Chapter 11
Collaboration in Immersive and Non-immersive Virtual Environments

Anthony Steed and Ralph Schroeder

Abstract There is a huge variety of tools for synchronous collaboration including instant messaging, audio conferencing, videoconferencing and other shared spaces. One type of tool, collaborative virtual environments (CVEs), allows users to share a 3D space as if they are there together. Today, most experiences of virtual environments (VEs), including games and social spaces, are constrained by the form of non-immersive interfaces that they use. In this chapter we review findings about how people interact in immersive technologies, that is large-screen displays such as CAVE-like displays, and how they provide a number of advantages over non-immersive systems. We argue that modern immersive systems can already support effective co-presence in constrained situations and that we should focus on understanding of what is needed for effective and engaging collaboration in a broader range of applications. We frame this discussion by looking at the topics of co-presence, representations of users and modalities of interacting with the VE. Different types of immersive technologies offer quite distinct advantages, and we discuss the importance of these differences for the future of CVE development.

Keywords Synchronous collaboration • Collaborative Virtual Environments (CVE's) • 3D Space • Interaction • Social space • Immersive technologies • Co-presence

A. Steed (✉)
Department of Computer Science, University College London,
Gower St, London WC1E 6BT, UK
e-mail: A.Steed@ucl.ac.uk

R. Schroeder
Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, UK
e-mail: ralph.schroeder@oii.ox.ac.uk

M. Lombard et al. (eds.), *Immersed in Media*, DOI 10.1007/978-3-319-10190-3_11

11.1 Introduction

What is most people's experience of synchronous collaboration at a distance? The most common experience is voice over the telephone or text messaging. Over the past few years video conferencing and other forms of web-based collaboration tools have become more popular (Hinds and Kiesler 2002). We individually might have preferences for some or other of these tools, but we would all agree that using these tools is nothing like being there together with our collaborators. For example, the problems of maintaining shared references with video-conferencing have been well understood for two decades (Gaver et al. 1993). However such technologies are very convenient since, even if the software might need some configuration, there is little or no per-user configuration required.

An emerging collaboration technology is shared or collaborative virtual environments (SVEs or CVEs). CVEs have developed rapidly over the last decade or so. Apart from applications in a few niche industrial projects and a variety of academic demonstrator projects, the most widespread uses of CVE are shared spaces for socializing and gaming such as Second Life and World of Warcraft. In this chapter, we will focus on collaboration in VEs and the effective and engaging use of these immersive spaces. Whether these will become as widespread as the leisure uses of non-immersive VEs is a question we will leave to one side, although one point to make at the outset is that even if online gaming and socializing continue to lead the uses of SVEs, the requirements of workaday uses of CVEs will need to be tackled if immersive (and indeed non-immersive) systems will be able to deliver on their dual promise of bridging distance between people and allowing them do things in spatial environments together. Therefore, regardless of whether future developments come from the 'pull' of applications, or from the 'push' of more powerful and less expensive immersive systems – one of the arguments that will be made here is that we need a better understanding of the benefits of immersion and of how people are able to interact with each other and with the environment using media.

In this chapter we will cover a range of CVE technologies. The range of technologies can be characterised in two ways: the *spatial extent* that is shared and the *degree of user modelling*. Just as there are different models for audio collaboration (e.g. point to point versus conference call) or text messaging (SMS versus Twitter), a CVE has a model with a particular spatial extent. We distinguish two particular spatial extents: *face-face extent* and *extended extent*. In a face-face extent the CVE simulates the situation of being across the table from a user. Examples include the Spin3D system (Louis Dit Picard et al. 2002) and the Office of the Future system that we will discuss in more detail in a later section (Fig. 11.1) (Raskar et al. 1998). Both these systems simulate a particular situation of a pair or a small group around a table (e.g. Spin3D, see below). Virtual objects can be shared in the common space in front of the users. Unlike videoconferencing, this type of CVE allows proper capture of or simulation of eye-gaze between users and objects in the common space. Extended extent refers to the majority of CVEs where users can independently navigate through complex information, walkthroughs of buildings and landscapes, and manipulate a range of objects.

Fig. 11.1 Two illustrative CVE systems that simulate a face-face situation. *Left*, Spin3D system (Image courtesy of Laboratoire d'Informatique Fondamentale de Lille). *Right*, Office of the Future system (Image courtesy of Department of Computer Science, University of North Carolina at Chapel Hill)

The degree of user modelling is also illustrated by the Spin3D and Office of the Future systems. The former uses a set of pre-modelled avatars. Users interacting with the system indirectly control the avatar, effectively acting as puppeteer. In Office of the Future though, the system completely reconstructs a representation of the user in real-time. In between these two extremes is a spectrum of systems that track some of the movements of the user in order to manipulate an avatar representation. We refer to these three types as *puppeteered*, *reconstructed* and *tracked*.

These two characterisations of CVEs pose many technical challenges and opportunities. It might seem that ideally we would support extended extent and reconstructed avatars, but this is an incredible technical challenge. If we take a step back to either face-face/reconstructed or extended extent/tracked we find that the technical challenges are much more tractable. In any case, we will see that there are already many opportunities and configurations of systems for enhanced communication.

In this chapter we explore the opportunities and challenges in more detail. To this end we go back to what is known about how people interact with each other and with the environment, both for immersive and non-immersive CVE systems. One area that has been investigated extensively is *presence*, or how people experience 'being there' in the environment (see Scheumie et al. 2001 and other chapters in this volume). There have been extensive debates about how to measure presence, but people tend to experience a greater sense of presence in immersive as opposed to desktop systems. *Co-presence*, the 'experience of being with others' is much more difficult to gauge (see also Schroeder 2011). One way to understand co-presence is by looking for situations when it is absent or much reduced – such as when using instant messaging or a phone call. In these situations it can be difficult to keep attention on the conversation and misunderstandings can occur in ways that don't in real conversations.

