoe, W., et al. (2008). Communicating eye gaze across a distance without rooting participants to the spot. In *Proceedings of the 2008 12th IEEE/ACM international symposium on distributed simulation and real-time applications-volume 00* (pp. 111–118).

Wolff, R., Roberts, D., & Otto, O. (2004). A study of event traffic during the shared manipulation of objects within a collaborative virtual environment. *Presence: Teleoperators & Virtual Environments*, *13*(3), 251–262.

Yee, N., Bailenson, J., Urbanek, M., Chang, F., & Merget, D. (2007). The Unbearable Likeness of Being Digital: The Persistence of Nonverbal Social Norms in Online Virtual Environments. *CyberPsychology & Behavior*, *10*(1), 115–121.

Author Note

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

Table 1

Selected data from an interaction sequence taken from the truth and deception experiment documented in Chapter 6. Particular lines taken from the log file have been chosen to highlight significant moments during the interaction. In this sequence, the user is required to lie to questions issued by a partner. In this case, they are asked "What is your first name?"

Time	*Experimenter's Markup*	*Head Direction*	*Gaze Direction*	*Pupil Size*	*Is Blinking*	*Is Talking*
0.00	General Questions/Lying/Question 01	Partner-Body	Grid-E2	0.08	0	0
1.81	Question issued	Partner-Face	Partner-Face	0.27	1	0
2.65	Answer start	Grid-D4	Grid-B4	0.39	0	1
5.17	Answer end	Partner-Body	Partner-Face	0.20	0	1
5.65	-	Partner-Body	Partner-Arm-R	0.31	0	0
7.33	General Questions/Lying/Question 02	Partner-Face	Partner-Face	0.13	1	0

Table 2

Selected data from an interaction sequence taken from the object-focused experiment documented in Chapter 5. In this sequence, the tracked user is solving a simplified Rubik's Cube puzzle with two partners.

Time	*Experimenter's Markup*	*Head Direction*	*Gaze Direction*	*Hand Position*	*Is Talking*
0.00	Search	Wall-01	Cube-04	-	0
2.17	Grab and Query	Cube-04	Partner-Body	Cube-04	1
4.81	Search	Wall-03	Cube-02	-	0
5.65	Search	Cube-02	Cube-05	-	0
6.17	Query	Cube-05	Partner-Face	-	1
7.48	Grab	Cube-05	Cube-05	Cube-05	0
9.32	Position	Cube-05	Cube-05	-	1