One reason for raising the topic of co-presence is that so far, co-presence has been studied as a psychological state, by asking the user, or otherwise ascertaining their state of mind at a particular time or for a particular experience. But, from the user's point of view, it is not the psychological 'state' that is important (however measured), but what they experience in terms of being able to interact with the other person and with the environment. In other words, the study of co-presence will need to become much more complex: it is not just that co-presence depends on the 'context', the application or the setting, but that several factors will affect co-presence. It may be, for example, that the spatial experience of the environment and the experience of being there with another person (the spatial versus communication uses of CVEs) will require quite different lines of investigation.

Howsoever this research is undertaken and whatever its findings may be, ultimately the factors affecting co-presence will need to be brought into a single model so that a body of cumulative research can be built up – as it has for presence. Yet the task of studying co-presence is made more difficult and uncertain by the fact that technology development and the uses or applications of the technology are indeterminate. We shall argue later that we can nevertheless foresee what the end-states of immersive CVEs will be, and this mitigates this uncertainty and indeterminacy and will allow us considerable insight into the effectiveness of different systems.

In the rest of this chapter we talk first about technologies for collaboration. We then give some initial observations about the impact that user representations have, and how these are used in collaboration. In Sect. 11.4 we introduce studies of co-presence, and we cover a three-way classification of factors that affect co-presence: modality, realism and context. Next we discuss end-states of collaboration technologies and we claim that CVE technology is actually heading in at least two different directions. We conclude with a short list of challenges for CVE developers.

11.2 Technologies

As mentioned in the introduction, one of the characteristic features of any CVE is the degree of user modelling. In this section we give a more detailed characterisation of the different degrees: puppeteered, tracked and reconstructed.

11.2.1 Puppeteered Avatars

Recently the burgeoning market for online 3D games has pushed this type of avatar into the limelight. Two common genres are first-person shooter (FPS) games and massively-multiplayer online games (MMOGs). The former are well known and described in non-academic writing, examples include the Halo series from Microsoft, the Quake and Doom series from Id Software and the Unreal

Fig. 11.2 Eyes of a high-quality avatar suitable for real-time rendering. Eye blinks, eye gaze and pupil dilation are all modelled as part of the behaviour of the avatar (Courtesy of Will Steptoe, UCL)

series from Epic Games. The latter genre has attracted more attention in academic literature (e.g. contributions by Persky and Blascovich, Jakobsson, Yee 2006; Brown and Bell, and Steen et al. in Schroeder and Axelsson 2006; Williams et al. 2006). Important examples include Everquest from Sony, Second Life from Linden Lab, Lineage II from NCSoft Corporation and World of Warcraft from Blizzard Entertainment.

In both genres of games the user is typically represented in the world by an avatar and the user explores the virtual environment by using that avatar. Figure 11.2 represents an example high-end avatar figure that typifies those in games in 2014. These days these avatars are obviously sophisticated enough that they could represent the gender, identity, role, emotional state and intentions of the user dynamically over time. But crucially these avatars are like puppets: they do not directly represent the actual player, because the appearance of the avatar is constrained by the visual metaphor of the environment and the constraints of the animations built in to the avatars.

Players will go to great length to customise these avatars, even creating representations that look like themselves (Cheng et al. 2002) but still these avatars have to be controlled through an interface.

11.2.2 Tracked Avatars

The most common use of tracked avatars is with immersive systems. In 2014 most high-end immersive systems are using Cave Automatic Virtual Environments (CAVE) -like displays, though there is renewed interest in high field of view head-mounted displays driven by consumer technology. Figure 11.3 shows a 3D model and a view into a four-walled CAVE-like system, in the lab of one of the authors. Such a facility is typical of those in academic labs, though there is increasing usage of these technologies in industrial applications (e.g. Weaver 2010).

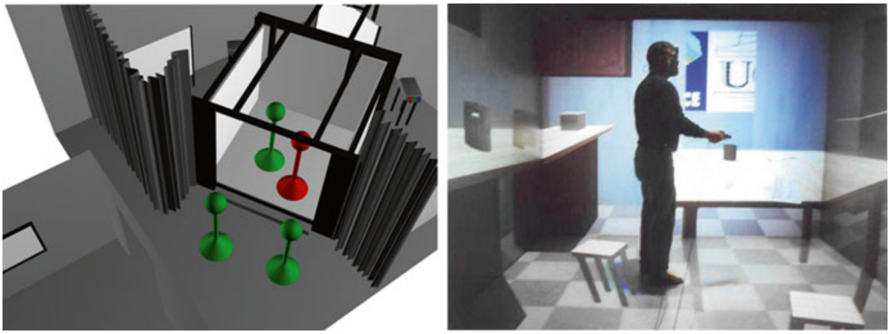

Fig. 11.3 *Left*, a 3D model of UCL's immersive systems representing the four walls and a number of users. *Right*, a view into the system with a user in front of the walls

The key components of this technology are that the images are in stereo on the walls and the head is tracked. This combination provides the ability to create images that show correct parallax when the head moves, creating the illusion of depth in objects. Unlike some other 3D stereo technologies, the limits of parallax are quite high so objects can appear to be distant and proximate to the user, in particular objects can appear to be inside the walls. Because of this property and because of the size of the screens, this technology is highly "immersive" in that it can create imagery that surrounds the user and isolates them from the real world. It provides the capability to represent objects at a one-to-one scale, and in particular people can be represented at a one-to-one scale.

The head needs to be tracked to create the correct imagery on the screen, but a side effect of this is that the user's position is known. Usually between one and three additional tracked points on the person are known, typically at least the dominant hand, and often both hands and the torso. This very limited tracking information allows us to generate a 3D model of the user of the system (e.g. Badler et al. 1993). This tracking can be seen as a limited form of motion capture. Motion capture is a technology most commonly used in the animation industry to create animation sequences for rendering offline (Jung et al. 2000). It typically uses quite a few tracked points all over the body in order to track deformations of all major limbs. Such systems can be integrated into CAVE-like systems, but current technologies are usually limited by the discomfort and inconvenience of "dressing" in sensors or markers before entering the system. Later in this chapter we will come back to experimental evidence from studies of collaborative tasks that show that simple tracked avatars can create a highly expressive representation of another person. For the moment, it suffices to note that the perception literature shows us that we can recognise human motion from very little information. For example, it has been shown that from a few moving point lights on the wrists and ankles we can tell not only gender of a subject, but aspects of their mood (Pollick et al. 2002). This suggests, and our later review will provide more evidence, that limited motion capture conveys a lot of the important information about a user's behaviour and state.

11.2.3 Reconstructed Avatars

Motion capture provides information about user motion, but can't provide real-time information about appearance. We will need to capture full 3D models in real-time order to satisfy our requirements of being able to place the user inside the virtual model. Currently detailed 3D models can only be captured offline, and whilst the resulting model is animated, this is tricky to do accurately. Of course appearance can change quickly and such animated models might not capture the subtlety of face expression, eye-gaze and so on.

What we would like is systems that can capture the 3D model of the user's appearance as well as movement in real-time. This has been a goal of computer vision for decades, and recently we have started to see the integration of these techniques into immersive virtual environments. We will briefly discuss two systems: the Office of the Future project and the Blue-C system.

The Office of the Future project (see Fig. 11.1, *right*) integrated real-time 3D model capture with head-tracked video display (Raskar et al. 1998). A number of demonstrations have been done, the key theme of the research being real-time reconstruction of the user in front of the screen. To date only one-way systems have been built; that is, one user is reconstructed and presented remotely to another user, but it is expected that advances in capture and processing equipment will make this easier. Figure 11.1, right showed an example of a real-time reconstruction. The background is statically captured, and the user is updated at interactive rates. The view of the remote user is somewhat blocky. This is a facet of the underlying algorithms which creates a "voxel" representation of the user – effectively a reconstruction out of small virtual cubes. The technology works by using an array of cameras around the screen to take the video of the user.

The Office of the Future system simulates the situation of being across a desk from the other person. For more general immersive systems we have to deal with capturing a user standing up in a more immersive display. The Blue-C system (see Fig. 11.4) is an example of a system that manages to combine vision-based reconstruction with an immersive format display (Gross et al. 2003). The system is able to reconstruct a 3D volumetric model of the avatar inside a CAVE-like system of three walls. The key enabling technology is a type of display surface that can be switched from transparent to opaque, see Fig. 11.4, left. The walls are turned transparent at a high frame rate to capture the user, and when opaque the user's view is blocked and the environment displayed. Simultaneously images from around the user are captured and these are turned into a 3D volumetric model. Figure 11.4, right shows a view of a user standing in front of their own reconstruction.

Recently, with the availability of depth cameras, there has been a lot of interest in reconstruction of static and dynamic scenes. At the time of writing, the state of the art in real-time reconstruction of avatars is typified by the work of Dou et al. (2013). They are able to reconstruct a 3D mesh representation of a person based on a sequence of captured scans from a Microsoft Kinect camera, and then animate that 3D mesh depending on live data from that camera.

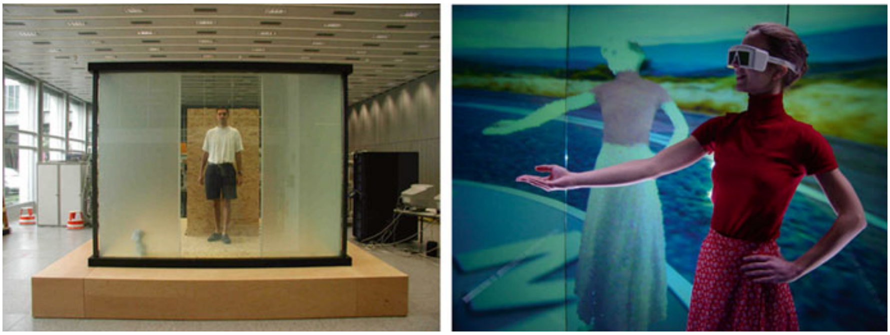

Fig. 11.4 *Left*, the walls of the Blue-C system. *Right*, a user standing in front of their own reconstruction (Both images courtesy of Markus Gross, the blue-c project, ETH Zürich)

Such systems provide us a way to capture a representation of the user into our virtual environment in real-time. However once we have this representation, it is hard to change it. There are two immediate reasons we might have for wanting to change the representation: making the representation appear visually consistent with the virtual environment into which it is inserted, and masking or changing the representation to change the identity or apparent role of the user. In many online games, for example, although users are expected to customise their avatars, customisation is done within some limits imposed by the theme of the world; many of them have strong science fiction or fantasy themes and players are forced, either by the customisation tools, or by the social rules of the system, to build appropriate avatars. More generally, when we look at potential applications, we see that there is a dichotomy emerging: reconstructing the user because this is the easiest way of capturing their posture and emotion; and wanting to hide aspects of this reconstruction such as actual appearance and perhaps even mask or tone down the actual emotion or posture. In the rest of this chapter we argue that even simple geometric avatars can support very successful collaboration between people, and that reconstruction and motion capture might be considered separately to be two "ideals" of immersive environments.

11.3 Impact of Avatars

In the previous discussion we focussed on how a single user is represented within the system. Now we turn to surveying evidence of the impact that representations have on other users. We start by looking at the potential response of a user to a simulated audience. This generates a very effective response, but is a very constrained social situation. In the second section we turn to evidence about interaction between immersed users. We then discuss what is different when we display a modelled or reconstructed avatar, and go on to give some specific examples of comparing different types of avatar representation.

11.3.1 Individual Response

We know that games have a significant impact on their players, and much of this comes from the interaction between players and avatars (Williams et al. 2008). Obviously, no matter the technology, the presence and representation of another person can have significant impact; we see such impacts in visual media such as film and TV. Here we do not want to get into the argument about differences in the impact of media representations, rather we just want to see what the potential space of impacts of avatars can be.