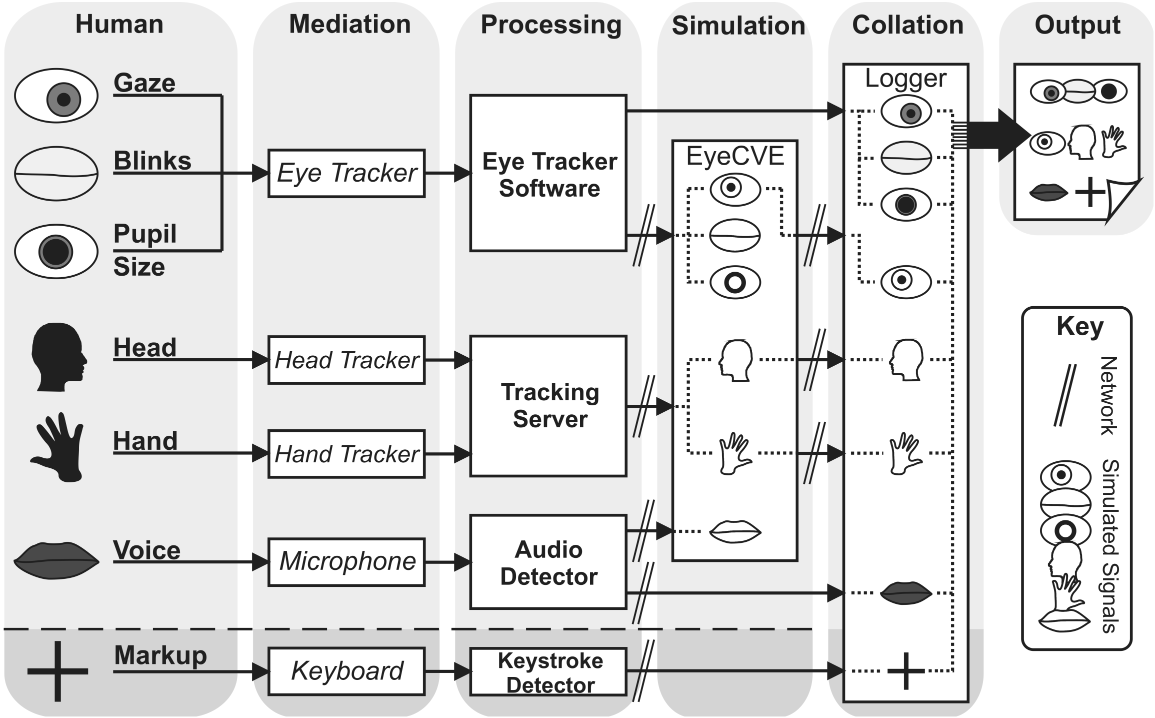
Figure Captions

Figure 1. ∗TWO COLUMN WIDTH∗Multimodal data collection pipeline. The six columns describe the flow of information, from left to right. Initial expression of *Human* behavior is captured by *Mediation* devices. These signals are synthesized and *Processed* by software running on machines local to each device. Input channels relevant to EyeCVE's real-time avatar animation system are passed to EyeCVE for *Simulation* and display at remote sites. The specific simulated data streams (green icons) of gaze, head, and hand movement are then transferred across a network to the logging system running on the machine local to the eye tracker. This data is *Collated*, along with raw gaze coordinates, blinks, pupil size, verbal signals, and additional markup data. These raw data streams all have inherent meaning outside of the 3D context of EyeCVE's VE, so do not require simulation. Finally, all data is *Output* in human-readable form, together with timing information, to a single line of a log file at 60 Hz (16 ms).

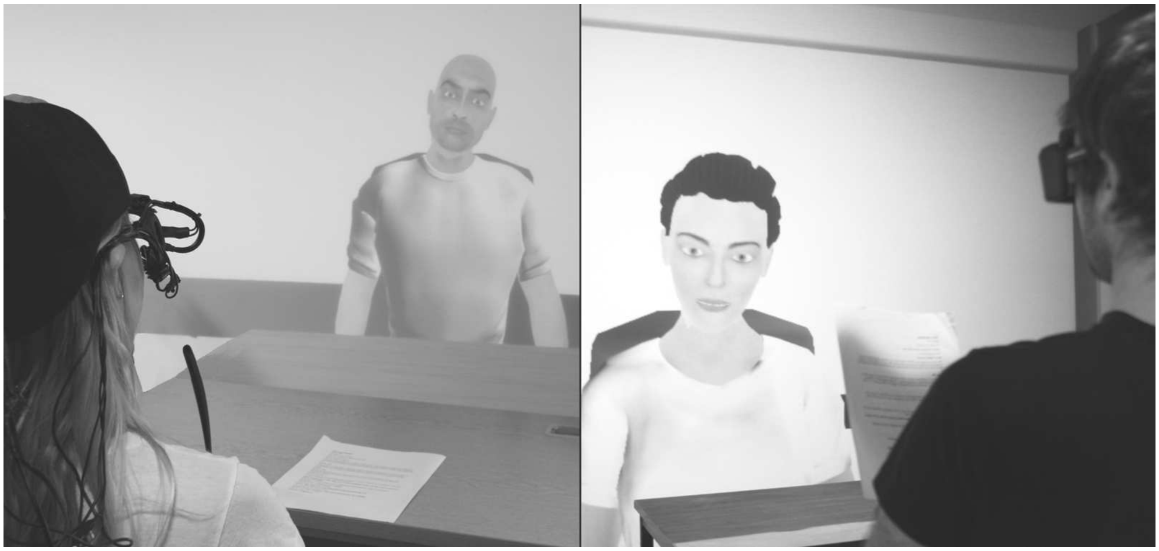
Figure 2. ∗ONE COLUMN WIDTH∗Users engaged in immersive AMC involving truthful and deceptive conversation, presented in (Steptoe et al., 2010). Both users viewed their partner's avatar embodiment in a VE depicting a meeting room. Avatars were animated in real-time using the same tracking devices used to record data eye tracking, head tracking, hand tracking, and an audio microphone.

Figure 3. ∗TWO COLUMN WIDTH∗A sequence of interaction recorded from the perspective of a user engaged in triadic object-focused AMC. Relating to Table 2: *Left* T=0.00; *Middle* T=6.17; *Right* T=9.32. Avatars were animated in real-time using the same tracking devices used to record data (eye tracking, head tracking, and hand tracking).

, Figure 1


Human
Mediation
Processing
Simulation
Collation
Output
Gaze
Blinks
Pupil
Size
Head
Hand
Voice
Markup
Eye Tracker
Head Tracker
Hand Tracker
Microphone
Keyboard
Eye Tracker
Software
Tracking
Server
Audio
Detector
Keystroke
Detector
EyeCVE
Logger
Key
Network
Simulated Signals

, Figure 2



, Figure 3