The first evidence we present about the power of avatar representation comes from studies of autonomous audiences of avatars. In a series of studies, Pertaub, Slater and colleagues have used simulations of audiences to investigate phobia of speaking in public (e.g. Pertaub et al. 2001). They simulate a variety of meeting scenarios using a small group of autonomous avatars (avatars with individually programmed behaviours). This is a mediated environment that causes many people, even experienced speakers, some mild anxiety. Experimental subjects who speak in front of an audience that is scripted to behave badly generally have a negative response to the situation on measures of social anxiety. Subjects who speak in front of an audience scripted to behave well, generally have a positive response to the experience. It should be noted that in those experiments, the avatars are not even reacting to the subject, but are following a fixed script of actions that range from applause (in the well-behaved audience) to muttering and turning away from the speaker (in the badly-behaved audience). See Fig. 11.5 for examples of audiences used in later studies in the series.

This system and variations of it have been used for initial trials as tools to assist with the treatment of certain types of mild phobias. Potential paradigms for this include exposure to a series of audiences that react in a more and more hostile manner. What this tells us is that having the avatars there can have an impact, even if the avatars are autonomous. What is uncharacteristic about this situation for the purposes of this chapter is that the user has no clue about the identity of the avatars. The subject might speculate that the avatars represent other individual people, or

Fig. 11.5 *Left*, an attentive audience of avatars. *Right*, a less attentive audience of avatars

that they might be controlled by the experimenter, but this is not supported or encouraged by any information that they are given. So it is left open whether the audience actually represents a group whilst it is in fact almost completely autonomous. The social situation is also constrained so that the subject doesn't attempt to engage with the audience or interact one on one. Of course these are exactly the properties that we need to support in a CVE. In fact, simulating more complex scenarios is very difficult, and the use of avatars even in structured conversations is hard to do satisfactorily (Johnsen et al. 2005).

11.3.2 Responses to User Avatars

Non-immersive CVEs are becoming quite prevalent and services like XBox Live make it very easy for players to log on to network services and find friends or enemies to socialise and play with. Such services have been available for much longer for PC and workstation class machines (e.g. Alphaworld from the Activeworlds Corporation has been active since 1995). Such worlds are well studied and they continue to attract media attention as well as academic attention (e.g., Schroeder 2011; Wardrip-Fruin and Harrigan 2004). However the interaction of people in the CVE and with each other is patently not like interaction in the real world. At one level this is obvious: virtual worlds are not based on real physical laws and social constraints, so why should we expect people to interact with them in that way? At another level it is controversial: obviously they are actually collaborating with another person, so we should rather ask whether this interaction is "normal". Certainly the type of interface has an effect. With systems similar to the Office of the Future system, a smile is captured and transmitted automatically, whereas with a typical game, if it is possible to make the user's avatar smile, this will have to be achieved through some user interface or inferred from the content of the conversation and gesture.

So far, most studies of collaboration in virtual environments have dealt with desktop systems (a variety of studies can be found in Churchill, Snowdon and Munro 2002; Schroeder 2011: 131–38). Further, the focus has typically been on the way in which the individual interacts with the system in order to collaborate rather than on the collaboration itself. This overlooks the complex interplay of the interactions between the avatars inside the virtual environment, though some recent work has examined how avatars interact with each other in terms of the social dynamic (Schroeder 2011: 61–91).

Hindmarsh et al. (2000) showed that collaboration on desktop systems has severe limitations due to the limited field of view and difficulties in referencing parts of the world. The study also shows that participants have problems in being able to take their partner's point view inside the environment. Typical errors that users would make include misinterpreting a pointing gesture or not realising that the other user can not see the object being pointed at. In immersive systems, many of these problems are overcome because of the better capture of participant behaviour through tracking and the wide field of view of the displays (Heldal et al. 2005). This means that participants are much more peripherally aware of their collaborator. Peripheral awareness supports communication about the task at hand but it also supports the maintenance of the collaboration itself since the participants rarely lose track of their collaborator.

A few studies have investigated how collaboration is affected by the use of various combinations of display system. A number of studies have shown that immersed participants naturally adopt dominant roles when collaborating with desktop system participants – even when they don't know what type of the system the other persons are using (Slater et al. 2000; Heldal et al 2005). Studies by Schroeder et al. (2001) and Roberts et al. (2003) have investigated the effect of display type on collaboration of a distributed team. Schroeder et al. (2001) showed that doing a spatial task together using a CAVE-like system, in this case a Rubik's cube type spatial puzzle, can be practically as good as doing the same task face-to-face, whereas the same task takes considerably longer on desktop systems. Roberts et al. (2003) have shown that it is possible to successfully do a construction task (building a gazebo) in networked CAVE-like systems, a task that requires that partners work closely together and in a highly interdependent way. With the cube task and gazebo tasks mentioned above, perhaps the most notable aspect of the interaction is the amount of movement that the users make when gesturing. In the cubes trials we would often see the users making very rapid pointing gestures simultaneously with voice gestures – something that is very hard to synchronise on a puppeteered interface. Users make quite complex spatial references relative to their own body ("on my left"), the body of the other user ("down by your feet") and objects in the environment ("next to the red and blue one"). Breakdowns of these types of reference are rare because it is easy to see whether your collaborator is following your gesture by watching their gaze. Figure 11.6, left shows an example view of two users in CAVE-like systems collaborating over the cube puzzle. Figure 11.6, right shows tracks of the head and hand gestures from a network trial where two users collaborate to build a gazebo (Wolff et al. 2004). The amount of head and hand gesturing is very apparent, and in fact we can even tell a difference between instructor (right) and pupil (left): the instructor makes many more gestures to indicate to surrounding objects and they even pick up a tool to help point. Spatial references of these types are discussed in Steed et al. (2005) and Heldal et al. (2005).

11.4 Presence and Co-presence

In the previous two sections we have discussed technology that affords what we have claimed to be novel styles of collaboration at a distance and we have given preliminary evidence of the impact of these technologies. We now turn to a broader discussion of the factors that might affect co-presence, or interpersonal interaction more broadly conceived. These factors can be grouped into three categories: *modality*, *realism* and *context*.

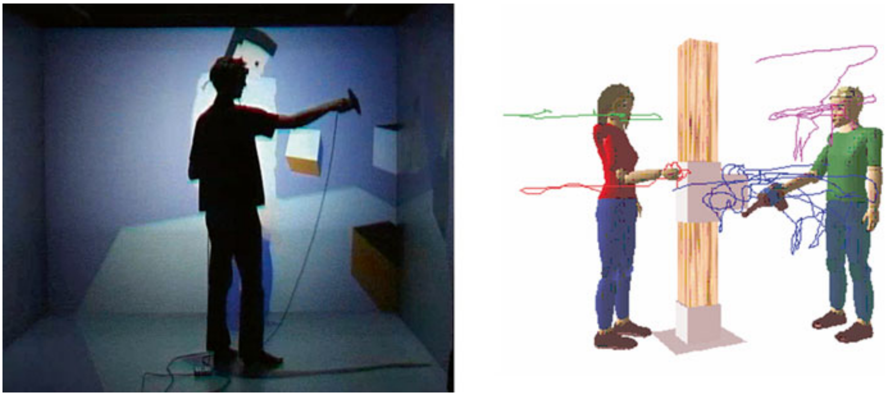

Fig. 11.6 *Left*, two users, one in foreground and one on the screen, in a CAVE-like system collaborating with the representation of another user in the cube task (Image courtesy Ilona Heldal, Chalmers University of Technology, Gothenburg). *Right*, a visualisation of two users in the gazebo task with tracks indicating recent head and hand motion of both (Image courtesy of Robin Wolff, The Centre for Virtual Environments, University of Salford)

11.4.1 Modality

The sensory modality whereby users interact with the system is a good starting point because it is relatively straightforward. The vast majority of systems are visual and auditory. Haptic systems and systems for smell and taste have been developed, and haptic systems will be used in certain settings (Kim et al. 2004), but this essay can confine itself to visual and auditory systems. These two sensory modalities also provide us with the bulk of our information in our face-to-face encounters with others in the physical world.

Two findings are important for CVEs: one is that people ‘compensate’ for missing cues. For example, when they cannot see certain parts of their interaction with each other, they put this part of interaction into words. Conversely, they may use exaggerated body movements to underline something they are saying. How, and under what circumstances they do this, has not been systematically investigated, though there are several potential methods for capturing and analyzing interaction (Schroeder et al. 2006). It is noteworthy that this is something that people will often be unaware of. But clearly, in this respect interaction in immersive CVEs is quite different from face-to-face interaction, and immersive systems differ in terms of how they support auditory and visual interaction. This ‘compensating’ behaviour (which will be quite different for situations with tracked as opposed to reconstructed avatars, for example) is perhaps the single most important aspect of interaction requiring research. Compensating is possibly the wrong term here, since users are also able to ignore the absence of many cues: it would be easy, for example to list a host of visual and auditory cues that users do not comment on as being ‘missing’. Conversely, they are able to make creative use of the ‘superpowers’ that CVEs afford them without finding this remarkable – for example, picking up oversize objects.

The second important aspect of sensory modality is how the senses relate to one another in CVEs. Sallnas (2004), for example, has shown that 'voice' outweighs (or overshadows) the visual sense in the setting that she studied. This finding has important ramifications. Anecdotally (e.g. Finn et al. 1997), the greatest obstacle, or the most annoying feature of, videoconferencing is the sound quality – not the image of the other person. The balance between the two will vary with the applications. But a considerable amount of effort has been devoted to achieving realistic 3D sound, not to speak of realistic visual environments: What if these are far outweighed by being able to hear the nuances in the other person's voice with high fidelity? Much research remains to be done on the interrelation between these two – most common – modalities.

11.4.2 Realism

Realism can be subdivided into several components: eye gaze, facial expressions, body movement and gesture, and the overall appearance of the environments. But apart from these different elements, the critical distinction here is between appearance and behavioural realism (Garau et al. 2003, see also Blascovich 2002), or between faithfulness of the representation of how the avatar looks and how they behave (move, blink their eyes, etc.).

It is well known that eye gaze is critical for interpersonal interaction. Various means of tracking eye gaze have been developed. Note that one basic obstacle for immersive systems (such as the CAVE-like and blue-c systems discussed earlier) is that, if users need to wear 3D glasses to see a 3D space, the system will need to be designed to track the eyes behind the glasses. Garau et al. (2003) showed that a simple model of eye-gaze that takes into account, for example, average eye saccade frequencies, changes the perceived realism, but obviously such a model can't convey important information such as attention.

Eye gaze and facial expression are critical for interpersonal interaction, and bodily movement and gesture for successful instrumental interaction. Note, however that in many circumstances, people seem to be able to cope with highly unrealistic avatars or not to pay much attention to them (Heldal al. 2005).

As for the environment, this is important for orientation. Note that in the environment, cues can be missing in a way that is different from real-world environments. For example, when people walked around in a landscape where many features are similar and where there is no obvious horizon, people complained about not knowing whether they had been to particular landmarks before, and found it difficult in general to orient themselves (Steed et al. 2003; Heldal et al. 2005). In the equivalent real-world scenario, it is much harder to experience this kind of confusion because so many cues in a landscape tell us where we have been (horizon, different experience of objects in relation to each other, etc.) The use of landmarks or other tools for orientation (or footprints to mark where one has been) are easy to implement, but again, a key question is in which circumstances these are needed and effective.

11.4.3 Context

The importance of context is obviously multifaceted; unlike the other two which are clearly delimited, this is a catchall category. Therefore context can be broken down into subcomponents:

What is the relation to the other person(s)? Are they people one is familiar with, or people one is interacting with for the first time (Steed et al. 2003)? What is the task? Perhaps it is unspecific socializing, in which case it seems inappropriate to call it a 'task'. And finally, but not least, what is the size of the group? If, for example, one is interacting with a larger group, it is difficult in a CVE, unlike in the real world, to monitor the behaviour of several copresent others simultaneously. Put differently, when one is interacting with several other people in the VE, does the attention one can pay to any one of the other people become 'diluted'? (This is much more likely in a VE because mutual awareness is more difficult).

One reason for making these distinctions is that they highlight the combinations of features that CVEs need, as well as those that are unlikely. For example, in the various applications used in the Strangers and Friends trial (Steed et al. 2003), there are many examples when the tracked bodies and gestures were critical to joint coordination, but the absence of eye gaze and facial expressions was not an important obstacle in this set of tasks.

This draws attention to a crucial point: in *immersive* collaborative systems, the task will likely be one in which people have to focus their attention on the space and the objects in it (which includes, for joint orientation, the other person(s) avatar body), but in these systems people may not need to focus on each other's facial expressions. Furthermore, they may not need realistic-looking bodies; it will be sufficient to be able to follow the other's movements and gestures – their appearance is irrelevant for tasks such as manipulating objects together, building things together, exploring the space and the like (Steed et al. 2003). One way to underline this is by noting that if there is more than one other person in the immersive space, the most important feature of the avatar bodies of others is that the user is able to tell them apart, not what they look like. Note that these features – a small group of tracked life-size avatars, their bodies perhaps distinguished by being different colours (Mr. Blue, Mr. Green, etc.) – will, in turn, have an important, perhaps 'overshadowing', influence on co-presence.

If we now add that immersive spaces are likely to contain only a small number of (non- co-located) people at any given time, it is possible to get a sense of the requirements of immersive spaces for collaboration: for instrumental tasks, all those aspects of the environment that facilitate joint orientation and manipulation should be adequate to the task (whereas appearance of the avatar, including expression, is relatively insignificant). In contrast, for tasks mainly involving interpersonal communication, facial expressions will be important – *but*, it is unlikely that these will play a dominant role in a shared immersive *space*: after all, people will not spend much time in close face-to-face contact in these spaces. When eye gaze *is* useful in this case, it will be mainly for people to indicate to the other person where they are looking (as opposed to, say, conveying their mood or emotional state) (Steptoe et al. 2009). Finally, there are various ways to design expressive avatar faces that have the capability to facilitate interaction without relying on capturing the user's real facial expression or their eye gaze (Bailenson and Beall 2006; Garau et al. 2003).

In immersive spaces then, the expressiveness of faces (including eye gaze) is likely to be highly context-dependent: the office in which one collaborates with another person in a trauma counselling or public speaking training or acting session (where facial expressions are critical) will be quite different from that required for a molecular visualization or vehicle design session (where joint orientation and referencing objects is most important). Perhaps an avatar face with the possibility to express only certain emotions or certain acknowledgements of the other person's effort will not only be sufficient in immersive space – but superior since it will reduce the 'cognitive load' in the task.

11.5 End-States

Many of the issues in the study of co-presence and collaboration can be illuminated by considering two end-states of CVE technology: captured versus puppeteered or tracked (for the following, see also the extended discussion in Schroeder 2011: 275–92). In the following discussion we will use the term *simulated avatars* to refer collectively to puppeteered or tracked avatars.

In the simulated avatars end-state, the environment can be configured so that any appearance and different behaviours are possible. In particular the appearance of the avatar is modelled prior to the experience so that it can fit with the visual appearance of the world. For example, everyone in a game such as World of Warcraft has a user avatar that fits with the overarching fantastic visual theme of that world. With captured avatars, such as the capture of the person and the scene in blue-c, appearance is limited to a faithful recreation of real world. This latter will have some advantages from the user's point of view: since they know what to expect, they can experience the environment (and also the devices that they use and that are used to create it) naturally and behave accordingly. The point is, however, that even the other end-state, of completely computer-generated artificial worlds with simulated avatars, will need to be designed so as to put constraints and possibilities into the environment that the user experiences as being at ease with; an environment that they feel at home in and that they can establish good interpersonal relations in. And here, as we have seen, users are able to accept certain 'unnatural' features of CVEs (not caring about avatar appearance), they adapt easily to some others (absence of touch), and find yet others impossible or difficult to cope with (being unable to distinguish between others' avatars). Nevertheless, CVEs will need to provide them with a place for being there together in which they are able to do things and interact with each other as they need to, for a variety of technologies and situations.

A simple point that highlights this difference between the two end-states is that in a captured environment, people will be certain of another person being there, just as in a videoconference (they are, after all, being captured). In generated CVEs with simulated avatars, on the other hand, mechanisms need to be put in place to ensure that users are 'really there' since the presence of avatar is not sufficient to establish that the person that was controlling that avatar is still connected to the system. Even if the avatar is moving, it may be automated or someone else may have taken control. This is taken to its logical conclusion in experiments in the BEAMING project, where avatars can blend between control by a human through to complex automated behaviour (Friedman and Tuchman 2011).

If we think about general captured and simulated immersive environments and what they may one day develop into, then it becomes clear that much of the technology is already in place, and that two end-states will be quite different: captured environments will take the form of 3D holographic videoconferencing. In other words, they will be similar to the blue-c system, except that they will be able to capture larger extended spaces accurately and put many interacting people into the shared space without the encumbrances of 3D glasses and the like. Simulated environments, on the other hand, will be extensions of today's immersive systems, though again, the environments and avatars will appear completely realistic (including in behaviours) and again, the encumbrances of 3D glasses and position-tracking equipment and the like will be minimized. In other words, both types of systems will provide perfect presence, co-presence and interaction with the environment – except that in the one case, the environment will reproduce persons and the world around them in 3D, and in the other, it will generate persons' likenesses and virtual worlds.

A more realistic expectation is that there will be a variety of systems that approximate these end-states, and these approximations are *unlikely* to be simply steps towards either *completely* realistic computer-generated or 3D video-captured systems and environments. Instead, they will reflect the combination of particular features that are required for successful interpersonal interaction and interaction with the environments. For example, there may be environments that combine captured faces with generated environments, or vice versa. Additionally the environments will have different spatial extents: some will display the face-to-face extent plus perhaps some nearby objects that people are working on together, others will display the extended extent of a large space that needs to be jointly visualized or explored. Again, these may not be realistic environments, but, for example, environments which focus on the fidelity of certain parts of the environments and not others, feature certain facial characteristics that convey essential information but leave out a host of information that is conveyed in face-to-face information, and consist of environments designed to facilitate easy orientation and mutual awareness by means of various 'artificial' features. These 'artificial' features may, for example, consist of facial expressions that are 'enhanced' to facilitate interpersonal awareness, or 'enhanced' to provide a better awareness of the environment.

It is possible then to recognize that the two end-states, with their quite different possibilities and constraints, may be combined in some way. It may be that the computer-generated end-state has distinct advantages in being much more flexible in terms of which features of modality, appearance (the face, body and environment) and context can be combined to support interaction in different ways. The constraint in this case is that the lack of realism will need to be compensated for in particular ways. The video-captured end-state, on the other hand, offers different possibilities, for example providing a realism that the user can trust in a different way, but it is constrained by capturing the real appearance of people and of the environment without being able to enhance or reconfigure them in a powerful way.

The combination of thinking about two end-states and thinking about systems for captured and simulated environments on the way towards them therefore allows us to recognize that there are different types of affordances and requirements that will be necessary for various scenarios for CVEs. We are still far from a good understanding of the likely future uses and configurations of immersive CVE systems. However, we can channel research towards forms of CVEs and CVE uses that will yield insights about the end states we have identified. These insights can then benefit the improvement of tools that support collaboration at-a-distance.

11.6 Challenges

We have described the range of current CVE technologies from computer games consoles through to highly-immersive CAVE-like systems that support real-time capture of the user standing within them. Given the fact that people invest so much time in them, collaboration through desktop interfaces has the capability to be compelling, though it is easy to see that in many ways people do not collaborate together in a similar way as they would in the real world. In an immersive system we see some evidence of people behaving as if the situation were the real world – that is, using voice and gesture as they might in a similar situation in the real world. We also see complex gestures and very fast paced interaction of types that are impossible in other media.

The question we have opened up is how CVE technology will develop in the long term. There is a push towards making real-time captured avatar systems, where the users have a faithful 3D representation of their collaborators. However we have argued that supporting presence and co-presence can be done with simulated avatars, and in some situations these will be preferred.

Aside from obvious technical challenges in further developing captured, tracked and puppeteered avatars, there are many challenges in studying collaboration with these technologies and designing to support better collaboration. We can do this by looking at remaining misunderstandings, looking at personality bias arising from collaborative situations and studying how people use these technologies over longer periods. One challenge we would highlight is understanding how well collaborators understand the intention of the others as this is one key to successful communication – being able to tell what the other person intends to do based on the subtle gestures and eye gaze alongside their speech.

References

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

Bailenson, J., & Beall, A. (2006). Transformed social interaction: Exploring the digital plasticity of avatars. In R. Schroeder & A.-S. Axelsson (Eds.), *Avatars at work and play: Collaboration and interaction in shared virtual environments* (pp. 1–16). London: Springer.

Blascovich, J. (2002). Social influence within immersive virtual environments. In R. Schroeder (Ed.), *The social life of avatars: Presence and interaction in shared virtual environments* (pp. 127–145). London: Springer.

Brown, B., & Bell, M. (2006). Play and sociability in there: Some lessons from online games for collaborative virtual environments. In R. Schroeder & A.-S. Axelsson (Eds.), *Avatars at work and play: Collaboration and interaction in shared virtual environments* (pp. 227–246). London: Springer.

Cheng, L., Farnham, S., & Stone, L. (2002). Lessons learned: Building and deploying shared virtual environments. In R. Schroeder & A.-S. Axelsson (Eds.), *Avatars at work and play: Collaboration and interaction in shared virtual environments* (pp. 90–111). London: Springer.

Churchill, E., Snowdon, D., & Munro, A. (Eds.). (2002). *Collaborative virtual environments: Digital places and spaces for interaction*. London: Springer.

Dit Picard, S. L., Degrande, S., Gransart, C., & Chaillou, C. (2002). VRML data sharing in the spin-3D CVE. In *Proceeding of the seventh international conference on 3D Web technology* (Tempe, Arizona, USA, February 24–28, 2002). Web3D'02 (pp. 165–172). New York: ACM Press.

Dou, M., Fuchs, H., & Frahm, J.-M. (2013). Scanning and tracking dynamic objects with commodity depth cameras. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 99–106). IEEE.

Finn, K., Sellen, A., & Wilbur, S. (Eds.). (1997). *Video-mediated communication*. Mahwah: Lawrence Erlbaum.

Friedman, D., & Tuchman, P. (2011). Virtual clones: Data-driven social navigation. In *Intelligent virtual agents* (Lecture notes in computer science, Vol. 6895) (pp. 28–34). London: Springer.

Garau, M., Slater, M., Vinayagamoorhty, V., Brogni, A., Steed, A., & Sasse, M. A. (2003, April 5–10). The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In *Proceedings of the SIG-CHI conference on Human factors in computing systems* (pp. 309–316). Fort Lauderdale: ACM.

Gaver, W. W., Sellen, A., Heath, C., & Luff, P. (1993, April). One is not enough: Multiple views in a media space. In *Proceedings of INTERCHI'93* (pp. 335–341). Amsterdam: ACM.

Gross, M., Würmlin, S., Naef, M., Lamboray, E., Spagno, C., Kunz, A., Koller-Meier, E., Svoboda, T., Van Gool, L., Lang, S., Strehlke, K., Moere, A. V., & Staadt, O. (2003). Blue-c: A spatially immersive display and 3D video portal for telepresence. *ACM Transactions on Graphics, 22*(3), 819–827.

Heldal, I., Schroeder, R., Steed, A., Axelsson, A.-S., Spante, M., & Widestrom, J. (2005a). Immersiveness and symmetry in copresent scenarios. In *Proceedings of IEEE VR* (pp. 171–178). Bonn: IEEE.

Heldal, I., Steed, A., Spante, M., Schroeder, R., Bengtsson, S., & Partanan, M. (2005b). Successes and failures in copresent situations. *Presence: Teleoperators and Virtual Environments*, *14* (5), 563–579.

Hindmarsh, J., Fraser, M., Heath, C., Benford, S., & Greenhalgh, C. (2000). Object-focused interaction in collaborative virtual environments. *ACM Transactions on Computer-Human Interaction (ToCHI), 7*, 477–509.

Hinds, P., & Kiesler, S. (Eds.). (2002). *Distributed work*. Cambridge, MA: MIT Press.

Johnsen, K., Dickerson, R., Raij, A., Lok, B., Jackson, J., Shin, M., Hernandez, J., Stevens, A., & Lind, D. S. (2005, March 12–16). Experiences in using immersive virtual characters to educate medical communication skills. In *Proceedings of the 2005 IEEE conference 2005 on Virtual Reality* (pp. 179–186). Washington, DC: VR. IEEE Computer Society.

Jung, M., Fischer, R., Gleicher, M., Thingvold, J. A., & Bevan, M. (Eds.). (2000). *Motion capture and editing: Bridging principle and practice*. Natick: A K Peters.

Kim, J., Kim, H., Tay, B. K., Muniyandi, M., Jordan, J., Mortensen, J., Oliveira, M., Slater, M., & Srinivasan, M. A. (2004). Transatlantic touch: A study of haptic collaboration over long distance. *Presence: Teleoperators and Virtual Environments, 13*(3), 328–337.

Pertaub, D.-P., Slater, M., & Barker, C. (2001). An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators and Virtual Environments, 11*(1), 68–78.

Pollick, F. E., Lestou, V., Ryu, J., & Cho, S. B. (2002). Estimating the efficiency of recognizing gender and affect from biological motion. *Vision Research, 42*, 2345–2355.

Raskar, R., Welch, G., Cutts, M., Lake, A., Stesin, L., & Fuchs, H. (1998). The office of the future: A unified approach to image-based modeling and spatially immersive displays. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques SIGGRAPH'98* (pp. 179–188). New York: ACM Press.

Roberts, D., Wolff, R., Otto, O., & Steed, A. (2003). Constructing a Gazebo: Supporting team work in a tightly coupled, distributed task in virtual reality. *Presence: Teleoperators and Virtual Environments, 16*(6), 644–657.

Sallnas, E.-L. (2004). *The effect of modality on social presence, presence and performance in collaborative virtual environments*. Ph.D. thesis, Royal Institute of Technology, Stockholm.

Scheumie, M. J., van der Straaten, P., Krijn, M., & van der Mast, C. (2001). Research on presence in virtual reality: A survey. *Cyberpsychology and Behaviour, 4*(2), 183–201.

Schroeder, R. (Ed.). (2002). *The social life of avatars: Presence and interaction in shared virtual environments*. London: Springer.

Schroeder, R. (2011). *Being there together: Social interaction in virtual environments*. Oxford: Oxford University Press.

Schroeder, R., & Axelsson, A.-S. (Eds.). (2006). *Avatars at work and play: Collaboration and interaction in shared virtual environments*. London: Springer.

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

Schroeder, R., Heldal, I., & Tromp, J. (2006). The usability of collaborative virtual environments and methods for the analysis of interaction. *Presence: Journal of Teleoperators and Virtual Environments, 15*(6), 655–667.

Slater, M., Sadagic, A., Usoh, M., & Schroeder, R. (2000). Small group behaviour in a virtual and real environment: A comparative study. *Presence: Teleoperators and Virtual Environments, 9*(1), 37–51.

Steed, A., Spante, M., Schroeder, R., Heldal, I., & Axelsson, A.-S (2003, April 27–30) Strangers and friends in caves: An exploratory study of collaboration in networked IPT systems for extended periods of time. In *ACM SIGGRAPH 2003 Symposium on Interactive 3D Graphics* (pp. 51–54). Monterey: Lawrence Erlbaum.

Steed, A., Roberts, D., Schroeder, R., & Heldal, I. (2005). Interaction between users of immersion projection technology systems. In *Proceedings of Human Computer Interaction International 2005*, 22–27 July, Las Vegas.

Steptoe, W., Oyekoya, O., Murgia, A., Wolff, R., Rae, J., Guimaraes, E., Roberts, D., & Steed, A. (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). IEEE Computer Society.

Wardrip-Fruin, N., & Harrigan, R. (2004). *First person: New media as story, performance, and game*. Cambridge: MIT Press.

Weaver, A. (2010). *How the Jaguar Land Rover headquarters tests new vehicles.* Wired, UK. http://www.wired.co.uk/magazine/archive/2010/12/start/car-design-goes-virtual. Accessed 16 Jan 2014.

Williams, D., Ducheneaut, N., Li, X., Zhang, Y., Yee, N., & Nickell, E. (2006). From tree house to barracks: The social life of guilds in world of Warcraft. *Games and Culture, 1*, 338–361.

Williams, D., Yee, N., & Caplan, S. (2008). Who plays, how much, and why? A behavioral player census of virtual world. *Journal of Computer Mediated Communication, 13*, 993–1018.

Wolff, R., Roberts, D. J., & Otto, O. (2004, October). Collaboration around shared objects in immersive virtual environments. In *Proceedings of 8th IEEE international symposium on Distributed Simulation and Real-Time Applications (DS-RT'04)* (pp. 206–209). Budapest: IEEE.

Yee, N. (2006). The psychology of massively multi-user online role-playing games: Motivations, emotional investment, relationships and problematic usage. In R. Schroeder & A.-S. Axelsson (Eds.), *Avatars at work and play: Collaboration and interaction in shared virtual environments* (pp. 187–208). London: Springer.